BUCLD 36 Proceedings
To be published in 2012 by Cascadilla Press
Rights forms signed by all authors

Factors Affecting Talker Recognition in Preschoolers and Adults

Sofia R. Jiménez and Sarah C. Creel

1. Introduction

Alongside learning how words sound in their language, children are learning about how *people* sound. How do they sort out what acoustic variability is relevant to identifying words vs. identifying talkers? Though long overlooked, talker variability is receiving increasing attention for its role in phonological development (Rost & McMurray, 2009, 2010) and sociolinguistic development (Kinzler, Dupoux, & Spelke, 2007).

Earlier studies (Bartholomeus, 1973; Mann, Diamond, & Carey, 1979) suggested that preschoolers encode new voices poorly relative to adults, distinguishing only gross differences like gender and accent. However, the data are somewhat problematic to interpret. Studies showing worse child talker recognition have used atypical voices (Jerger, Spence, & Rollins, 2002), have not tested both children and adults, have not equated voice exposure between children and adults (Bartholomeus), or have used tasks subject to working memory interference (Mann et al.). Jerger et al., for instance, found good performance by preschool children in recognizing familiar cartoon voices, but cartoon voices tend to be acoustically-distinctive relative to the normal human range of voice variability. A recent voice-learning study showing moderately good child performance in learning voices (Moher, Feigenson, & Halberda, 2010) used only one recording of each voice throughout learning and testing, making it unclear whether children overlearned a single utterance per talker, or actually extracted voice-specific properties. Additionally, Moher et al. imposed a learning criterion, which resulted in discarding a number of participants in each study. Eliminating poorer-performing participants may have led to the appearance of better performance than was actually the case.

The current study adapted a word-learning paradigm to investigate preschoolers' and adults' voice-learning abilities, but used *different* utterances during learning and testing. Thus, to succeed, listeners had to extract voice-specific characteristics. Voice exposure was controlled precisely, and amount of

* Sarah C. Creel and Sofia R. Jiménez, University of California San Diego; s2jimene@ucsd.edu, creel@cogsci.ucsd.edu. Research was supported by a Hellman Foundation Fellowship and an NSF CAREER Award (BCS-1057080) to SCC.

exposure was identical for children and adults. Working memory demands were minimized by providing training trials, rather than asking listeners to hold a voice in short-term memory for comparison to a later voice (as in Mann et al.).

2. Experiment 1

In this experiment, children and adults were exposed to voice-character pairings. Voices were drawn from a set of ten (described below). Characters were novel cartoons (Figure 1).

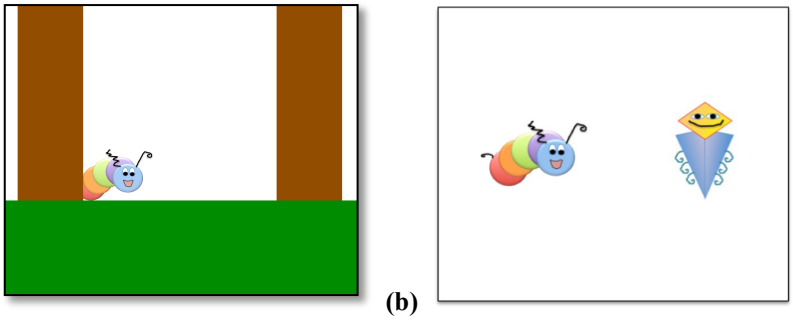
2. 1. Method

Participants. Children (*n* = 26, 3-6 years) from local day cares and preschools took part. Adults (*n* = 10) were undergraduates recruited from the UCSD human participant pool; they received course credit for participation. All participants spoke English, but not all of them were monolingual.

Stimuli.

Pictures. Two brightly-colored animate figures were created in PowerPoint (Figure 1). They were designed to be engaging and distinguishable. Each figure was saved as a .jpg file and resized to fit in a 200 x 200 pixel square for experimental presentation.

Selection of talkers. We chose Experiment 1 talkers based on formant-frequency (vocal-tract) differences. We recorded ten young, female, same-dialect talkers. Each talker produced: English vowels in the frame "Say the word h_d now" (where the blank represents a vowel); four training utterances (e.g., *Look at me jumping! Whee!*); four test sentences, which were distinct from the training utterances (e.g. *Can you find me?*); and several sentences for other studies. For Experiment 1, we selected talkers with distinct vocal tracts by choosing the two talkers whose vowel spaces, measured in the "h_d" utterances,



(a) **(b)**

Figure 1. On learning trials (a), a creature appeared on screen and spoke one of four phrases, such as "See where I am, behind the tree?" On testing trials (b), both creatures appeared and children heard one of the four test phrases, such as "Point at me!", from one talker.

differed most (i.e., had the most distant point vowels in Euclidean vowel (logF1-logF2) space). Point vowels are plotted in Figure 2a, with fundamental frequency (f0) characteristics in Figure 2b.

Procedure. Training and testing blocks alternated throughout. The first training block was 16 trials long, and the second and third were each 8 trials long. Test blocks 1-3 (following Training 1-3, respectively) were each 8 trials long. On each training trial, one of two cartoon creatures appeared onscreen (Figure 1a) and "spoke" with one of the two voices. Each creature spoke four different training phrases in each training block. Assignments of voices to pictures was counterbalanced across children. In *testing* trials, both creatures appeared side by side (Figure 1b). One creature spoke, and children were asked to point to it. In each test block, each creature spoke four test phrases, which were all different from the training phrases. Each creature occurred equally often as the target, and equally often on each side of the screen. Eye tracking data were obtained from a subset of participants (8 adults and 11 children) to gauge whether a more implicit measure might be more sensitive to recognition than pointing accuracy.

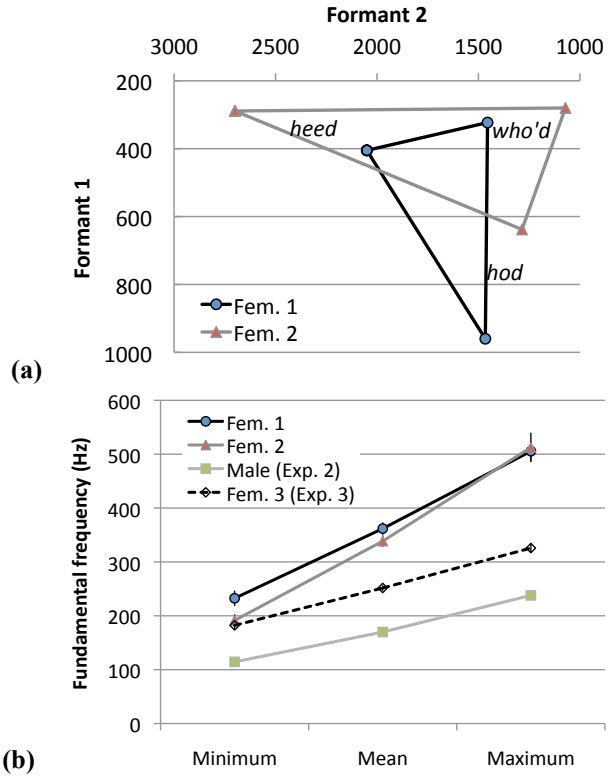


Figure 2. (a) Vowel spaces of the two talkers used in Experiment 1. (b) Pitch characteristics of all talkers, with standard errors (very small).

2.2. Results

Accuracy. Children's pointing accuracy (Figure 3) was 60.7% overall (*SD* = 17.5%). Adults reached 90% accuracy (*SD* = 10.6%). While children exceeded chance ($t(25) = 3.08$, $p = .005$), they were less accurate than adults ($t(34) = 4.82$, $p < .0001$), who approached ceiling (90% ± 11%; greater than chance, $t(9) = 11.35$, $p < .0001$). Children and adults differed numerically but not statistically in the first test block, and significantly in the second and third test blocks.

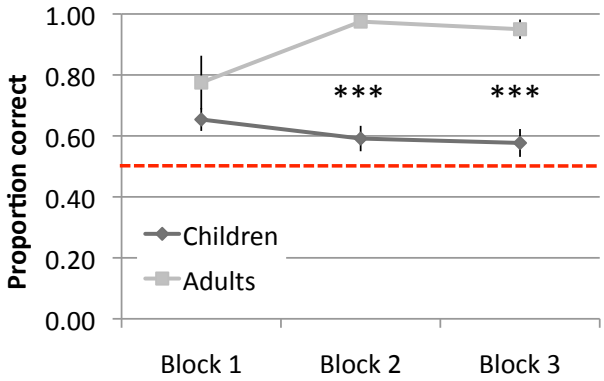


Figure 3. Accuracy of talker identification (± standard errors) over blocks in Experiment 1. Dashed line indicates chance performance. ***$p < .0001$

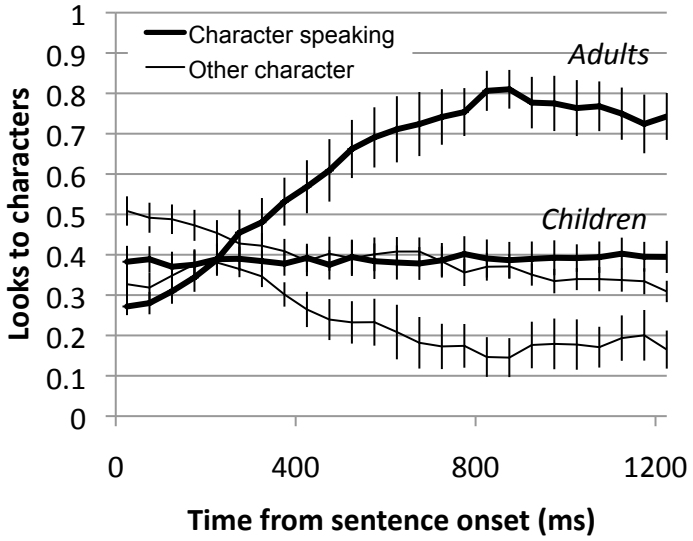


Figure 4. Looks to characters in Experiment 1 (11 children, 8 adults).

Eye tracking. Children's looks to targets (Figure 4) were not distinguishable from chance. Adults' looks to the correct character began to exceed chance between 400-450 milliseconds (ms) into the utterance ($t(7) = 2.59$, $p = .04$), suggesting fairly rapid identification of talkers.

2.3. Discussion

Children were above chance, but far below adult accuracy levels at mapping different-formant voices to characters. Eye tracking data provided no indication that children's implicit recognition abilities exceeded their pointing accuracy. These data suggest children have trouble distinguishing similar voices. However, it is also possible that children are simply uninterested in voice differences (or the experiment). To exclude this possibility, Experiment 2 paired a male voice with each Experiment 1 voice—a highly-salient difference. If children in Experiment 1 were uninterested in voice differences, then children in Experiment 2 should still show only moderate accuracy. However, if they were having difficulty distinguishing between the voices, then children in Experiment 2 should be highly accurate since voices are even more discriminable.

3. Experiment 2

3.1. Method

Participants. Children ($n = 16$, 3-6 years) from the same pool as in Experiment 1 took part.

Stimuli. The visual stimuli were the same as in Experiment 1. The female talkers were also the same (heard by 8 children each), but a new male talker (heard by all 16 children) was recorded.

Procedure. This was the same as in Experiment 1, except that instead of hearing two female voices, each child heard one of the two female talkers from Experiment 1, paired with a new male talker. The new male talker differed from the females in mean, minimum, and maximum pitch (p's $< .05$; see Figure 2b) and in formant frequencies.

3.2. Results

Accuracy. Children approached ceiling accuracy for different-gender voices ($M = 92\%$, $SD = 16\%$; $p < .0001$; Figure 5). That is, children easily mapped two acoustically-dissimilar voices to characters in the same paradigm where children in Experiment 1 had difficulty. This suggests the paradigm itself provided minimal difficulty.

Eye tracking. Children's looks to the character who was speaking (Figure 6) increased as the test sentence elapsed, exceeding chance at 350-400 ms ($t(15) = 3.01$, $p = .009$). This suggests that they rapidly recognized the characters' identities when acoustic cues to the characters were markedly different.

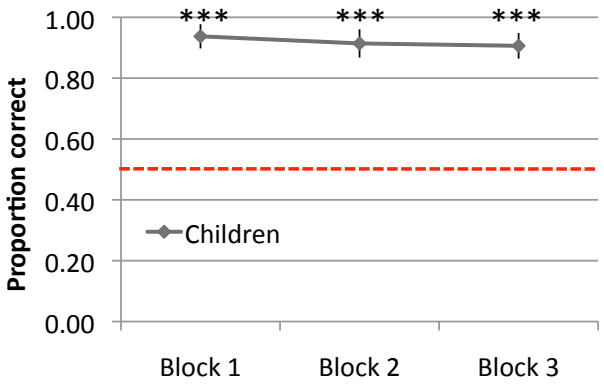


Figure 5. Accuracy of talker identification (±std. err.) over blocks in Experiment 2. Dashed line indicates chance performance. ***$p < .0001$

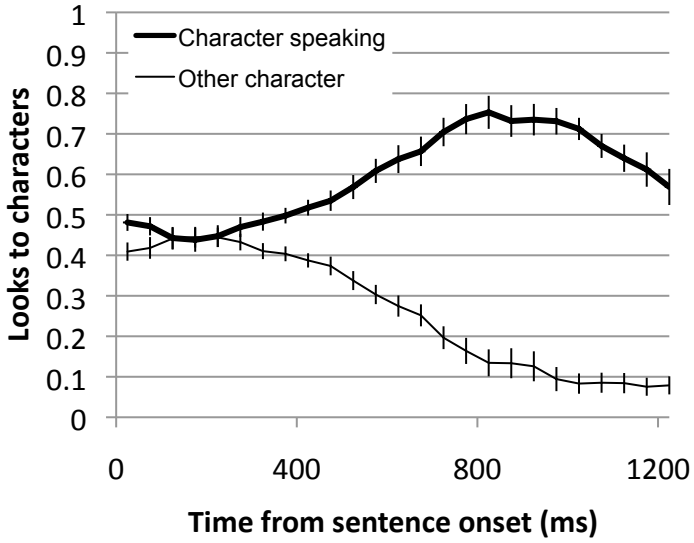


Figure 6. Looks to characters in Experiment 2.

3.3. Discussion

These data suggest that children map voices to characters more reliably when the voices are acoustically—and, interestingly, socially—distinct. Another possibility is that children can distinguish some acoustically- and socially-similar voices, but the acoustic differences in Experiment 1 (formants) were not salient to children. To address this possibility, Experiment 3 examined whether

children find pitch differences more salient. The talkers used differed strongly in pitch characteristics: each of the Experiment 1 talkers was paired with a third "low-prosody" voice from the original set of recordings.

4. Experiment 3

4.1. Method

Participants. Children (n = 24, 3-6 years) and adults (n = 16) from the same groups as in previous experiments took part.

Stimuli. Visual stimuli were the same as before. A new talker, female 3 (drawn from the original recordings of 10 females), was substituted in for the male talker from Experiment 2, so that each child heard female 3 and *one* of the Experiment 1 talkers. We selected female 3 by having 9 adults in the lab rate all ten original in terms of prosody. They gave each speaker a score from 1-7; the low-prosody speaker received an average score of 2.3 whereas our two original speakers received an average score of 5.9. Measurements of fundamental frequency (Figure 2b) bore out these impressionistic ratings: the two original female voices both differed from the new voice (female 3) in minimum-pitch, maximum-pitch, and pitch-range (maximum: minimum ratio; all $p < .05$).

Procedure. This was the same as in Experiments 1-2. All listeners were eye tracked; one child's eye tracking data file was inadvertently deleted, so this participant is not included in eye tracking analyses.

4.2. Results

Accuracy. Children's accuracy (54.3% ± 14.4% Figure 7) did not exceed chance ($t(23) = 1.44$, $p = .16$). While this was not significantly worse than children's performance in Experiment 1 ($t(48) = 1.38$, $p = .17$), it was lower than adults' accuracy ($t(38) = 11.15$, $p < .0001$). Adults were at ceiling (97.9% ± 6.1%; exceeding chance, $t(15) = 30.55$, $p < .0001$). Unlike Experiment 1, adults' advantage was already evident in the first test block.

Eye tracking. As in Experiment 1, children's low accuracy was also evident in their visual fixations (Figure 8). Adults rapidly fixated the correct picture, reaching significance by 250-300 ms ($t(15) = 2.63$, $p = .02$).

4.3. Discussion

Child and adult listeners learned to map prosodically-dissimilar female voices to characters. As in Experiment 1, adults far outperformed children in the task, and adult but not child visual fixation data showed rapid, accurate visual fixations to the character than was speaking. This suggests that large fundamental frequency differences, though quite salient to adults, do not facilitate voice-character mapping by children.

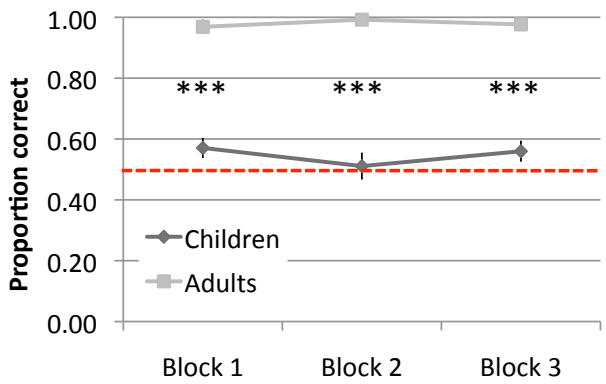


Figure 7. Accuracy of talker identification (±std. err.) over blocks in Experiment 3. Dashed line indicates chance performance. ***$p < .0001$

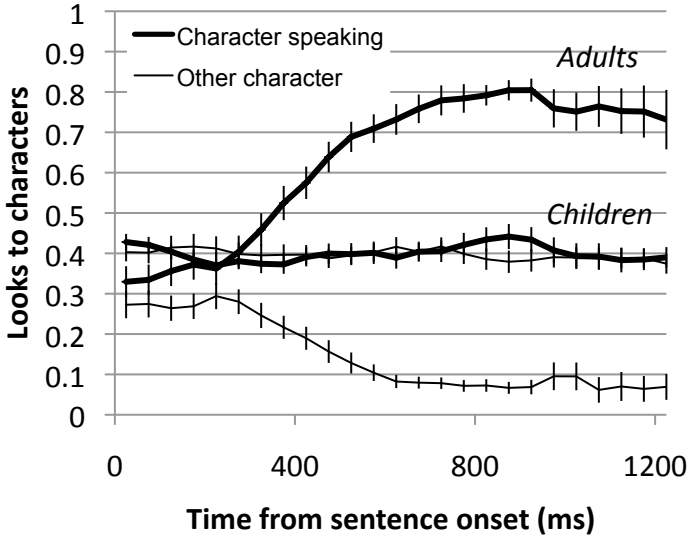


Figure 8. Looks to characters in Experiment 3 (16 adults, 23 children).

5. General Discussion

We asked whether preschool children could learn to map voices to characters, and how these abilities compared to adults. Children performed best when voices were the most acoustically distinct (male vs. female; Experiment 2). Children were above chance but far less accurate than adults in mapping

voices that differed spectrally (Experiment 1). Children did not exceed chance for voices that differed in pitch (Experiment 3), though adults were at ceiling.

5.1. Performance differences across age

The experiments conducted suggest that children have more difficulty than adults in representing different voice qualities. Why might this be the case? One possibility is that children attend to voice differences that distinguish major social categories like gender, but do not attend to voices otherwise. Other authors (Hirschfeld & Gelman, 1998; Kinzler et al., 2007) have found that children use familiar vs. unfamiliar language from a speaker to make social inferences, and Creel (in press) found that preschoolers interpreted spoken sentences differently depending on the speaker's gender. These studies suggest that children use acoustic cues to social categories to make social and linguistic decisions.

Of course, social similarity and acoustic similarity tend to covary, meaning that acoustic similarity alone could be driving the effects. Importantly, our cartoon-character design ensured that all voice pairs used were mapped to the same fictional creatures, which contained few visual cues to gender or other social characteristics. That is, children could not use *visual* information to distinguish gender in Experiment 2, meaning that elevated performance in that experiment was based on acoustic information.

It is of interest that, despite children's weaker performance on similar voices, a subset of children performed quite well: 9 of 50 children in Experiments 1 and 3 reached 75% correct or higher. This is unlikely to be due to chance, as *no* children received scores this extreme in the negative direction (25% or fewer correct). What distinguished these children? Gender appeared not to be a factor, as roughly half (4 of 9) were female. Seven of 9 were bilingual, indicating a potential bilingual benefit, though this relationship is not holding up in ongoing work in our lab. Age is a factor: the high-scoring children were on average 4.9 years old (SD = .9), while the average age of the rest was 4.4 years (SD = .9). The correlation of accuracy and age was significant (r = .41, $t(47)$ = 3.08, p = .003; note that df = 47 because one child's age was missing). This suggests that accuracy in mapping similar voices improved with age.

At a broader level, these results match with other studies of young children's uptake of paralinguistic information. For instance, Morton and Trehub (2001) found that children even older than the current sample had difficulty identifying a talker's vocal affect when verbal content conflicted (such as "I made an A at school" spoken in a sad voice). Specifically, they seemed to rely almost exclusively on verbal content. Quam and Swingley (in press) found that children could not reliably identify vocal emotion at all—even though they could identify visual cues to emotion—until age 4-5. At an even younger age (17 months), children have trouble mapping words to referents if the words are extremely similar (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002). The commonality across these different tasks is that children

seem to have difficulty mapping similar acoustic or acoustic-phonetic information onto concepts (object; character; emotional state).

5.2. Strategic differences

Another factor distinguishing preschoolers and adults is task meta-awareness. Did adults succeed simply based on greater understanding of the experimental situation? Experiment 1 suggested a role of meta-awareness: adults' performance did not differ significantly from children until after the first test. This raised the possibility that adults did not encode voices any better than children until they became aware, via the first block of test trials, what they were expected to learn. However, adults in Experiment 3 exceeded children even in the first test, suggesting that the voice differences were highly salient to them without directing attention to it. Children showed no such pattern of improvement over blocks in any experiment—if anything, they showed the opposite pattern, becoming slightly less accurate during the course of the experiment. On the one hand, this might mean that children had too little meta-awareness of the task to intuit what sound properties they were supposed to be attending to. On the other hand, it seems unlikely that, if children did possess sufficient meta-awareness, they might not have been able to use it effectively if they were unable to discern the sound characteristics to which they needed to attend.

5.3. Implicit vs. explicit measures of voice-character mapping

A final question addressed here was whether eye tracking, an implicit measure of voice recognition, might be more sensitive than accuracy, an explicit measure of voice recognition. This does not seem to be the case: the two measures yielded remarkably consistent results. When listeners had difficulty explicitly *identifying* characters' voices (children in Experiments 1 and 3), they showed no patterns of *looking* more to the correct character than to the incorrect character. Inversely, when listeners easily identified characters' voices (adults in Experiments 1 and 3, children in Experiment 2), looks to the correct character over the incorrect character were rapid and robust. This is interesting in that most investigations of voice perception in even-younger children have used implicit measures (Houston & Jusczyk, 2001; Johnson, Westrek, Nazzi, & Cutler, 2011). Though those studies found infant sensitivity to voice (particularly Johnson et al.), we did not. Of course, we asked listeners to learn *mappings* of voices to characters. As discussed above, mapping may be a more difficult problem than simple discrimination (e.g. Stager & Werker, 1997). We suggest that, rather than difference between implicit and explicit awareness, differences in infant and older-child performance in voice recognition reflect task differences. This generates the prediction that infants who are capable of dishabituating to a voice change (as in Johnson et al., 2011) would not be able to learn to map those voices to characters, though this has not been tested.

5.4. Remaining questions

Data thus far suggest that children's difficulty in voice-character mapping is in perceptually distinguishing the voices. Nonetheless, other possibilities remain. For instance, children may detect acoustic differences, but have difficulty pinpointing the acoustic characteristics which define voices across changes in verbal content. If so, children should distinguish female voices more readily when trained and tested on a single utterance. Another question concerns the relevance of social category distinctions: is gender the only attribute that children encode strongly, or are other social categories, such as age, also readily encoded? Ongoing work in our lab is examining these questions.

6. Conclusion

This study extends previous work on child voice recognition by directly comparing children and adults in a naturalistic learning task with controlled exposure, and in requiring generalization to novel utterances. More generally, this work emphasizes the multifaceted learning problem that the speech signal presents to children.

References

Bartholomeus, Bonnie (1973). Voice identification by nursery school children. *Canadian Journal of Psychology*, *27*(4), 464-72.

Creel, Sarah C. (In press). Preschoolers' use of talker information in on-line comprehension. *Child Development.*

Hirschfeld, Lawrence A., & Gelman, S. A. (1997). What young children think about the relationship between language variation and social difference. *Cognitive Development*, *12*, 213-238.

Houston, Derek M., & Jusczyk, Peter W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570-1582. doi:10.10371/0096-1523.26.5

Kinzler, Katherine D., Dupoux, Emanuel, & Spelke, Elizabeth S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577-80.

Johnson, Elizabeth K., Westrek, Ellen, Nazzi, Thierry, & Cutler, Anne (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002-1011.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology*, *27*, 153-165.

Moher, Mariko, Feigenson, Lisa, & Halberda, Justin. (2010). A one-to-one bias and fast mapping support preschoolers' learning about faces and voices. *Cognitive Science*, *34*(5), 719-751.

Morton, J. Bruce, & Trehub, Sandra E. (2001). Children's understanding of emotion in speech. *Child Development*, *72*(3), 834-843.

Quam, Carolyn, & Swingley, Daniel (In press). Development in children's interpretation of pitch cues to emotions. *Child Development.*

Rost, Gwyneth C, & McMurray, Bob. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*, 339-349.

Rost, Gwyneth C., & McMurray, Bob. (2010). Finding the Signal by Adding Noise: The Role of Noncontrastive Phonetic Variability in Early Word Learning. *Infancy*, *15*(6), 608-635.

Spence, Melanie J., Rollins, Pamela R., & Jerger, Susan. (2002). Children's recognition of cartoon voices. *Journal of Speech, Language, and Hearing Research*, *45*(1), 214-222.