Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties

Linda Polka[1], Matthew Masapollo[1] & Lucie Ménard[2]

McGill University[1] and Université du Québec à Montréal[2]

Keywords: Speech Perception, Infancy, Categorization

Running Head: INFANTS' PERCEPTION OF INFANT VOWELS

Article Type: Research article

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

Abstract

Little is known about infants' abilities to perceive and categorize their own speech sounds or vocalizations produced by other infants. In the present study, prebabbling infants were habituated to /i/ ("ee") or /a/ ("ah") vowels synthesized to simulate men, women, and children, and then were presented with new instances of the habituation vowel and a contrasting vowel on different trials, with all vowels simulating infant talkers. Infants showed greater recovery of interest to the contrasting vowel than to the habituation vowel, which demonstrates recognition of the habituation vowel category when it was produced by an infant. A second experiment showed that encoding the vowel category and detecting the novel vowel required additional processing when infant vowels were included in the habituation set. Despite these added cognitive demands, infants demonstrated the ability to track vowel categories in a multitalker array that included infant talkers. These findings raise the possibility that young infants can categorize their own vocalizations, which has important implications for early vocal learning.

Keywords
speech perception, infancy, vowel categorization, babbling, talker variability

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

Understanding spoken language involves categorization processes at many levels. The perception of phonemes, the smallest speech units that carry meaning, entails recognizing phonetic categories in the context of enormous acoustic variability. One of the largest sources of acoustic variability that the perceiver must contend with is due to the wide variation in physical attributes of different talkers related to age and gender (Kuhl, 2004). Developmental research has demonstrated that young infants can differentiate phonetic categories across different talkers in some test situations (Kuhl, 1979; 1983; Jusczyk, Pisoni, & Mullennix, 1992). Other research, however, suggests that this process is initially difficult for infants, particularly when the task is more complex or the talkers are more acoustically dissimilar (e.g., Houston & Jusczyk, 2000). Current research is focused on understanding how the ability to process phonetic information in complex multi-talker contexts develops and improves during infancy. This work is motivated to understand the acquisition of speech *comprehension*. The ability to track phonetic information across talker differences has not been addressed from a perspective that considers the perceptual resources that infants need to acquire speech *production* skills. It is not enough for infants to contend with the talker differences they encounter in communicative interactions. They must also recognize the phonetic equivalence between speech sounds produced by their caregivers and the sounds they *themselves* produce, which entails perceiving speech produced by an *infant* talker. Research has been silent on this issue because infant talkers have been silent in the lab.

Infants must accurately perceive and categorize their own self-produced speech in order to effectively monitor and refine their productions during vocal learning (Kuhl, 1979; Kuhl & Meltzoff, 1996; Rvachew, Mattock, Polka, & Ménard, 2006; Doupe & Kuhl, 2008). However, the existing data suggest two different hypotheses regarding the interaction between the development of speech perception and production capacities, which make different assumptions about the perceptual resources that young infants can harness to perceive infant speech at the onset of vocal learning.

In the "*high resource/imitation*" view infants' perceptual skills develop well in advance of production skills and provide a critical infrastructure supporting emerging production skills that are guided by imitation (Kuhl & Meltzoff, 1996). Under this view, young infants have sophisticated perceptual capacities that allow them to perceive phonetic categories across different talkers, including infants. These perceptual skills are in place before infants begin producing speech; infants can rely on these skills to learn to accurately imitate phonetic categories in their ambient language input. This account is bolstered by evidence that 3- to 5-month-olds modified their vocalizations in response to vowels produced by an adult talker on a television screen (Kuhl and Meltzoff, 1996).

In the "*low resource/interaction*" view speech perception and production skills develop concurrently, guided by exchanges in an interactive context (e.g., Howard & Messum, 2011; Zlatin & Koenigsknecht, 1976). Under this view, infants require experience producing and comparing their own vocalizations with caregiver vocalizations, during face-to-face social interactions to learn how their speech output corresponds to "target" sounds in the ambient language. Accordingly, the ability to recognize phonetic category equivalence across talkers is only partially developed in young infants; they must initially rely on their caregiver's imitative and

affective responses to indicate when *their* productions perceptually match native language sounds. Supporting this view, studies show caregivers frequently imitate the vocalizations of their young infants (Pawlby, 1977) and provide social stimulation (e.g., smiling and/or touching) that facilitates more advanced vocal behavior (Goldstein *et al*., 2003).

Importantly, the ability to recognize phonetic categories in infant speech is a prerequisite in the "high resource/imitation" view, but not in the "low resource/interaction" view. Presently, there are no data pertaining to infant perception of infant speech to select between these alternative hypotheses. Using technical advances in speech synthesis to generate infant speech, we investigate for the first time, how infants perceive vowels produced by an *infant* talker. Several additional factors suggest why the development of this skill is challenging. First, vowels produced by an infant are acoustically distinct because the infant's vocal folds and vocal tract are much shorter than those of an adult or child (Kent & Murray, 1982; Kuhl & Meltzoff, 1996; Rvachew *et al*., 2006; Vorperian & Kent, 2007; Ménard *et al*., 2004). The fundamental frequency (corresponding to voice pitch) and the formant frequencies (corresponding to the vocal tract resonances) observed in infant vocalizations fall well above the values found in adult or child speech. Infant vowels span a large acoustic space, which overlaps only partially with the adult and child acoustic space. This is illustrated in Figures 1 and 2. Introducing infant vowels thus increases the range of acoustic variation that infants encounter in their input. Second, until they begin to babble, most infants will have limited experience listening to infant speech. Third, tracking phonetic information across acoustically dissimilar talkers is cognitively demanding for infants (Houston & Jusczyk, 2000). Together, these factors suggest that perceiving vowels in a multi-talker context that includes infant talkers may be particularly demanding for young infants.

We address these issues in two experiments designed to determine whether young (4-to-6-month-old) infants, who are not yet producing canonical babble, can recognize the same vowel when produced by adult, child, *and* infant talkers. We synthesized isolated vowels, /i/ ("ee") and /a/ ("aw"), to simulate productions by men, women, children, and infants. At 4-6-months infants are developing productive control of fully resonant vowels; their vowel productions typically fall within a narrow range of the vowel space with front, low and central vowels preferred. Expansion of the infant vowel space to stabilize the corner vowels - /i/, /u/, /æ/ and /a/ - requires another year of practice (Rvachew, Slawinski, Williams, & Green, 1996). Thus, infants in this study will have had only limited exposure (from their own speech) to the full range of vowel qualities that can be produced by an infant vocal tract (Kent & Murray, 1982).

We chose the vowels /i/ ("ee") and /a/ ("aw") because previous research has shown that 5-to-6-month-olds can perceive this phonetic contrast in a multi-talker context. Kuhl (1979) tested infants with vowels synthesized to emulate men, women and children, using the conditioned head-turn procedure. Using vowels synthesized to control for acoustic dimensions that are not related to talker age and gender, we tested infants using the visual habituation procedure which does not involve conditioning or training. In this procedure, infants were first habituated to a set of vowel exemplars (e.g. /i/) produced by different talkers. As soon as the infant has

habituated we present four test trials - two with new exemplars from the same (familiar) vowel category (e.g., /i/) and two with new exemplars from a novel vowel category (e.g., /a/). Importantly all vowels in test trials (familiar and novel) were produced by a new talker who was not encountered during habituation. If infants have formed a category representation for the habituation vowel, we expect them to recognize the novel vowel as belonging to a different phonetic category, and to show this by recovering interest and listening longer to the novel vowel than to the familiar vowel. Infants may also recognize the change in talker and show an increase in listening to the familiar vowel (produced by a new talker) compared to the final habituation trials. However, despite recognizing the talker change, infants should show a larger recovery of interest to the novel compared to the familiar test vowel, if they recognize that the familiar test vowel is another instance of the vowel presented during habituation.

Experiment 1

Methods

Participants. Data was analyzed for fifty-six infants, aged 4-to-6-months (36 male, *M* = 161 days, *R* = 138 – 197 days); all were exposed to languages in which /i/ and /a/ are phonemic with the majority from English- or French-speaking families. Twenty-three additional infants were excluded due to fussiness (15), caregiver interference (3), or experimental error (5). All were full-term with no known health problems.

Stimuli. We selected 8 /i/ and 8 /a/ isolated vowels from a large corpus of vowels synthesized using the Variable Linear Articulatory Model (VLAM), described in Ménard, Schwartz, & Boe (2004). This corpus simulates talkers across a broad age range from infancy to adulthood. The selected vowels emulate eight different talkers: two 6-month-olds, an 8-, 10-, and 12-year-old, two adult females, and one adult male. Details of the VLAM synthesis, acoustic description of the vowels, and audio samples are provided in the supplemental online material.

All vowels were 500 ms long and matched in intensity and intonation contour. The stimuli were judged to be intelligible, natural-sounding exemplars of each vowel category by English- and French-speaking adults. Adults also accurately identified the age and gender differences simulated in the stimulus set. For testing, we created 16 stimulus files (one per vowel); each 30-second file included 20 repetitions of the same vowel with a 1000 ms inter-stimulus interval.

Procedure. Infants were tested using the visual habituation (look-to-listen) procedure (Polka, Jusczyk, and Rvachew, 1995). The infant

sat on the caregiver's lap at a distance of about 150 cm facing a 21-inch television monitor in a dimly lit curtained soundproof booth. Audio TRAK BSI-90 loudspeakers and a Sony digital video camera were located behind the curtain just below the TV screen. An experimenter observed the infant outside of the testing room on a monitor linked to the video camera. The caregiver wore noise-canceling headphones and listened to masking music during the entire procedure to avoid influencing the infant's behavior. Experimental stimuli were presented and looking/listening[1] times monitored using the software Habit 2000 (Cohen, Atkinson & Chaput, 2000). At the start of each trial, a red flashing light is presented to direct attention followed by a black-and-white-checkerboard. The experimenter (who could not hear the stimuli) pressed a key when the infant fixated on the checkerboard, which activated the presentation of the auditory stimulus, providing an index of the infant's listening time. When the infant looked away for more than 2 seconds, the sound stopped and the screen went black. The minimum look time for a trial was 1 second. If the infant looked away for less than 2 seconds the sound continued to play but the look away time was not included in the looking time for that trial. The trial was terminated when the infant looked away for more than 2 seconds or when the complete stimulus file (30 s long) had played. After a brief pause, the attention-getter returned to start the next trial.

Design. The experiment consisted of four consecutive phases: pre-test, habituation, test, and post-test (see Figure 3), with no break or pause between. Instrumental music was presented during pre- and post-test trials. On each habituation and test trial infants heard a vowel produced repeatedly by the same talker; a different talker was presented in each trial. The vowel presented during habituation was counterbalanced across subjects. During habituation the order of talkers was block-randomized so that every 3 trials included a man, woman, and a child talker. The software tracked a running average of listening time across a 3-trial window. The habituation criterion was met when the running average dropped below 65% of the longest 3-trial average. Thus, the number of habituation trials varied across infants; however, all infants completed at least 4 habituation trials (most completed 6 or more) and were exposed to all three talker types (man, woman, child).

During the test phase, four trials containing infant vowels were presented; two with the same vowel as habituation (F=familiar) and two with the contrasting vowel not heard during habituation (N=novel). Test trials were presented in one of two fixed orders: FNNF or NFFN. Infants were assigned to 4 conditions (2 habituation conditions; 2 test orders) as shown in Figure 3.

Results

The two habituation groups were combined because they showed no differences in total habituation time, number of habituation trials or post-test listening times. For each infant, listening time was averaged across the last two trials in habituation and for each test trial type (novel; familiar). Group means (collapsed across habituation condition and test trial order) are shown in Figure

4a. The scores were submitted to a mixed ANOVA with test trial order (FNNF vs. NFFN) as between-subjects factors, and trial type (habituation vs. familiar vs. novel) as a within-subjects factor. There was no main effect of test trial order or interaction with trial type, F < 1. There was a main effect of trial type, $F(2, 108) = 23.81$, $p = < .001$, $\eta_p^2 = .306$. Post hoc comparisons showed that infants listened longer to the novel (M_N = 11.7 s, SD = 6.3) than the familiar test vowel, (M_F = 9.8 s, SD = 6.1), $t(55) = -2.46$, $p = .017$, $r^2 = .314$, and longer to the familiar test vowel than the habituation vowel (M_{Hab} = 6.5 s, SD = 3.0), $t(55) = 4.30$, $p < .001$, $r^2 = .501$.

Discussion

Experiment 1 shows that young infants can track a change in vowel category and a change in talker in a multi-talker context that includes an infant talker. Thus, young infants have some ability to recognize vowel categories across acoustically and perceptually distinct talkers. Because infants encountered the infant vowels at the end of the task, Experiment 1 provides little insight into the processing demands involved in this task. Experiment 2 addresses this issue.

Experiment 2

There is clear evidence that perceiving speech in a multi-talker context entails some processing costs for infants as well as adults (e.g., Mullenix & Pisoni, 1990; Jusczyk, Pisoni & Mullennix, 1992; Houston & Jusczyk, 2000; Schmale & Seidle, 2009). Moreover, stimulus complexity affects infant processing and category formation in many domains (Hunter & Ames, 1988). Thus, we reasoned that because infant speech augments stimulus complexity in a multi-talker scenario, it might increase the cognitive demands associated with perceiving speech in a multi-talker context. We tested this prediction in Experiment 2 using the same task and stimuli as Experiment 1, but with one change – the infant vowels were included in the habituation set (replacing the adult male vowels), and the adult male vowels were presented in the test phase (replacing the infant vowels). This manipulation increased the acoustic variability present in the habituation set, but not the number of talkers. We predicted that this change would impact stimulus complexity and increase cognitive demands in the categorization task.

If this is the case, then infants will need more time to encode the phonetic category information in the habituation stage well enough to recognize the novel category in the test phase. This leads to two specific predictions. First, infants will listen longer during habituation in Experiment 2 compared to Experiment 1 before reaching the habituation criterion. Second, infant categorization performance will be affected by their level of engagement during the habituation, i.e., total listening time during habituation will be *positively* correlated with the magnitude of the novelty effect observed in test trials. It is important to note that prior categorization studies using this paradigm show that individual differences in categorization ability are typically *negatively* correlated with total

habituation time. That is, in both auditory and visual tasks, infants who are better categorizers process stimuli more efficiently and thus require less time to habituate to a stimulus set (Polka, Rvachew & Molnar, 2008: Arterberry and Bornstein, 2002). Thus, a positive correlation between total habituation time and novelty score reflects an increase in stimulus complexity and processing load rather than individual differences in categorization ability. Total habituation time and magnitude of novelty scores were not correlated in Experiment 1 (r^2 = .209, p = .122).

Method

Participants. Data from twenty-seven 4- to 6-month-olds (10 males) were analyzed (M = 158 days, R = 132 – 195 days; language background was the same as Experiment 1. Nine additional infants were excluded due to fussiness (7), failure to habituate (1), or experimental error (1). All were full-term with no known health problems.

Stimuli and Procedure. Same as Experiment 1.

Design. The design was identical to Experiment 1 expect that all infants were habituated to /i/, the habituation trials included vowels simulating an infant (but not a man) and vowels presented on test trials simulated a man.

Results

Total listening time during habituation (summed across habituation trials) was computed for each infant. Group means for infants in Experiment 2 and Experiment 1 (/i/ habituation condition) are plotted in Figure 5 which shows that infants listened significantly longer during habituation in Experiment 2 compared to Experiment 1, $t(55) = -3.41$, p = .001, r^2 = .417. Thus, as predicted infants required more time to encode the variability associated with the different talkers when infant vowels were present in the habituation set (M = 118.1 s, SD = 12.9) compared to when only adult and child vowels were present (M = 71.6 s, SD = 5.7); the number of talkers was the same in both experiments. There was no difference across experiments in the number of habituation trials to reach criterion or in post-test listening times.

For each infant, listening time was averaged across the last two habituation trials and for each test trial type (novel; familiar). Group means for these scores (collapsed across test trial order) are shown in Figure 4b. As in Experiment 1, these scores were submitted to a mixed ANOVA with test trial order (FNNF vs. NFFN) as a between-subjects factor, and trial type (habituation vs. familiar vs. novel) as a within-subjects factor. There was no main effect of test trial order or interaction with trial type, $F < 1$. There

was a main effect of trial type, $F(2, 50) = 3.57$, $p = .035$, $\eta_p^2 = .125$. Post hoc comparisons showed that infants listened longer to the novel than to the familiar test vowel ($M_N = 11.9$ s, $SD = 6.7$; $M_F = 8.6$ s, $SD = 4.0$), $t(26) = -2.29$, $p = .031$, $r^2 = .409$. Listening times to the habituation vowel ($M_{Hab} = 9.3$ s, $SD = 3.8$), and the familiar test vowel were not significantly different, $t(26) = .633$, $p = .532$..

To compare the novelty response across experiments, a mixed ANOVA was conducted with experiment (Exp 1- /i/ habituation group vs. Exp. 2) and as a between-subjects factor, and test trial type (familiar vs. novel) as a within-subjects factor. There was no effect of experiment or interaction with test trial type, $F < 1$. There was only a main effect of test trial type, $F(1, 55) = 5.10$, $p = .028$, $\eta_p^2 = .085$, showing that the novelty response was comparable across experiments. As predicted total habituation time was positively correlated with the size of the novelty score (listening time on novel test trials minus familiar test trials); $r^2 = .460$, $p = .014$). A follow-up analysis showed that roughly half of the infants in Experiment 2 ($n = 13$) displayed total habituation listening times comparable to Experiment 1 (within 1 *SD*) and half ($n = 14$) listened much longer with levels more than 1 SD above the mean of Experiment 1 ($M = 164.19$; $SD = 68.95$). As shown in Figure 6 a reliable novelty effect was observed *only* in the long listener subgroup in Experiment 2, $t(12) = -2.42$, $p = .032$. Thus increased listening time during habituation is clearly linked to successful recognition of the novel vowel in this task.

Discussion

As predicted, processing demands increased when the infant vowels were part of the stimulus set that infants needed to encode to form a habituation category. Despite the added demands infants were able track changes in vowel quality and displayed a novelty response comparable to the effect observed in Experiment 1. However, unlike Experiment 1, in Experiment 2 the magnitude of the novelty score was directly related to the amount of listening time invested during habituation. Overall, Experiment 2 shows that including infant vowels in a multi-talker context increases processing demands but the added costs fall within the cognitive abilities of infants.

General Discussion

The perception of talker variability is a focal issue in speech perception research with adults and infants. Until now infant speech has been left out of the picture despite its relevance for infant development and for conceptual views of speech perception. In the present study, we examined pre-babbling infants' ability to recognize vowel categories in a multi-talker context that includes an infant talker. Our findings show that infants' ability to track vowel categories across talkers extends to *infant* vowel productions.

Given the limited overlap between infant vowels and adult or child vowels (Figure 2), it is not possible for infants to generalize

the vowel category on the basis of raw acoustic patterns. Moreover, we observed this skill in infants who are not yet producing canonical babble, and do not have the motor skills required to produce the target vowels in a controlled way. Unless they engage in frequent interactions with older babies, they will have limited exposure to infant speech. Infants displayed this skill in the visual habituation procedure, which relies on infants' spontaneous listening behavior with no conditioning or reinforcement to support or guide them. Thus without experimental training or reinforcement and with minimal exposure, infants were able to extrapolate beyond their immediate experience and adapt quickly to large acoustic shifts related to talker size.

The precise mechanisms contributing to perceptual skills measured in this study are not fully understood. Infants noticed the talker change in Experiment 1 and were clearly affected when there was a change in the talker set during habituation. These findings indicate that infants rely on processing mechanisms that involve encoding both talker and phonetic information in the speech signal. The results also show that pre-babbling infants have some phonetic categorization skills in place that they can harness immediately to assess the phonetic quality of their own vocal output. This suggests that young infants have the perceptual skills required to support vocal learning through imitation of stored target representations, consistent with the "high resource/imitation" view of infant speech development. Tracking vowel quality across talker differences related to physical attributes (size) may be a natural process involving innate auditory processes (Smith *et al*., 2005).

Although infants recognized the novel vowel in both experiments, listening time during habituation increased markedly when infant vowels were part of the habituation set, and longer habituation times were clearly linked to recognition of the novel vowel. This suggests that cognitive demands increased when infant speech was part of the multi-talker array during habituation. In other studies, infants have not adapted quickly and effectively to talker differences involving much smaller acoustic differences. The robust, rapid adjustments observed in this study may reflect a natural response to the sharp shift in acoustic variability or novelty tied to the infant signals, a perceptual bias favoring infant speech, or some combination of these factors. It is noteworthy that high pitch and expanded vowel space characterize both infant speech and infant-directed speech, which clearly draws infant attention and facilitates speech processing (Fernald & Kuhl, 1987). This raises the intriguing possibility that infant-directed speech may help prime the infant perceptual system for processing of their own vocalizations.

In summary, the present study provides the first evidence that pre-babbling infants can recognize infant-produced vowels as phonetically similar to adult and child productions. We now have our first glimpse at the perceptual resources that infants can access to process their own speech. The present study breaks new ground. Exploring how infants perceive speech with infant vocal properties opens a window into the interplay between speech perception and production in early language development.

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

Acknowledgements

This work was supported by a grant to L. P. from the Natural Sciences and Engineering Research Council of Canada. We thank Athena Vouloumanos, Susan Rvachew, Kris Onishi, Michael Tyler, and Janet Werker for their helpful comments on a previous version of this manuscript.

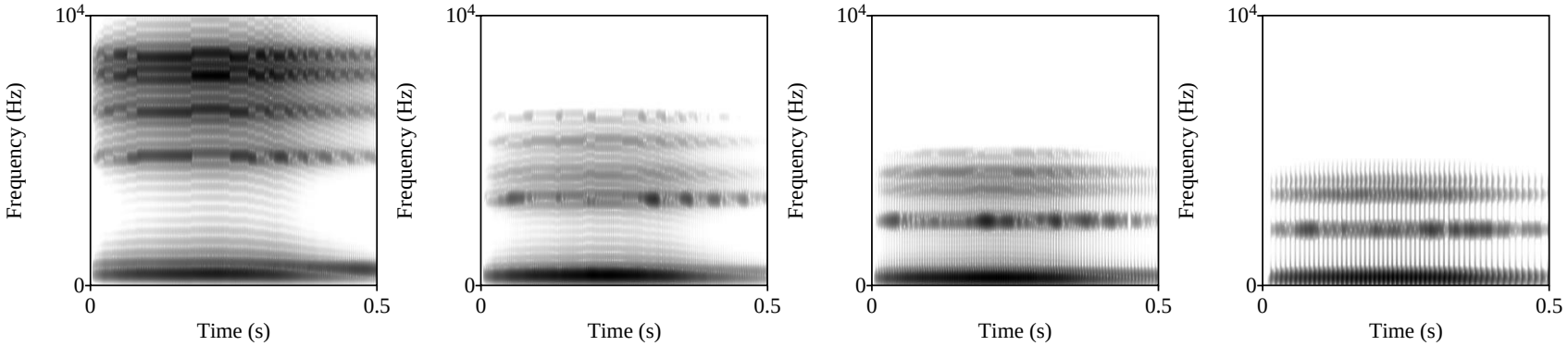


Fig. 1. Spectrograms of vowel /i/ (“ee”) with the vocal properties of (from the left) a 6-month-old infant, an 8-year-old child, an adult female and an adult male.

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who’s Talking Now? Infants’ Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

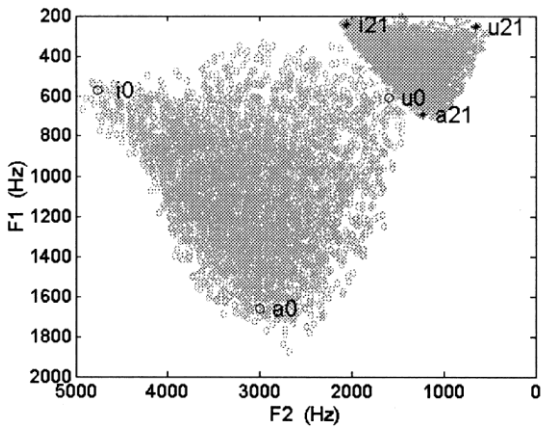
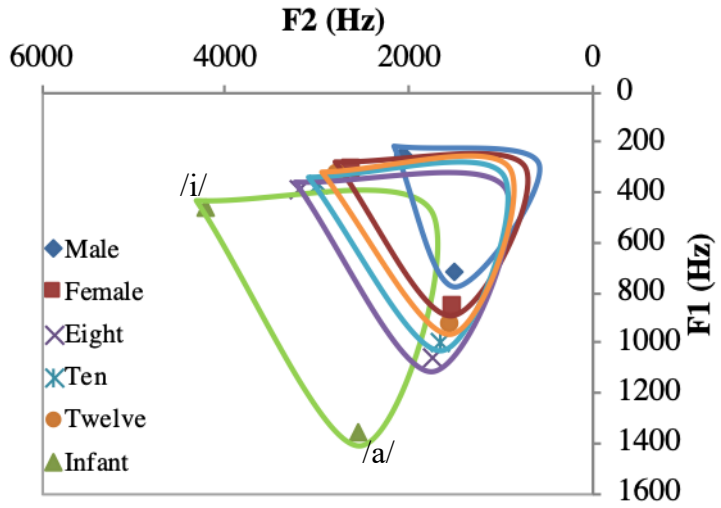


Fig. 2. A simulation of the maximal acoustic vowel spaces (F1/F2 Hz) corresponding to an adult male and an infant. The first formant frequency (F1) is plotted on the x-axis and second formant (F2) on the y-axis. The corner vowels /i u a/ are represented by circles and labeled "i0, a0, u0" for the infant, and represented by stars and labeled "i21, a21, u21" for the adult (borrowed with permission from Ménard *et al*., 2002). As the simulation illustrates, infant vowels span a large acoustic space, which overlaps only partially with the adult and child acoustic space (see also Figure 3).

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571



	Pre-test	**Habituation** Talkers: Male (m), female (f_1, f_2), child (c_8, c_{10} , c_{12}) Number of trials varies	**Test** Talker: Infant (i_1, i_2) 4 trials	Post-test
Experiment 1	Music	i_m ic_8 if_2 i_m ic_{12} if_2 … until criterion	ii_1 ai_2 ai_2 ii_2 OR ai_2 ii_2 ii_2 ai_2	Music
	Music	a_m ac_8 af_1 a_m ac_{12} af_2 … until criterion	ii_1 ai_2 ai_2 ii_2 OR ai_2 ii_2 ii_2 ai_2	Music

		Talkers: Female (f_1, f_2), child (c_8, c_{10} , c_{12}), infant (i) Number of trials varies	Talker: Male (m) 4 trials	
Experiment 2	Music	ii_1 ic_8 if_1 ii_2 ic_{12} if_2 … until criterion	im am am im OR am im im am	Music

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

Fig. 3. (top) Plot of F1/F2 frequencies for individual tokens of /i/ and /a/ in Hz for each speaker type. (bottom) vowel stimuli presented in each stage (Pre-test, Habituation, Test, Post-test) for Experiments 1 and 2.

PostPrint – Polka, L. Masapollo, M. & Menard, L. (2014) Who's Talking Now? Infants' Perception of Vowels with Infant Vocal Properties, *Psychological Science*, *25*(7), 1448-1456. http://dx.doi.org/10.1177/0956797614533571

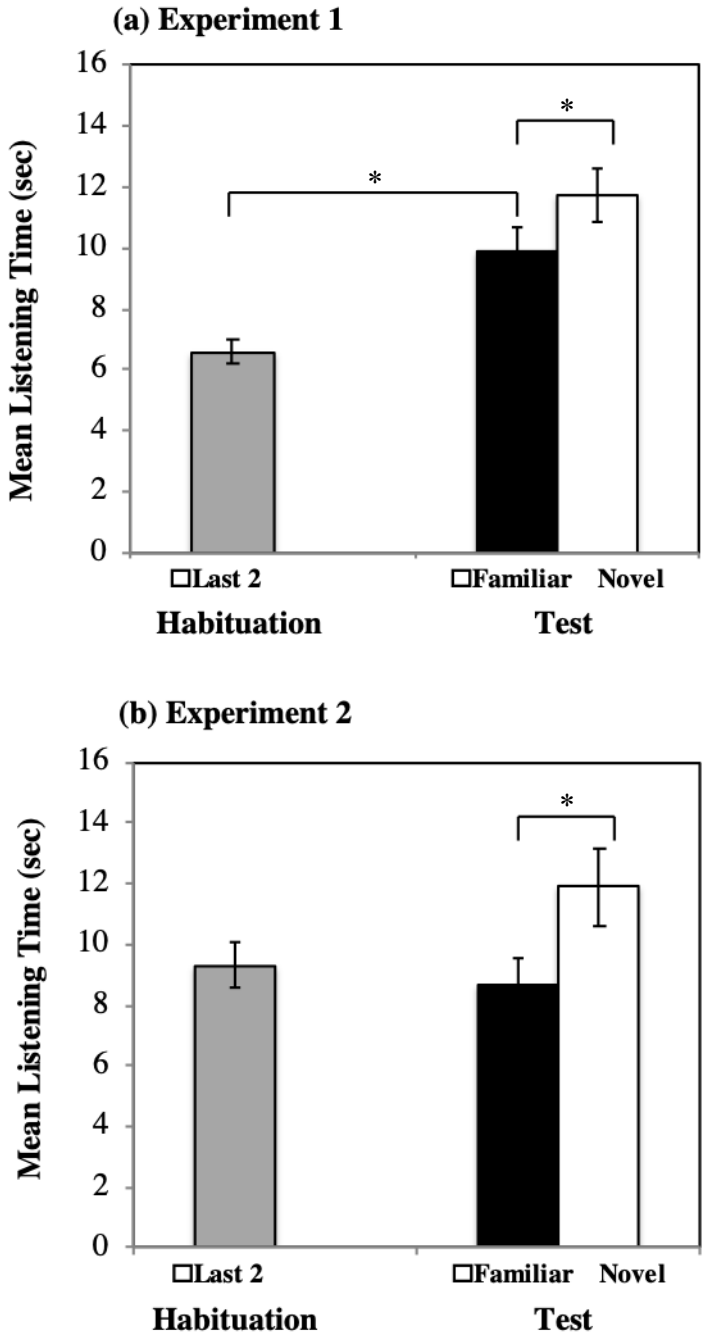


Fig. 4. Mean listening times (sec) to the last two habituation trials and to the novel and familiar test trials in Experiment 1 (a) and Experiment 2 (b). Error bars = std. errors, single asterisk = $p < .05$.

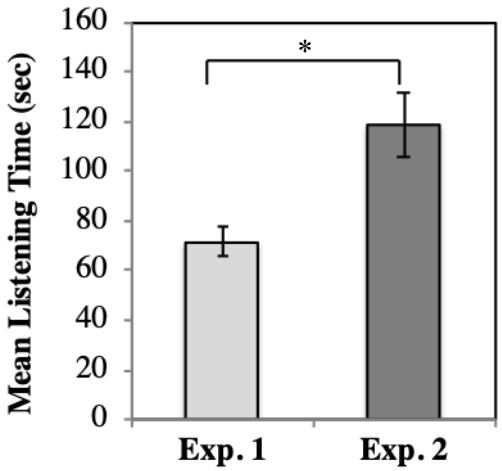


Fig. 5. Mean listening times (sec) during habituation in Experiment 1 (/i/ habituated group) and Experiment 2. Error bars = std. errors, single asterisk = $p < .05$.

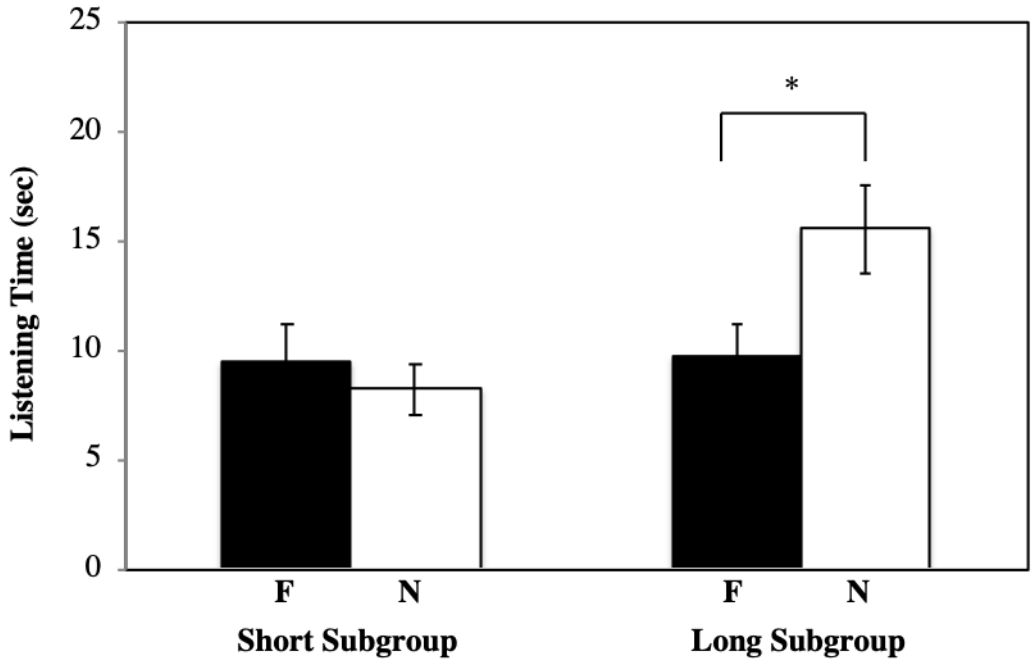


Fig. 6. Mean listening times (sec) during familiar and novel test trials for infants in the habituation subgroups of Experiment 2. Error bars = std. errors, single asterisks = $p < .05$.

Supplemental Material

Stimuli

Vowel stimuli that conform to the articulatory and acoustic properties of an infant vocal mechanism (vocal tract geometry and vocal fold vibration) were synthesized using the Variable Linear Articulatory Model (VLAM). VLAM integrates the vocal tract growth data currently available (Goldstein 1980) into a previous model already existing for the adult (Maeda 1979; Maeda 1990). The latter is based on a statistical analysis of 519 mid-sagittal cineradiographic images of a French speaker uttering ten sentences (Bothorel *et al*., 1986). A principal components analysis, guided by knowledge of the physiology of the articulators, revealed that seven articulatory parameters (P_i, $i \in \{1 \ldots 7\}$) could account for 88% of the variance of the tongue contours (Boë *et al*., 1995). The parameters included labial protrusion, labial aperture, tongue tip position, tongue body position, tongue dorsum position, jaw height, and larynx height. These parameters are directly interpretable in terms of functionally organized articulatory blocks. Each parameter is adjustable at a value in the range of ±3.5 standard deviations around the mean values for this articulator in the cineradiographic images.

VLAM integrates non-uniform vocal tract growth, in the longitudinal dimension, by two scaling factors: one for the oral cavity and another for the pharyngeal cavity, the zone in-between being interpolated. The values of the factors, from 0.3 to 1.2, were calibrated year-by-year and month-by-month based on Goldstein's (1980) length data. The anatomical measurements of vocal tracts generated by the model are consistent with MRI data, and the acoustic targets are in the range of the mean values of +/- 1 standard error reported for vowels from 3-year old to adult speakers in Hillenbrand (1995). The target scaled f0 values follow those given by Beck (1996). Thus, VLAM is constrained to generate only speech that can realistically be produced by a human vocal tract (see Boe, 1999, for a complete review of VLAM).

For each synthesized vowel, a mid-sagittal contour is obtained. The conversion models from the two-dimensional mid-sagittal function to the three-dimensional area function and the transfer function are then applied (Badin & Fant, 1984). The poles of the transfer function are excited through a 5-formant cascade synthesis system (Feng 1983), by a pulse train generated by a source according to the LF model (Fant *et al*., 1985).The source parameters (glottal symmetry quotient, and open quotient) are equal to 0.8, and 0.7, respectively, and remain unchanged for all the growth stages. The resulting signal is sampled at 22,050 Hz. The f0 contours and intensity envelopes were extracted from natural productions of isolated vowels uttered by an adult male.

In the Ménard *et al*. (2004) corpus, VLAM was set to generate five-formant synthesized vowels for speakers ranging from infancy to adulthood. The acoustic targets for the vowels of each speaker type were introduced by manipulating certain anatomical dimensions of the respective vocal tracts. The longitudinal dimension of the vocal tract was modified by varying the ratio between the oral and pharyngeal cavities. The following vocal tract lengths were implemented: 7.70 cm (infant), 11.91 cm (8-year-old), 12.65 cm (10-year-old), 13.52 cm (12-year-old), 15.36 cm (adult female), and 17.45 cm (adult male). Tongue length was calculated to be proportional to palate length. Vocal tract shape was determined using data from Goldstein (1980). Different talkers for each speaker type were then generated by manipulating fundamental frequency values using the following f0 values for each speaker type: 360 and 450 Hz (infant), 270 Hz (8-year-old), 337 Hz (10-year-old), 207 Hz (12-year-old-speaker), 210 and 240 Hz (adult female),

and 110 Hz (adult male). Acoustic values for all of the vowel tokens across the different talker types and audio files of each stimulus are shown in PolkaTableS1 below.

Talker	Vowel	f0	F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
Infant (1)	/i/	450	459	4220	5106	6412	8500	61	131	188	468	130
Infant (2)	/i/	360	459	4220	5106	6412	8500	61	131	188	468	130
Child (8-yr)	/i/	270	387	3213	3756	5173	6574	57	60	239	309	110
Child (10-yr)	/i/	337	362	2997	3571	4919	6104	56	56	224	288	110
Child (12-yr)	/i/	207	338	2778	3384	4644	5654	54	52	206	264	116
Adult female (1)	/i/	210	313	2628	3115	3704	5222	57	33	151	192	115
Adult female (2)	/i/	240	313	2628	3115	3704	5222	57	33	151	192	115
Adult male	/i/	110	247	2062	2951	3832	4359	58	29	58	146	229
Infant (1)	/a/	450	1359	2563	5106	6412	8500	61	131	188	468	130
Infant (2)	/a/	360	1359	2563	5106	6412	8500	61	131	188	468	130
Child (8-yr)	/a/	270	1062	1753	3756	5173	6574	57	60	239	309	110
Child (10-yr)	/a/	337	999	1647	3571	4919	6104	56	56	224	288	110
Child (12-yr)	/a/	207	933	1545	3384	4644	5654	54	52	206	264	116
Adult female (1)	/a/	210	856	1658	3115	3704	5222	57	33	151	192	115
Adult female (2)	/a/	240	856	1658	3115	3704	5222	57	33	151	192	115
Adult male	/a/	110	721	1512	2951	3832	4359	58	29	58	146	229

PolkaTableS1 Fundamental frequency f0 (Hz), mean center frequencies (Hz), and bandwidths (Hz) of the first five formants (F1, F2, F3, F4, F5) of the /i/ and /a/ vowels for the 8 talkers simulated for this study.

References:

Beck, J.M. (1996). Organic variation of the vocal apparatus. In W. J. Hardcastle & J. Laver (Eds.), Handbook of phonetic sciences, Cambridge, England: Blackwell, pp. 256-297.

Boe, L-J., Schwartz, J-L., & Valle, N. (1995). The prediction of vowel systems: Perceptual contrast and stability. In: Keller, E. (ed.) Fundamentals of speech synthesis and speech recognition. John Wiley, London, 185-213.

Bothorel, A., Simon, P., Wioland, F., and Zerling, J. P. (1986). Cinéradiographie des Voyelles et Consonnes du Français [Cineradiographic study of French vowels and consonants], Institut de Phonétique de Strasbourg, Strasbourg, France.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four- parameter model of glottal floqw.Royal Institute of Technology, Speech Transmission Laboratory – Quarterly Progress and Status Report, 4, 1-13.

Feng, G. (1983). Vers unc synthese par la method de poles et des zeros [Toward a synthesis method using poles and zeros]. Proceedings of the Journees d'Etude sur la Parole, Groupe Francophone de la Communication Parlee, 155-157.

Goldstein, U.G. (1980). An articulatory model for the vocal tract of growing children.Thesis of Doctor of Science, MIT, Cambridge, MA.

Maeda, S. (1979). An articulatory model of the tongue based on a statistical analysis. Journal of the Acoustical Society of America, 65, S22.

Maeda, S. (1990). Copensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.L. Hardcastle & A. Marshal (Eds.), Speech production and speech modeling (131-149). Dodrecht, The Netherlands: Kluwer Academic.

Ménard, L., Schwartz, J-L. & Boe, L.-J. (2004). The role of vocal tract morphology in speech development:

Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. Journal of Speech, Language, and Hearing Research, 47, 1059- 1080.