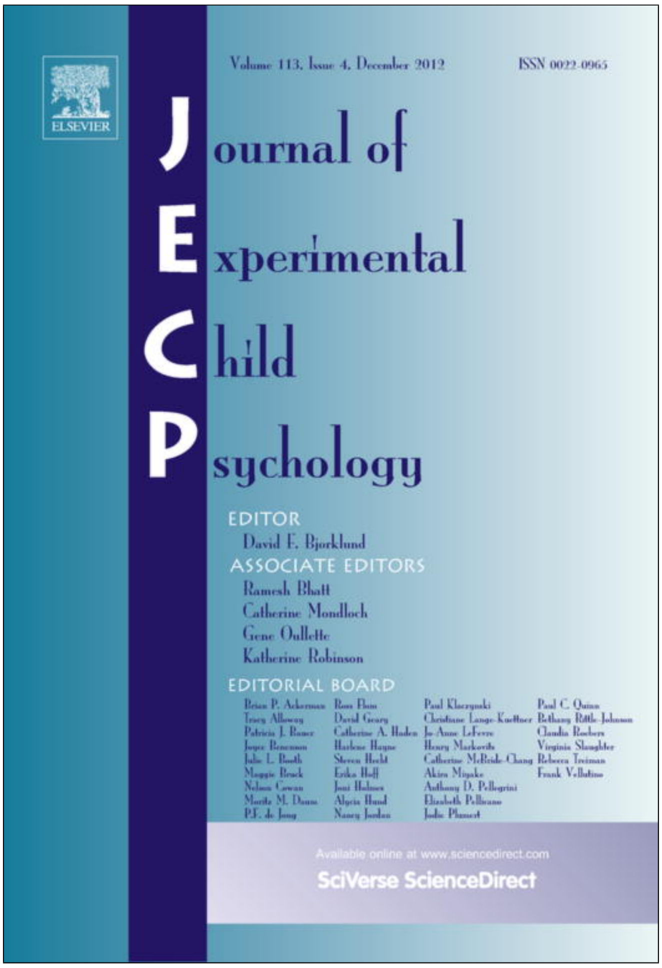


Volume 113, Issue 4, December 2012 ISSN 0022-0965

JECP

Journal of Experimental Child Psychology

Journal of Experimental Child Psychology 113 (2012) 487–509

Contents lists available at SciVerse ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Differences in talker recognition by preschoolers and adults

Sarah C. Creel *, Sofia R. Jimenez

Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 19 December 2011
Revised 7 June 2012
Available online 5 September 2012

Keywords:
Talker variability
Perceptual learning
Auditory learning
Preschool
Language development
Voices

ABSTRACT

Talker variability in speech influences language processing from infancy through adulthood and is inextricably embedded in the very cues that identify speech sounds. Yet little is known about developmental changes in the processing of talker information. On one account, children have not yet learned to separate speech sound variability from talker-varying cues in speech, making them more sensitive than adults to talker variation. A different account is that children are less developed than adults at recognizing speech sounds *and* at recognizing talkers, and development involves protracted tuning of both recognition systems. The current research presented preschoolers and adults ($N = 180$) with voices linked to two distinct cartoon characters. After exposure, participants heard each talker and selected which character was speaking. Consistent with the protracted tuning hypothesis, children were much less accurate than adults when talkers were matched on age, gender, and dialect (Experiments 1–3), even when prosody differed (Experiment 5). Children were highly accurate when voices differed in gender (Experiment 2) or age (mother vs. daughter; Experiment 6), suggesting that greater acoustic dissimilarity facilitated encoding. Implications for speech sound processing are discussed, as are the roles of language knowledge and the nature of talker perceptual space in talker encoding.

Introduction

How do children sort out speech sound variability amid rampant talker variation? Decades of research have explored how children learn the speech sounds of their native language. Recent accounts describe speech sound acquisition as a perceptual learning problem; very young children acquire

* Corresponding author. Fax: +1 858 534 1128.
E-mail address: creel@cogsci.ucsd.edu (S.C. Creel).

0022-0965/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jecp.2012.07.007

language-specific knowledge by accumulating statistics on their speech input (e.g., Feldman, Griffiths, & Morgan, 2009; Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007). The input eventually statistically clusters into language-specific speech sounds via unsupervised or partially supervised learning (Yeung & Werker, 2009). However, we know much less about what happens to talker variability during this perceptual learning process. Infants appear to be highly sensitive to talker variation, to the point where they cannot generalize a familiarized word form to a new talker (Houston & Jusczyk, 2000), suggesting that attention to talker variability is detrimental to speech sound recognition. Nonetheless, adults still show residual sensitivity to talker variation in speech perception (Creel, Aslin, & Tanenhaus, 2008; Goldinger, 1996) and can recognize talkers with high accuracy (Van Lancker, Kreiman, & Emmorey, 1985).

What happens between infancy and adulthood that allows adults to perceive speech regardless of who is talking yet still recognize voices, an important skill in a social world (e.g., Kinzler, Dupoux, & Spelke, 2007)? Understanding the answer to this question is crucial for understanding how listeners represent speech itself. On the one hand, some evidence hints at a slow decline in talker sensitivity as children learn to *filter out* (or adjust for) (McMurray & Jongman, 2011) talker information during speech perception. This would predict that children might be *better* than adults at recognizing talkers, with some residual ability in adults. On the other hand, children may start out being sensitive to all types of sound contrasts and then slowly begin to tune different recognition processes—some for recognizing words, some for recognizing talkers. This *protracted tuning* hypothesis would predict a slow increase over development in ability to recognize talkers. The evidence for either pattern, reviewed below, is somewhat equivocal. In the next section, we describe the close relationship between talker variation and speech sound variation. Following that, we explore the evidence for talker recognition at different ages.

Talker variability and speech perception

Sorting out speech sound variability (e.g., what distinguishes a *peach* from a *beach*) from talker variability (e.g., what distinguishes Mom vs. Aunt Gertrude saying *beach*) is an extremely complex problem because speech sound and talker variability are highly intertwined in the speech signal. That is, nearly all speech sounds are realized differently by different talkers, even within a language. Talkers differ in their vowel formant frequencies (Peterson & Barney, 1952; see also Hillenbrand, Getty, Clark, & Wheeler, 1995), voice onset times for stop consonants (Lisker & Abramson, 1964; see also Allen, Miller, & DeSteno, 2003), and fricatives (McMurray & Jongman, 2011; Newman, Clouse, & Burnham, 2001; see also Jongman, Wayland, & Wong, 2000). McMurray and Jongman (2011) examined 24 cues to fricative identity and found that *all 24* cues were affected by talker variability. The implications of this variability are twofold: Speech variability should influence talker identification, and talker variability should influence speech sound identification.

Speech sound knowledge does in fact influence talker identification. Pioneering work by Bricker and Pruzansky (1966) found that talkers varied in identifiability depending on what speech sound was being produced, suggesting that there is no such thing as invariant "voice quality," even though this misconception persists to the present day. Further studies have indicated that familiarity with a language facilitates talker identification in that language (Goggin, Thompson, Strube, & Simental, 1991; Winters, Levi, & Pisoni, 2008). This appears to be driven by phonological knowledge in particular; Perrachione, Chiao, and Wong (2010) found that listeners recognized voices better within their own dialect than in a different dialect, controlling for verbal content. Furthermore, Perrachione, Del Tufo, and Gabrieli (2011) found a correlation between degree of phonological impairment in dyslexic listeners to degree of difficulty in recognizing voices.

Just as speech sound differences influence talker identification, talker differences influence speech sound identification. Identification of speech sounds can be influenced by the talker's perceived gender (Johnson, Strand, & D'Imperio, 1999), race (Staum Casasanto, 2008), or geographical origin (Niedzielski, 1999). Talker information can aid listeners in distinguishing phonologically similar words (Creel et al., 2008). Listeners are better at recognizing speech in noise from highly familiar talkers than from unfamiliar talkers (Nygaard, Sommers, & Pisoni, 1994). Listeners identify speech sounds and words more rapidly when the talker remains constant rather than changing from trial to trial (e.g., Magnuson

& Nusbaum, 2007; Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992). Finally, adults learn to distinguish non-native speech sound contrasts more accurately when they hear the sounds produced by a variety of talkers (Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991).

Taken together, these studies suggest that there is a strong relationship between the processing of speech sound variability and talker variability in the speech signal, such that better recognition of each facilitates processing of the other. This facilitative relationship suggests that talker recognition should improve along with speech sound recognition across development, consistent with the protracted tuning hypothesis. However, the evidence for such an improvement is sparse and contradictory.

Talker sensitivity across development

How does talker sensitivity change across development? Some work suggests that very young infants are exquisitely sensitive to individual differences between talkers but lose the sensitivity late in the first year of life. Neonates respond differentially to their mothers' voices versus strangers' voices after only prenatal exposure (DeCasper & Fifer, 1980; Kisilevsky et al., 2003). At 7 months of age, infants in a dishabituation paradigm detected a change from one female voice to another similar female voice, but only in the infants' ambient language (Johnson, Westrek, Nazzi, & Cutler, 2011; see related adult work by Perrachione et al., 2010, 2011). Infants' sensitivity to talkers is accompanied by difficulty in recognizing word forms over talker changes. Houston and Jusczyk (2000), Houston and Jusczyk (2003) found that at 7.5 months of age, infants familiarized with words from one talker did not recognize those words in the fluent speech of another dissimilar talker. By 10.5 months, infants did recognize the words from another talker (see similar but slightly later effects in accent variability by Best, Tyler, Gooding, Orlando, & Quann, 2009; Schmale, Cristià, Seidl, & Johnson, 2010; Schmale & Seidl, 2009). At 7 or 8 months of age, infants recognize familiarized words-in-noise from the mother's voice, but not from a stranger's voice (Barker & Newman, 2004). The waning of these talker specificity effects roughly parallels infants' loss of non-native speech sound contrasts during the first year of life (Werker & Tees, 1984), raising the possibility that, over development, children may "lose" talker contrasts as well as non-native speech sounds. Conversely, presenting words in *multiple* voices (Rost & McMurray, 2009; Rost & McMurray, 2010) aids 14-month-olds in learning similar words (see related work by Singh, 2008, on variability in vocal affect in younger infants). Rost and McMurray (2010) hypothesized that hearing multiple talkers told infants what cues *not* to link to word meaning; given multiple talkers but invariant speech sounds, infants accordingly weighted talker-related cues downward, or filtered them out, rather than attending to all cues. This filtering process might become automatized over the course of development, leading to worse talker recognition later in development.

However, relatively few studies have examined talker sensitivity in older children, and only some of those have compared children directly with adults. Spence, Rollins, and Jerger (2002) tested 3- to 5-year-olds' recognition of familiar cartoon characters' voices in a six-alternative forced-choice (6AFC) picture selection task. Accuracy ranged from 61% at age 3 to 86% at age 5. Performance was comparable to Van Lancker and colleagues' (1985) study of adults' recognition for famous voices, where Van Lancker and colleagues analyzed accuracy for the particular talkers that each individual listener thought he or she would recognize and found 68% accuracy in a 6AFC task. However, the comparison between the results of Spence et al. (2002) and those of Van Lancker et al. (1985) is imperfect; cartoon voices (Spence et al.) tend to have more exaggerated sound characteristics than adult male voices (Van Lancker et al.), which may make cartoon voices easier to recognize. Moher, Feigenson, and Halberda (2010) took the approach of teaching voices and found that 4- and 5-year-olds, given feedback, could link female voices (each saying, "Can you touch my nose?") to pictures. Accuracy in a 2AFC task ranged from 63% to 77% across three experiments. Creel (in press) showed that preschool-aged children use talker information in the speech signal to constrain the referential domain of sentences for different-gender talker pairs but not for same-gender talker pairs. This indirectly suggests that children might not distinguish same-gender voices. However, adults in Creel's study also failed to use the same-gender talker differences in sentence processing, making it possible that both groups were simply failing to attend to more subtle talker differences.

Only two studies have directly compared talker recognition in adults and children. Bartholomeus (1973) tested preschoolers' recognition of their classmates' voices along with the children's teachers.

Children achieved 57% free naming accuracy, whereas the four teachers identified child talkers at 68% accuracy. Both age groups were better at recognizing children's *faces* (97% for children, 100% for adults) than children's voices. Yet talker exposure could not be controlled across children and adults; the adults, as teachers, may have had more balanced interactions with individual children, whereas children themselves might have interacted verbally with a more limited group of peers. In a larger cross-sectional study, Mann, Diamond, and Carey (1979) tested 6-year-olds through adults in a standard comparison talker discrimination task. All talkers were unfamiliar females, similar in age (25–45 years) and accent, and were asked to match the prosody and timing of a model recording. Participants heard a standard talker speak a phrase, followed by two to four test talkers. They needed to decide whether each test talker was the standard talker or a ''mystery'' talker. The 6-year-olds were at chance in distinguishing same-accent female talkers, with gradual improvement into adulthood. However, this task puts strong demands on working memory; listeners needed to hold the standard talker's voice in mind while listening to the test talkers. This raises the question of how much the results were due to working memory constraints rather than talker processing. Children might perform better in a task that gives them more robust memory representations—that is, one that familiarizes them with the voices prior to testing them.

In summary, although infant research is consistent with a filtering-out account of talker variability, studies of talker recognition in older children suggest a possible developmental *increase* in talker recognition, which is more consistent with protracted tuning. However, these studies are either small in scale or are open to alternative explanations.

The current research

Do children filter out talker variability from the speech signal early in life, or do they slowly improve at talker recognition via protracted tuning of talker representations? It is difficult to tell because very few studies have compared talker recognition abilities of children directly with those of adults (Bartholomeus, 1973; Mann et al., 1979), and the results are inconclusive. Some studies (e.g., Houston & Jusczyk, 2000) suggest an early developmental decrease in talker sensitivity, consistent with the filtering-out hypothesis. However, the two studies to compare children directly with adults at *recognizing* talkers suggest that there may be an increase in sensitivity to talkers, consistent with the protracted tuning hypothesis. One of those two studies (Bartholomeus, 1973) tested only four adults and could not control for amount of voice exposure. The other study (Mann et al., 1979) found a developmental increase in talker recognition, but the auditory standard comparison task likely taxed children's working memory capacity, meaning that improvement with age may have reflected gains in working memory rather than in talker recognition. Thus, it is still unclear whether talker recognition declines or improves over development.

The current study aimed to understand developmental changes in talker processing by testing preschoolers' versus adults' abilities to map a natural range of talker differences to individuals. The first three experiments examined children's and adults' abilities to distinguish two female voices differing in timbre (spectral quality), with slight variations on the paradigm in each experiment. Experiment 4 tested children with a more salient voice difference—male versus female voices. Experiment 5 tested children and adults on female talkers differing in fundamental frequency characteristics. Finally, Experiment 6 tested children's and adults' abilities to distinguish voices matched in gender but differing in age. Experiment 6 also assessed whether children would recognize two children's voices more accurately than two adults' voices.

Experiment 1

This experiment tested preschool-aged children's abilities to map same-age, same-gender, same-dialect voices to characters. Children and adults viewed an engaging learning paradigm modeled on word learning experiments. We used unfamiliar voices to control listeners' amount of exposure, allowing direct comparison of child and adult performance. Talkers varied naturally in a range of acoustic–phonetic characteristics. Rather than testing and training on a single utterance, each of two cartoon ''talkers'' spoke multiple sentences (Table 1) during talker training and spoke a different set of sentences during the test, requiring listeners to generalize talker recognition across utterances. With exactly the same exposure, what do children learn about talkers, and how do they compare with adults?

Table 1
Sentences used in all experiments.

Training sentences	Testing sentences
Look at me now! Watch this!	Do you see me?
Can you see me now? Yay!	Can you find me?
See where I am, behind the tree?	Where am I?
Watch me jumping! Whee!	Point at me!

Method

Participants

The participants were 26 children (12 girls and 14 boys, 3.0–6.3 years of age, $M = 4.7$ years, 1 age unreported) from local day care and preschool facilities, and 10 adults drawn from the UCSD experimental participation pool. The sample throughout was predominantly Caucasian, but African, East Asian, and South Asian ethnicities were also represented. An additional 2 children were excluded due to ambient construction noise during testing. Children received a small toy, and adults received course credit, for participation.

Stimuli

Two distinct cartoon creatures, also used in word learning studies (principal investigator's lab, in preparation), were used (Fig. 1). Color cartoon figures were created in PowerPoint software and were exported as image files. They were further edited in an image editing program to fit into a 200 × 200-pixel square for computer presentation. During the simply animated learning trials, cartoons underwent translational motion on the screen (up, down, left, and right) but did not otherwise change in facial expression, body posture, or size.

Talkers

Ten female college-aged California native talkers were recorded in a sound-attenuated chamber. They were instructed to speak all materials in a child-directed manner. Each talker recorded the experimental sentences listed in Table 1. The talkers also recorded additional materials that we used to calculate measures of talker similarity, including citation forms of English vowels in the frame "Say the word h_d now" (heed, hid, head, had, etc.) (e.g., Hillenbrand et al., 1995). The h_d vowels were used in an analysis in Praat 5.1.44 (Boersma & Weenink, 2010) to find the two talkers with the most dissimilar point vowels ("ee", "ah", and "oo")—a correlate of vocal tract properties—to obtain the most distinct vocal timbres. All remaining voice measurements were taken directly from the actual experimental stimuli in Table 1. The selected talkers, Fem1 and Fem2, had similar degrees of child-directed prosody (similar mean and maximum fundamental frequency [f0]; see Fig. 2A) but differed in their vowel triangles (F1 and F2; see Fig. 2B) and center of gravity (Fig. 2C). Higher formants contributed to the subjective percept that Fem1 sounded "higher" than Fem2, although mean f0 did not differ. (Note Fem1's similarity to child vowels later in Fig. 4B.) Additional voice measures are plotted in Appendix A, including shimmer, in which Fem2 exceeded Fem1 (see Appendix C for spectrograms of all voices used).

Procedure

Adults sat in a chair in front of the computer in the lab, and children sat in an unbuckled car seat in a quiet area of their school facility. Sounds were presented over Kidz Gear child-sized headphones (http://www.gearforkidz.com) for children and over Sennheiser Pro HD 280 headphones (http://www.sennheiserusa.com) for adults. The task alternated between training and testing. Training Block 1 (16 trials, 2.4 min) presented each talker 8 times. On each trial, a cartoon creature moved on-screen (Fig. 1B, left), paused, spoke a two-sentence passage, and then moved off-screen. Each creature spoke each passage equally often. Across children, each creature occurred equally with each voice.

Just before Test Block 1 (8 trials), the experimenter instructed children to point to whichever creature was talking. Each test trial displayed the two creatures side by side on a white background (Fig. 1B, right), and one of them spoke (e.g., "Which one is me?"). If children did not point immediately, the experimenter prompted them. Test trials proceeded at children's own pace.

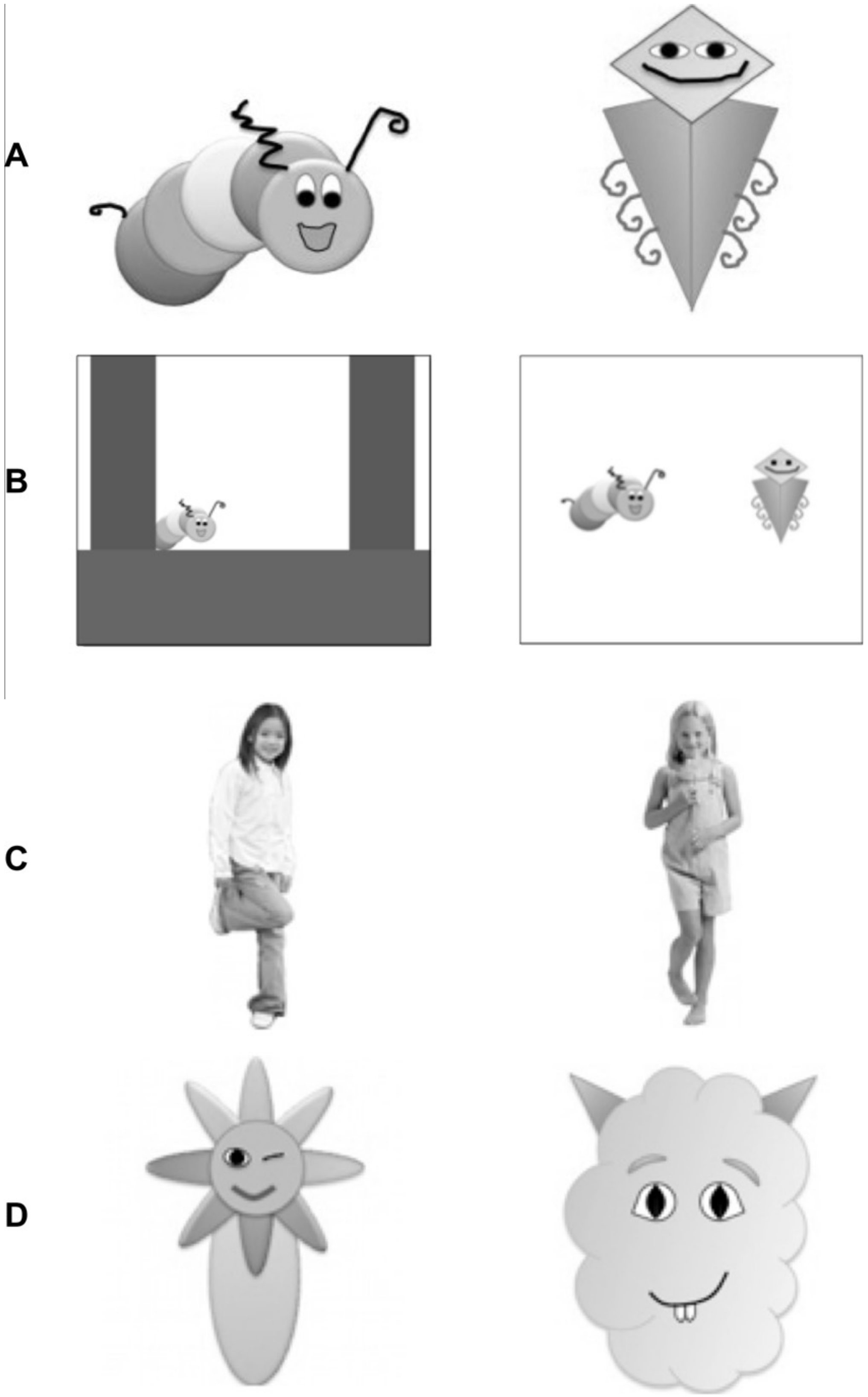


Fig. 1. (A) Cartoon characters used in all Experiments except Experiment 3. (B) Sample training, test trials. (C) Pictures used in Experiment 3. (D) Extra characters used in Experiment 6.

Following Test Block 1, there were 8 "refresher" training trials, then 8 more test trials, then 8 more training trials, and 8 final test trials. This design allowed us to assess effects of fatigue (worse performance later in the experiment) and increased exposure (better performance later). To maintain

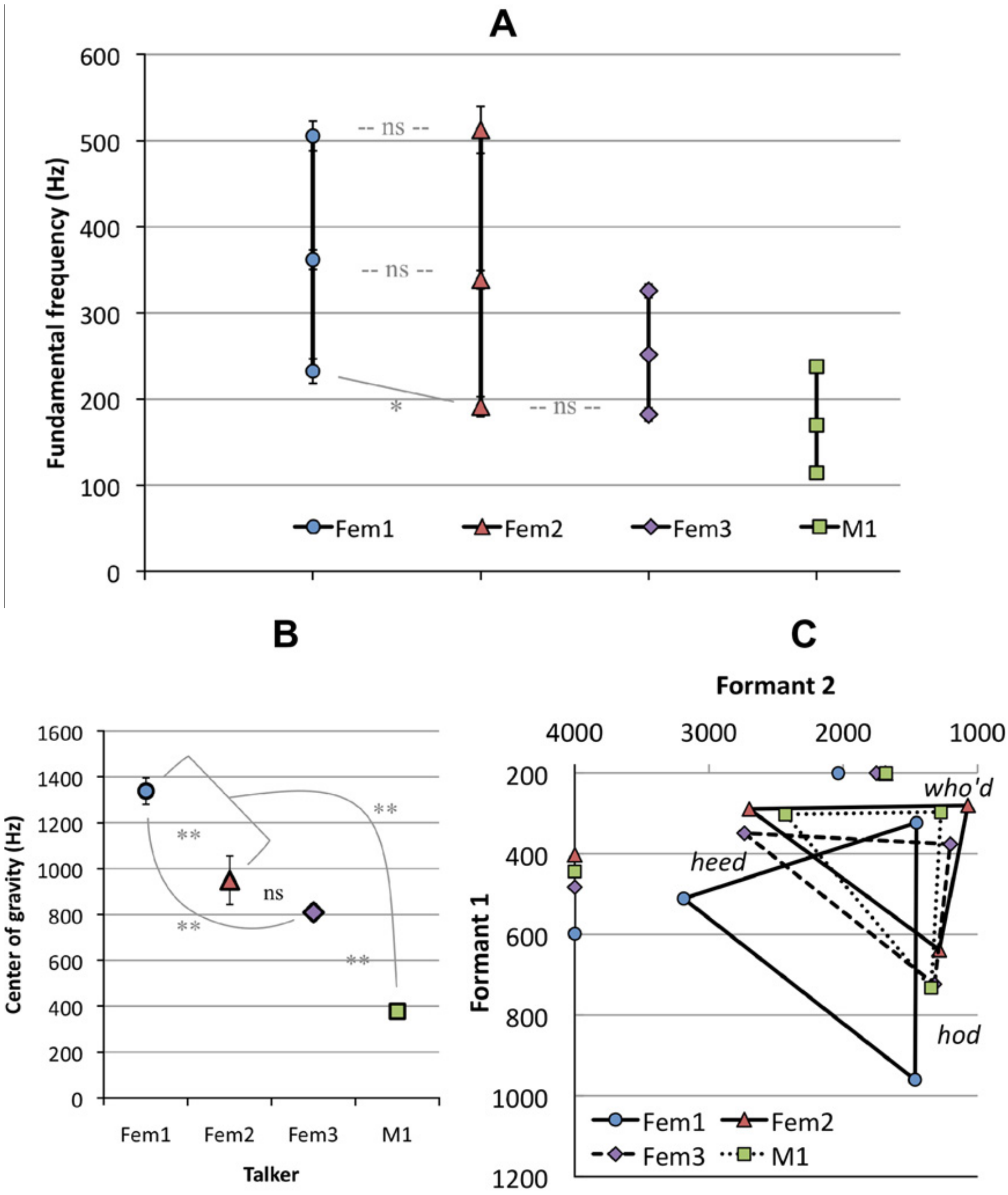


Fig. 2. Talkers from Experiments 1 to 5. Shown are measures of fundamental frequency (minimum, mean, and maximum; unmarked comparisons all significant, $p \leqslant .001$) (A), point vowels (B), and center of gravity (Hz) (C). In panels A and C, some standard errors are too small to see over point markers. In panel B, no error bars are shown because measurements were based on single observations. Points on axes show average F1 and F2 values. $^{*}p < .05$; $^{**}p < .01$.

interest, animated distracter sequences (moving pictures of appealing animals paired with clapping, cheering, baby giggling, or "Aaaaaah!" sounds) occurred after Training Trial 8 (two distracters) and after Test Block 1 (two distracters). Experimenters (and adult participants) recorded pointing responses by mouse-clicking the chosen picture. The entire experiment took 6 to 8 min to complete.

Results

Children achieved 60.7% (*SD* = 17.5) accuracy (Fig. 3), above chance but below adult performance (*M* = 90%, *SD* = 11). Responses were evaluated by a mixed logit model with mean-centered, standardized predictor variables, with age (child or adult) as a between-participants factor, block (1, 2, or 3) as a within-participants factor, and participant random intercept and slope (Participant × Block). Note that in this and following experiments, we also tried adding intercepts and slopes for phrase—the sentence spoken—to models. In no case were those random effects significant. The maximum correlation between fixed effects was −.523. The intercept term was significant (estimate = 1.20, *SE* = 0.18, $z = 6.51$, $p < .0001$), indicating that accuracy across all participants exceeded chance. An age effect (estimate = −1.14, *SE* = 0.21, $z = 5.42$, $p < .0001$) indicated higher adult accuracy. There was no block effect (estimate = 0.15, *SE* = 0.11, $z = 1.46$, $p = .14$), but there was an Age × Block interaction (estimate = −0.50, *SE* = 0.14, $z = 3.71$, $p = .0002$). Models were then computed at each age alone (maximum fixed effect correlations: adults, .134; children, .123). The Age × Block interaction apparently resulted from a marginal increase in adult performance (estimate = 0.97, *SE* = 0.52, $z = 1.85$, $p = .06$) across blocks and a nonsignificant decrease in child performance (estimate = −0.13, *SE* = 0.09, $z = -1.53$, $p = .13$) across blocks. Difference from chance performance was assessed by the significance of the intercept term of the adult model and the child model. Adults exceeded chance performance (estimate = 3.74, *SE* = 0.59, $z = 6.32$, $p < .0001$). Children, although less accurate than adults, also exceeded chance (estimate = 0.52, *SE* = 0.17, $z = 3.01$, $p = .003$).

Discussion

Children were above chance at recognizing talkers, but adults were better. This confirms earlier findings (Bartholomeus, 1973; Mann et al., 1979) that young children have more difficulty than adults in distinguishing highly similar talkers, but unlike Mann and colleagues' (1979) findings, our data suggest some capacity for similar-voice identification in preschoolers.

Perhaps children's difficulty is in *generalizing* talker-specific properties across several different utterances from the same talker. If children encoded individual sentences in great detail, they may have experienced difficulty in extracting talker-specific properties across the acoustic variability within a talker. To test this, Experiment 2 replicated Experiment 1 except that listeners were trained and tested on just one phrase (as in Moher et al., 2010). If children encode individual utterances in detail, they should improve at identifying talkers.

Experiment 2

Method

Participants

The participants were 24 new preschool-aged children (13 girls and 11 boys, 3.0–6.4 years of age, $M = 4.2$ years) from the same population as in Experiment 1.

Stimuli

The stimuli were the same as in Experiment 1.

Procedure

Each child learned on *one* testing phrase from Experiment 1. Across participants, each phrase was used equally often. Children were tested on the learned phrase and one other phrase to determine whether learning was phrase specific. The first test block used the original phrase only to assess learning before the novel phrase was presented in case its introduction created confusion.

Results

A mixed logit model with block as a within-participants factor and with participant slope and intercept terms was computed, with a maximum fixed-effects correlation of −.038. Children achieved 61.6% accuracy (*SD* = 19.1), exceeding chance (intercept estimate = 0.58, *SE* = 0.21, $z = 2.81$, $p = .005$). An additional question was whether hearing a new phrase led to a decrease in accuracy. To assess this, a second logit model was computed and assessed performance differences between the trained phrase and the new phrase in Blocks 2 and 3, where the novel phrase was introduced. Fixed effects were block (2 or 3) and phrase (old or new), and random effects of participant intercept and all slopes were included (maximum fixed effects correlation: −.148). Children were no more accurate on the learned

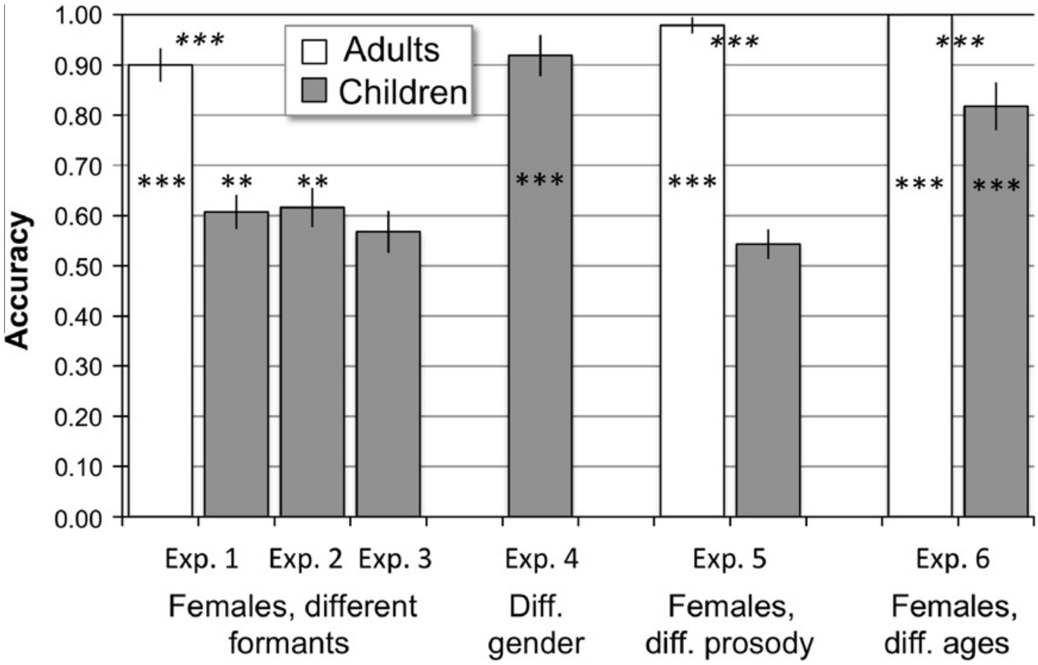


Fig. 3. Accuracy from Experiments 1 to 6. Italicized asterisks indicate child–adult participant comparisons, and the rest indicate difference from chance. $^{**}p < .01$; $^{**}p < .001$. Exp., experiment; diff., different.

phrase (59.9 ± 21.9%) than on the novel phrase (61.4 ± 23.4%) (estimate = −0.05, *SE* = 0.11, *z* = −0.50, *p* = .62).

Discussion

Children were no better at distinguishing talkers when trained and tested on a single phrase (current experiment) than on multiple phrases (Experiment 1). These results suggest that the locus of children's difficulty is not generalization across multiple utterances but instead discrimination of, or memory for, talker-related speech characteristics.

A remaining concern is ecological validity; children's cartoon viewing habits aside, perhaps it is unnatural or uninteresting to encode voice–character pairings when the characters are not human. Experiment 3 replicated Experiment 1 but with photographs of two female children instead of cartoon animals. If children map voices more readily to human referents, they should perform better in Experiment 3 than in Experiments 1 and 2.

Experiment 3

Method

Participants

The participants were 24 new preschool-aged children (11 girls and 13 boys, 3.4–6.0 years of age, *M* = 4.7 years, 6 ages unreported) from the same population as before.

Stimuli

Auditory stimuli were the same as in Experiment 1. Visual stimuli were human figures (Fig. 1C) selected from the Microsoft Office online clip art database (http://office.microsoft.com/en-us/images). Color images were edited to fit in a 200 × 200-pixel square like the cartoons. Figures differed in hair color, ethnicity (East Asian vs. White), clothing (color, length, and style), paraphernalia (one was holding a flower and one was not), and posture. Both were young females, making them plausible candidates for the voices. They were also close to the participants' age range and, thus, socially interesting. Note that although preschoolers are known to have better within-race face discrimination than other-race face discrimination (e.g., Pezdek, Blandon-Gitlin, & Moore, 2003), the difference in ethnicity

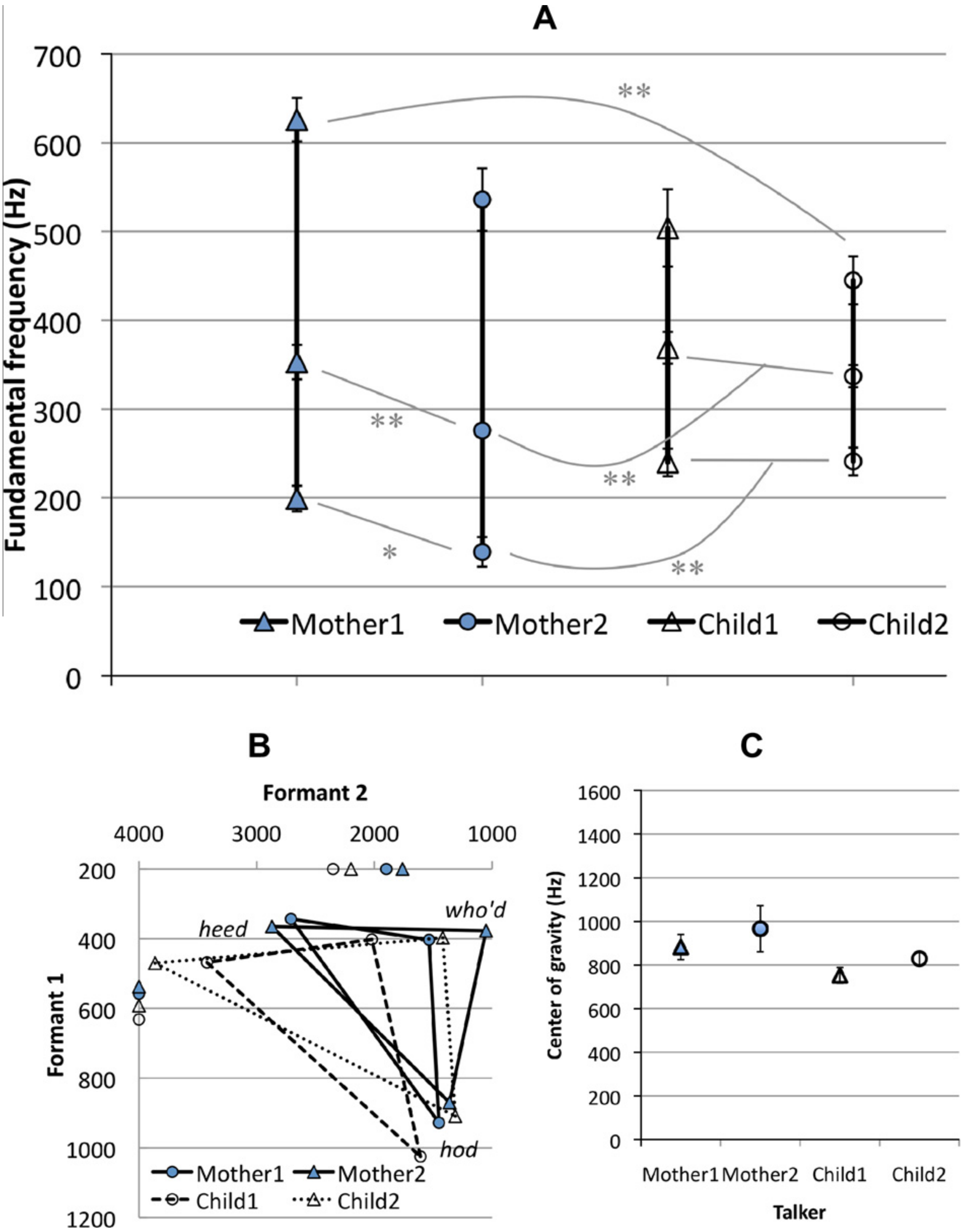


Fig. 4. Talkers from Experiment 6. Shown are measures of fundamental frequency (minimum, mean, and maximum) (A), point vowels (B), and center of gravity (Hz) (C). *$p < .05$; **$p < .01$.

between the two children should not have created learning difficulty because children were not asked to make distinctions *within* a less familiar ethnic category.

Procedure

The procedure was identical to that in Experiment 1.

Results

Overall performance was 56.8% correct (*SD* = 20.5). A logit model with block as a fixed effect and participant slope and intercept terms (maximum fixed effects correlation: −.210) showed that performance did not exceed chance (estimate = 0.26, *SE* = 0.17, *z* = 1.5, *p* = .13). A marginal block effect (estimate = −0.18, *SE* = 0.09, *z* = −1.95, *p* = .051) suggested a slight decrease in performance during the experiment.

To assess whether this accuracy level differed from children's performance in Experiments 1 and 2, new logit models were computed comparing Experiment 3 with each other experiment, with block and experiment as fixed effects and participant slope and intercept. Effects of experiment were nonsignificant when comparing Experiment 3 with Experiment 1 (maximum fixed effects correlation: −.033; estimate = −0.09, *SE* = 0.08, *z* = −1.04, *p* = .30) and when comparing Experiment 3 with Experiment 2 (maximum fixed effects correlation: −.125; estimate = −0.26, *SE* = 0.27, *z* = −0.96, *p* = .34). This suggests that performance was not significantly less accurate than in Experiments 1 and 2.

Discussion

Children were no more accurate in identifying two female voices when the voices' owners were visibly human (Fig. 2), supporting an acoustic–phonetic locus—not a visual one—for children's incomplete voice learning in Experiments 1 and 2. So far, our results suggest that voice–character mappings that adults learn easily are somewhat difficult for most children. Alternately, the task may simply be unnatural; children may be tuned to use visual cues to identify individuals rather than acoustic cues. If this were the case, they would be encoding the characters' appearances rather than their voices. T o assess this, we increased perceptual salience by using male–female talker pairs. If encoding talker information is simply unnatural, children should still show low accuracy. However, if the paradigm is easy but the talkers in Experiments 1 to 3 were too perceptually similar for children, children should be highly accurate.

Experiment 4

Method

Participants

The participants were 16 new preschool-aged children (10 girls and 6 boys, 3.2–5.7 years of age, *M* = 4.6 years) from the same population as before.

Stimuli

Visual figures were the cartoon characters used in Experiments 1 and 2. Each child heard Fem1 *or* Fem2 from Experiment 1 (8 children each) and new male talker M1 (all 16 children). M1 was lower in f0 and center of gravity (Fig. 2) and had longer syllables, less shimmer, more jitter, and greater harmonicity (Appendix A) than Fem1 and Fem2.

Procedure

The procedure was the same as in Experiment 1.

Results

Children approached ceiling performance at 92 ± 16% accuracy. A logistic model with block as a fixed effect and with random participant slope and intercept terms was computed (maximum fixed effect correlation: −.545). The intercept term was significant (estimate = 4.60, *SE* = 5.80, *z* = 5.77, *p* < .0001), reflecting children's well-above-chance accuracy. An effect of block (estimate = −0.79, *SE* = 0.29, *z* = −2.76, *p* = .006) suggested decreasing performance across blocks. An additional model with block and experiment effects and with participant slopes and intercepts compared performance with Experiment 1, which was identical in every respect except for the talkers used (maximum fixed effect correlation: .602). Performance exceeded that in Experiment 1 (estimate = 2.84, *SE* = 0.49, *z* = 5.76, *p* < .0001).

Discussion

Children adeptly map different-gender voices to cartoon characters (Fig. 2), suggesting that the talker mapping task itself presents little difficulty. This means that children's performance in Experiments 1 to 3 may actually reflect less adult-like encoding of talker information. It is interesting that few studies have looked at children's or infants' processing of adult male voices; most have used female talkers exclusively (including Johnson et al., 2011; Mann et al., 1979 [although they used a male vs. female voice pair during training with good results]; Moher et al., 2010). Work on infant face recognition suggests that discrimination of male *faces* shows up substantially later in development than discrimination of female faces (Ramsey, Langlois, & Marti, 2005). The current experiment suggests that distinguishing *between*-gender talkers is easy for children but does not assess *within*-gender discrimination of male talkers. A later trajectory for male voice learning than for female voice learning is an interesting topic for future research.

It is still somewhat puzzling that children are so poor at distinguishing female talkers if the task itself is so easy. A remaining possibility is that Fem1 and Fem2 did not differ in ways that matter to children such as prosody. Mann et al. (1979) asked talkers to *match* prosody to a single model utterance and found chance performance in 6-year-olds. On the other hand, Moher et al. (2010) explicitly instructed talkers to use *different* prosody and found slightly higher accuracy than we did in some experiments. Furthermore, even young infants distinguish languages based on prosodic characteristics (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Mehler et al., 1988). These studies suggest that prosody may be a particularly salient characteristic to young children. However, other studies suggest that preschoolers have difficulty in *interpreting* prosody (Berman, Chambers, & Graham, 2010; Morton & Trehub, 2001; Nelson & Russell, 2011; Quam & Swingley, 2012). Experiment 5 explored this issue and replicated the female talker experiments with new talker pairs using talkers differing substantially in prosodic characteristics. If prosody differences are more salient to children than spectral differences, children should be highly accurate.

Experiment 5

This experiment asked children to learn pairs of talkers with differing prosody. The talkers chosen for Experiment 1—Fem1 and Fem2—had very similar fundamental frequency characteristics, but other talkers we recorded had flatter prosody. This was due to differences in talkers' executions of our request for child-directed speech. The new talker, Fem3, had much lower minimum, mean, and maximum f0 (Fig. 2A) than Fem1 and Fem2.

Method

Participants

The participants were 24 new preschool-aged children (7 girls and 17 boys, 3.0–5.9 years of age, $M = 4.1$ years, 1 age unreported) and 16 adults from the same populations as in previous experiments.

Stimuli

We asked 9 adults in the lab to rate the prosody of the 10 original talkers on a scale of 1 (*low*) to 7 (*high*). Of the 10 original talkers, 2 were tied for the lowest prosody score at 2.3 ($SD = 0.78$), and we selected one of them. Fem1 and Fem2 received prosody ratings of 5.6 ± 0.46 and 6.3 ± 0.67, respectively, which were the highest and third highest average ratings. The f0 measurements (Fig. 2A) confirmed these impressionistic ratings. Each child heard either Fem1 or Fem2 ($n = 12$ children each) contrasted with Fem3 (all children). Visual figures were the cartoon characters used in Experiments 1, 2, and 4.

Procedure

The procedure was the same as that used in preceding experiments.

Results

Adults performed at ceiling (97.9 ± 6.1%), verifying that talkers were distinct to adult listeners, whereas children were less accurate (54.3 ± 14.4%) (Fig. 2). A logit model was computed with block

(1 or 2) as a within-participants fixed effect, age as a between-participants fixed effect, and participant intercept and slope random effects (maximum fixed effects correlation: −.858, between block and the interaction term). Age was significant (estimate = −1.93, *SE* = 0.22, $z = -8.97$, $p < .0001$), indicating higher accuracy for adults than for children. To compute performance relative to chance, individual models with block and participant slope and intercept were computed for each age (maximum fixed effects correlations: adults, −.313; children, .023). Adults exceeded chance (estimate = 7.90, *SE* = 1.90, $z = 4.15$, $p < .0001$). Children's accuracy, although quite low, approached significance (estimate = 0.21, *SE* = 0.12, $z = 1.74$, $p = .08$).

Discussion

Prosody differences did not increase children's success in recognizing talkers. This is a bit surprising given infants' sensitivity to prosody (Jusczyk et al., 1993; Mehler et al., 1988). However, our child participants' lack of prosody sensitivity makes more sense in light of studies where young children have difficulty in *interpreting* prosody (Cutler & Swinney, 1987; Morton & Trehub, 2001; Nelson & Russell, 2011; Quam & Swingley, 2012). Although infants can detect acoustic changes in prosody in a dishabituation paradigm, it may be much harder to map such acoustic patterns to different affective states or to individuals.

The emerging story is that preschoolers are less adept than adults at mapping fine-grained voice distinctions to individuals, even after a fair amount of exposure. Of course, the case where children approach ceiling is a gender difference, which is salient both socially and acoustically. Can children can make *other* socially relevant voice distinctions accompanied by acoustic differences such as age? To assess this, we conducted a final experiment where children learned female mother–daughter voice pairs contrasting on age (4–5 years vs. 35–40 years). Children also heard pairs of same-age talkers. For half of the children the talkers were their own age, and for the other half of the children the same-age talkers were of parental age. If children process vocal cues to age, they should easily distinguish mothers and daughters.

This design also permitted a test of own-age voice discrimination: Are children better at identifying voices of their age mates? Preschoolers may have been at a disadvantage in earlier experiments because the voices were those of 20-year-olds rather than of children their own age (see, e.g., Melinder, Gredebäck, Westerlund, & Nelson, 2010, for own-age effects in face perception). Members of our population—children in preschool or day care situations—have exposure to many more children's voices than adults' voices, which might provide them with more perceptual learning of child voices than of adult voices. Furthermore, individuating peers' voices may be more socially relevant to children than individuating adults' voices. This might provide children with more motivation to distinguish between child voices. If children excel at distinguishing own-age voices, they should learn the child talker pair better than the adult talker pair. However, if children simply make fewer voice distinctions than adults, own-age talkers should be as difficult to learn as other-age talkers.

Experiment 6

Method

Participants

The participants were 24 new preschool-aged children (15 girls and 9 boys, 3.1–5.3 years of age, $M = 4.2$ years) and 16 adults from the same populations as before.

Stimuli

Two mother–daughter pairs (4.5, 5.5, 35, and 41 years of age) were recorded in quiet rooms. Mothers had lower minimum f0 (Fig. 4A), and lower F1 and F2 (Fig. 4B), than daughters, but there were no difference in mean or maximum f0 (Fig. 4A) or center of gravity (Fig. 4C). Daughters showed less jitter (fine-scale f0 variability) and shimmer (fine-scale amplitude variability), and greater harmonicity, than mothers (Appendix B). Two more cartoon characters (Fig. 1D) were added to allow testing of multiple voices per participant.

Procedure

The procedure was modified as follows. To maximize data collection, each participant was trained on two voices (8 trials) and tested (8 trials) and then was trained on one old voice and one new voice (16 trials) and tested again (8 trials). The first set of training trials was intended to be 16 trials, but a programming error caused only 8 trials to run. Although the logit model did not indicate block effects, we inspected the data to see whether children performed worse after the shortened first block. Children were slightly *more* accurate after the short first training block, whereas adults were more accurate after the second (longer) training block, suggesting that the shortened exposure block did not deflate children's overall accuracy relative to previous studies. Due to fewer training trials, the experiment was slightly shorter than previous experiments (5–7 min rather than 6–8 min). Each participant heard a same-age pair (either mothers or daughters) and a different-age pair (one mother–daughter set). Order of pairs and voice–character assignments were distributed across participants. To ask whether talker learning is related to language ability, children completed a receptive vocabulary measure, the Peabody Picture Vocabulary Test–Fourth Edition (PPVT-IV; see Dunn & Dunn, 2007).

Results

Children did well on the age contrast pair, but adults still performed better (Figs. 3 and 5). A logit model of accuracy was computed with age match between talkers (different ages or same age) and block (1 or 2) as within-participants factors and with participant age (child or adult) and same-age pair (mother voice pair or child voice pair) as between-participants factors. Participant intercepts and slopes were included as random effects.

Due to adults' perfect performance in the different-aged talker condition, each effect was tested using change in log likelihood from the maximal random effects model rather than the Wald *z* test. The Wald *z* statistic, reported in previous experiments, becomes unreliable as one or more cells approach 0 or 1 (Agresti, 2002; Levy, Fedorenko, Breen, & Gibson, 2012, note 13). In these cases, change in log likelihood is the more reliable measure. To assess significance of individual effects, pairs of models were carefully constructed to contrast the full model with one that was the full model minus the effect of interest. Using this test, age was significant, $\chi^2(1) = 20.17$, $p < .0001$, reflecting higher adult accuracy. Age match was also significant, $\chi^2(1) = 61.61$, $p < .0001$, with higher accuracy for different-age talkers. Finally, age and age match interacted, $\chi^2(1) = 8.82$, $p = .003$, suggesting a larger log–odds difference between same-age and different-age conditions for adults than for children.

New models were computed to compare adult and child performances on parent–child, parent–parent, and child–child talker pairs. Each model used age as a between-participants fixed effect and participants intercepts as random effects (maximum fixed effects correlation: −.210). Additional models tested whether adult and child participants exceeded chance on particular talker pairs by using only the intercept and participants random effects (no fixed effects correlations because there was only one fixed effect). Adults exceeded chance for all talker pairs (parent–parent: estimate = 1.64, *SE* = 0.49, $z = 3.33$, $p = .0009$; parent–child: perfect performance; child–child: estimate = 0.89, *SE* = 0.30, $z = 2.94$, $p = .003$). Adults outperformed children in parent–parent pairs (estimate = 0.65, *SE* = 0.18, $z = 3.57$, $p = .0004$) and parent–child pairs ($\chi^2 = 16.19$, $p = .0004$), and they nearly outperformed children in child–child pairs (estimate = 0.32, *SE* = 0.17, $z = 1.91$, $p = .06$).

Children exceeded chance only on parent–child pairs (*M* = 81.8%, *SD* = 23.3; estimate = 2.26, *SE* = 0.44, $z = 5.22$, $p < .0001$). They were near chance on child–child pairs (*M* = 55.2 ± 11.9%; vs. chance: estimate = 0.04, *SE* = 0.20, $z = 1.02$, $p = .31$), which did not suggest any advantage for same-age talkers. How did this compare with previous experiments? We computed models with experiment as a between-participants factor and with participants intercept as a random factor. Children's accuracy on parent–child pairs in the current study exceeded children's accuracy on same-age female voices in Experiment 1 (estimate = 0.33, *SE* = 0.39, $z = 3.57$, $p = .0004$; fixed effects correlation: −.227) but was slightly lower than that on gender-mismatched voices in Experiment 4 (estimate = −0.54, *SE* = 0.27, $z = 2.04$, $p = .04$; fixed effects correlation: −.515).

Finally, we assessed whether the PPVT-IV predicted talker learning in children. There was a nonsignificant correlation between accuracy (overall or per condition) and the PPVT in 21 of 24 children who

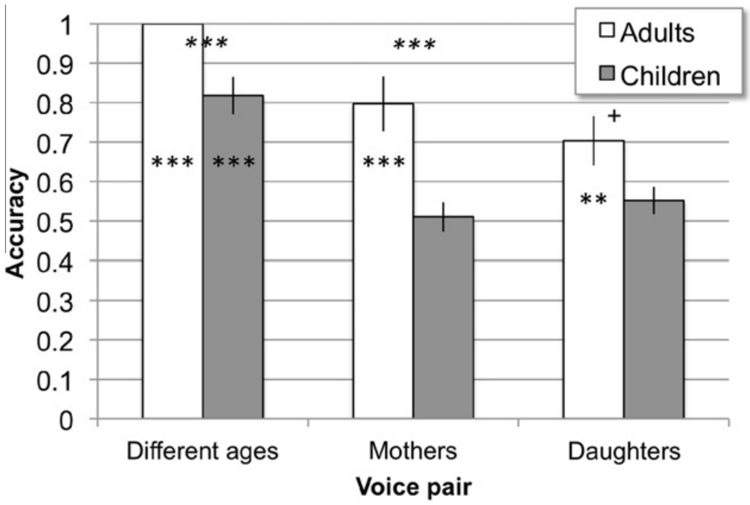


Fig. 5. Adult and child accuracy for different-age and same-age voice pairs from Experiment 6. Italicized asterisks indicate child–adult participant comparisons, and the rest indicate difference from chance. $^{+}p < .10$;$^{**}p < .01$; $^{***}p < .001$.

completed the PPVT (maximum $r = .23$, $p = .29$). The PPVT did correlate with age ($r = .57$, $p = .007$), suggesting that there was a reasonable amount of variance in scores.

Discussion

In this final experiment, preschoolers robustly learned a second socially and acoustically salient talker contrast—age. Nonetheless, adults continued to outperform children. In addition, children performed at chance for own-age voices, failing to support an own-age effect in talker encoding. This further confirms the general picture that preschoolers are not at adult proficiency at mapping voices to individuals, consistent with protracted tuning of talker recognition. In addition, children's vocabulary size did not significantly predict talker learning, which does not support a strong relationship between talker encoding and language processing.

Developmental trends

We were interested in looking at broader patterns of performance in cases where children had difficulty in encoding voices. In particular, was there evidence of more accurate talker encoding with increasing age? To assess age trends, we looked at performance (Fig. 6) across all of the young adult female talker experiments (Experiments 1–3 and 5, $n = 91$ children for whom age data were available). Performance showed a significant gain in accuracy with age ($r = .32$, $p = .002$). However, the relationship was not strong; some of the oldest children performed near chance, suggesting that talker encoding skills are far from asymptote at 6 years of age.

Another interesting aspect of the data is evident in this analysis: Accuracy seems to be bimodally distributed (Fig. 6, inset). Bimodal performance is consistent with one set of children performing at chance and another set performing well above chance. However, the bimodal pattern did not appear to result from any obvious participant characteristics. Means did not differ by gender (boys: 57 ± 17%; girls: 58 ± 19%); $t(89) = 0.26$, $p = .80$, or language status (monolingual: 57 ± 17%; bilingual or exposed to two languages: 58 ± 19%), $t(89) = 0.46$, $p = .65$. Thus, it is unclear what might explain this pattern. We return to this point in the General Discussion.

General discussion

We began by asking how talker information is processed across development. We specifically tested whether preschool-aged children were better at talker learning than adults, reflecting a gradual filtering out of talker variability, or whether adults exceeded children at talker learning, reflecting protracted perceptual tuning to talker variability. The clear answer is that children are far from adult performance, consistent with the protracted tuning hypothesis and inconsistent with developmental declines in talker sensitivity due to filtering out talker variation. Our 3- to 6-year-olds readily mapped different-gender (Experiment 4) and different-age (Experiment 6) voices to characters. However, children were less adept than adults at mapping two same-age female voices to talkers, whether voices differed in formants (Experiments 1–3) or fundamental frequency (Experiment 5). Performance was not aided by limiting verbal content variability (Experiment 2), by using more visually realistic child-aged characters (Experiment 3), or by using own-age (child) talkers (Experiment 6). These results suggest that, much like the emerging picture of speech sound acquisition, 3- to 6-year-olds are still learning a complex set of acoustic cues that map to talker identity. In the next sections, we consider how this learning process may occur.

Encoding talker information

If talker variability indeed requires protracted learning, what is being learned? It seems clear that talker representations take advantage of some of the same cues used in speech sound identification (McMurray & Jongman, 2011). Thus, one possibility is that speech sounds are used as a scaffold for encoding talker variation. This is substantiated by adult research suggesting that phonological knowledge in dyslexic listeners (Perrachione et al., 2011) and dialect familiarity in normal listeners (Perrachione et al., 2010) predict talker encoding. Given that children are less adult-like in recognizing speech sounds (e.g., Nittrouer, Lowenstein, & Packer, 2009; Ohde & Haley, 1997), they may have a weaker scaffold for encoding talker information.

Another possibility for improvement over age is that *talker* representations scaffold learning of new talkers, and children are worse than adults because they know relatively fewer talkers; they have a smaller "talker lexicon." One might construe this as children beginning with a small number of talker representations. Some research (e.g., Hayes & Taplin, 1993) suggests that children essentially represent *prototypes,* representations based on central tendencies in their experience. Sloutsky and Fisher (2004), on the other hand, suggested that 5-year-olds have more item-specific representations than adults, that is, more exemplar-like representations. On either account, children start with a limited number of representations and base additional talker representations on these representations. Which pattern, if either one, is more reflected in our data—the prototype account or the exemplar account? If talkers are encoded relative to prototypes of talkers such as mother, father, and child, we should expect young children to be uniformly bad at recognizing highly similar talkers. However, if talkers are

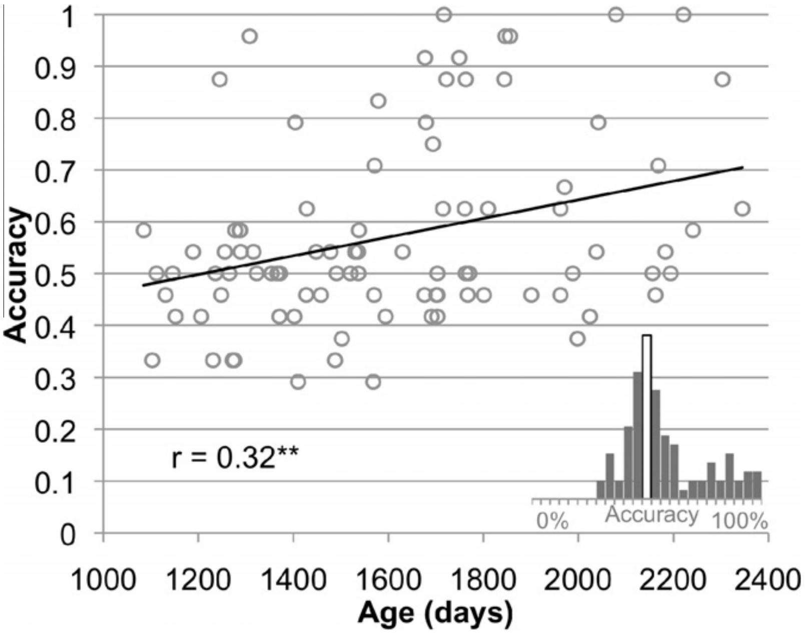


Fig. 6. Scatter plot of children's accuracy by age in Experiments 1, 2, 3, and 5 (two college-aged female voices). Note that the linear effect of age is nearly identical for female and male participants. Inset: Histogram of accuracy, where the white bar denotes chance performance (50%). $^{**}p = .002$.

encoded relative to representations of other specific talkers, one might expect that a few children might benefit from knowing talkers who sound similar to the talkers tested, leading to a bimodal pattern in children's talker encoding ability in our task. As shown in the inset of Fig. 6, a bimodal pattern appears in children's accuracy in learning similar voices. This is consistent with a small number of children having preexisting knowledge of voices like those used (although other explanations are certainly possible). Another point in favor of talker-specific encoding is that, if children are encoding talkers relative to other talkers rather than encoding talkers relative to language knowledge, we would not expect a strong relationship between language knowledge and talker learning. In Experiment 6, we failed to find a significant effect of PPVT-IV scores on talker learning.

However, some aspects of our data provide less support for talker-based talker encoding. For instance, the PPVT-IV might not have predicted voice learning because it measures vocabulary rather than directly measuring phonological knowledge; had we measured phonological processing, we might have found a stronger relationship. Furthermore, if children were encoding new talkers with respect to known talkers, children in Experiment 6 should have encoded child talkers easily due to their greater exposure to children's voices. However, child listeners were not better at recognizing child talkers than at recognizing adult talkers.

Recognizing talkers: early development

The current research suggests changes in talker processing from childhood to adulthood, but one question that still remains is how processing of talker variability changes between infancy and childhood. We have argued that there is continuous perceptual improvement from infancy through adulthood. However, complicating this account is an apparent discontinuity in development; infants under 1 year of age seem much more sensitive than older infants or adults to talker *variation.* Does this mean that young infants are better than young children at recognizing voices, implying a developmental discontinuity (unlike that observed in language development) (Friederici, 2005; Friederici, 2006)? Differences in tasks used make interpretation difficult (see Apfelbaum & McMurray, 2011, for an account of differing performances across infant tasks); studies showing talker *sensitivity* in infants have used dishabituation (Johnson et al., 2011; Rost & McMurray, 2009, 2010) or listening time measures (Houston & Jusczyk, 2000). To our knowledge, there are no studies on talker *learning* in infants, which would be most comparable to our task.

We propose that these results do not reflect superior infant acuity in distinguishing talkers but rather that young children (and adults) are being presented with a more taxing task—mapping. Evidence from multiple areas of speech processing suggests that mapping acoustics to meaning is far more difficult than simple acoustic change detection. For instance, although 6-month-olds react differently to different vocal affects (Singh, Morgan, & Best, 2002), children cannot map vocal cues to emotions until 4 or 5 years of age (Quam & Swingley, 2012). In addition, 8-month-olds discriminate *bih* from *dih,* but they cannot map those similar words to referents until 17 to 20 months of age (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002), although they succeed given talker variability (Rost & McMurray, 2009, 2010). Finally, 7-month-olds dishabituate to voice changes (Johnson et al., 2011), but our 3- to 6-year-olds had difficulty in *mapping* similar voices to characters. These patterns together suggest that mapping various cues to talkers may require more protracted learning than simple detection of an acoustic change.

Talker encoding as person identification

We have focused primarily on perceptual learning processes underlying child and adult differences in talker recognition. However, it is also fruitful to consider talker recognition in the broader context of recognizing individuals and social groups because these factors may catalyze attention to talker information in the speech signal. In this light, it is interesting that even *adults'* talker recognition abilities are not always impressive. For instance, Bartholomeus (1973) found that both children and adults were far better at recognizing children's faces than their voices. Van Lancker et al. (1985) found that adults identified famous talkers they thought they knew with approximately 68% accuracy—meaning that they *failed* to recognize those talkers 30% of the time in a closed-set task. Presumably, they would have been far more accurate at recognizing the faces (see Bahrick, Bahrick, & Wittlinger, 1975). There is some evidence that adults might not pay attention to talker differences if attention is directed toward the speech stream that they themselves are articulating in a shadowing (immediate speech repetition) task (Cherry, 1953). Infants also seem to listen preferentially to sounds that are of a complexity commensurate with their own level of language production (Lange-Küttner, 2010). Adults may even fail to pay attention to talker differences if they do not know that the talker is relevant to the task at hand (Vitevitch, 2003; see also Creel & Tumlin, 2011, Experiment 4) in a phenomenon analogous to visual change blindness (Simons & Levin, 1998). Do these studies imply that perceivers generally are worse at recognizing voices than faces, or do they simply reflect the high degree of complexity of talker encoding?

There are at least three aspects of the talker acoustic–perceptual space that may make talkers more difficult to encode than faces. First is the close relationship between speech sounds and talker variation (e.g., Bricker & Pruzansky, 1966; McMurray & Jongman, 2011); just as talker variability adds noise to speech sound distributions, speech sound variability adds noise to talker characteristics. Second, the space of talker-related cues may be much larger than the space of speech-related cues (see Van Lancker et al., 1985, for a similar argument). For instance, English talkers can vary in both phonemic aspects such as voice onset time (Allen et al., 2003) and in nonphonemic aspects such as a creaky voice. This may make the space of all possible talkers much larger and, thus, harder to learn.

A third possible reason for better face recognition than talker recognition is that talker perceptual space may be characterized by dense clusters, whereas face perceptual space is more evenly distributed. This may occur because people are much more successful at *sounding* like each other than at *looking* like each other. Voice characteristics are highly malleable to learning processes in the short term (adaptation to an interlocutor; e.g., Pardo, 2006) and in the long term (sociolinguistic communities; e.g., Labov, Ash, Ravindranath, Weldon, & Nagy, 2011). Faces, because they are less prone to learning effects, may be more evenly distributed across similarity space and, thus, more readily learned. In summary, there are multiple reasons why perceptual learning of talker variability may be a particularly difficult problem.

A final question is whether social categories have an influence on talker encoding. Do salient social differences between two speakers focus attention on differences between their voices, or would those voice differences be salient even without social difference? That is, is children's ability to distinguish voices enhanced by apparent social differences, or is all voice encoding a function of acoustic discriminability? Children seem adept at using speech characteristics to make social judgments (Hirschfeld & Gelman, 1997; Kinzler et al., 2007), suggesting that children link voice differences to social group differences. Furthermore, our child participants learned talkers best when talkers differed socially (in age or gender), although these talkers tended to be *acoustically* less similar as well. Of course, if children are so strongly socially driven, same-age pictures (Experiment 3) or same-age voices (Experiment 6) would presumably have facilitated talker learning, and they did not. This outcome is more consistent with acoustic distinctiveness, rather than social distinctiveness, driving effects. Problematically, social similarity and even physical similarity (Krauss, Freyberg, & Morsella, 2002) are confounded with acoustic similarity to some extent, requiring ingenuity in future studies to find even larger acoustic differences that do not imply (real or imagined) differences in social categories.

Conclusion

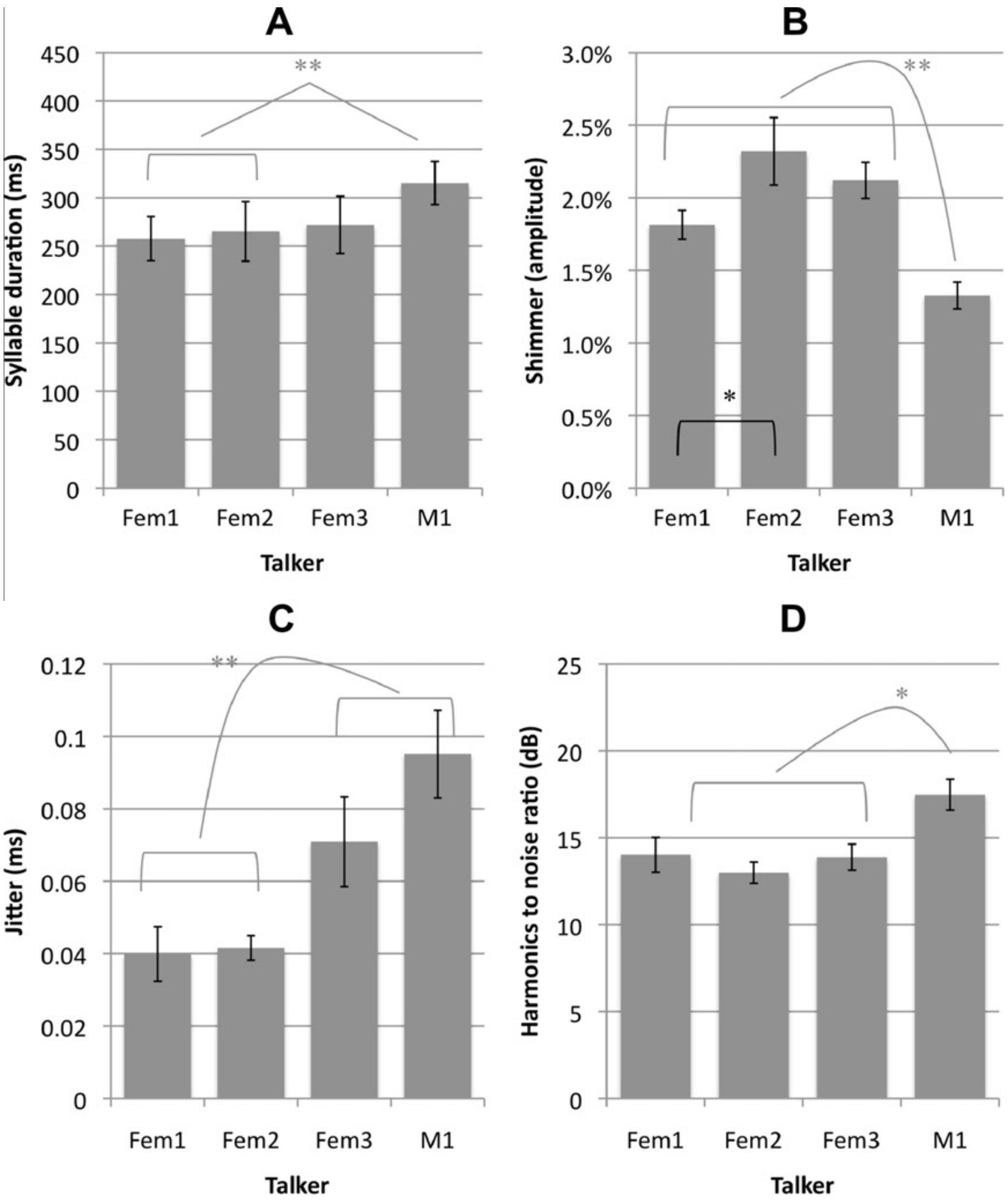
How is talker information in the speech signal processed during development? In this study, children mapped perceptually distinct voices to cartoon characters, but children were far less accurate than adult listeners when asked to learn voices that were more perceptually similar. Results suggest that talker identification undergoes protracted tuning over development. These results complement a growing literature suggesting lengthy perceptual learning of speech-related variability. Rather than listeners tuning out irrelevant characteristics of the speech signal during the first few years of life, the emerging picture is one of slow continuous improvements in young listeners' tuning to both speech sound categories and talkers.

Acknowledgments

Thanks go to Heather Pelton and other Creel lab members for running participants, to Rayna Friendly for valuable discussion, to individuals who provided voices, and to child participants, parents, and administrators for participation. S.C.C. was supported by a National Science Foundation (NSF) CAREER Award (BCS 1057080), and S.R.J. was supported by a Genentech Scholars summer research fellowship.

Appendix A

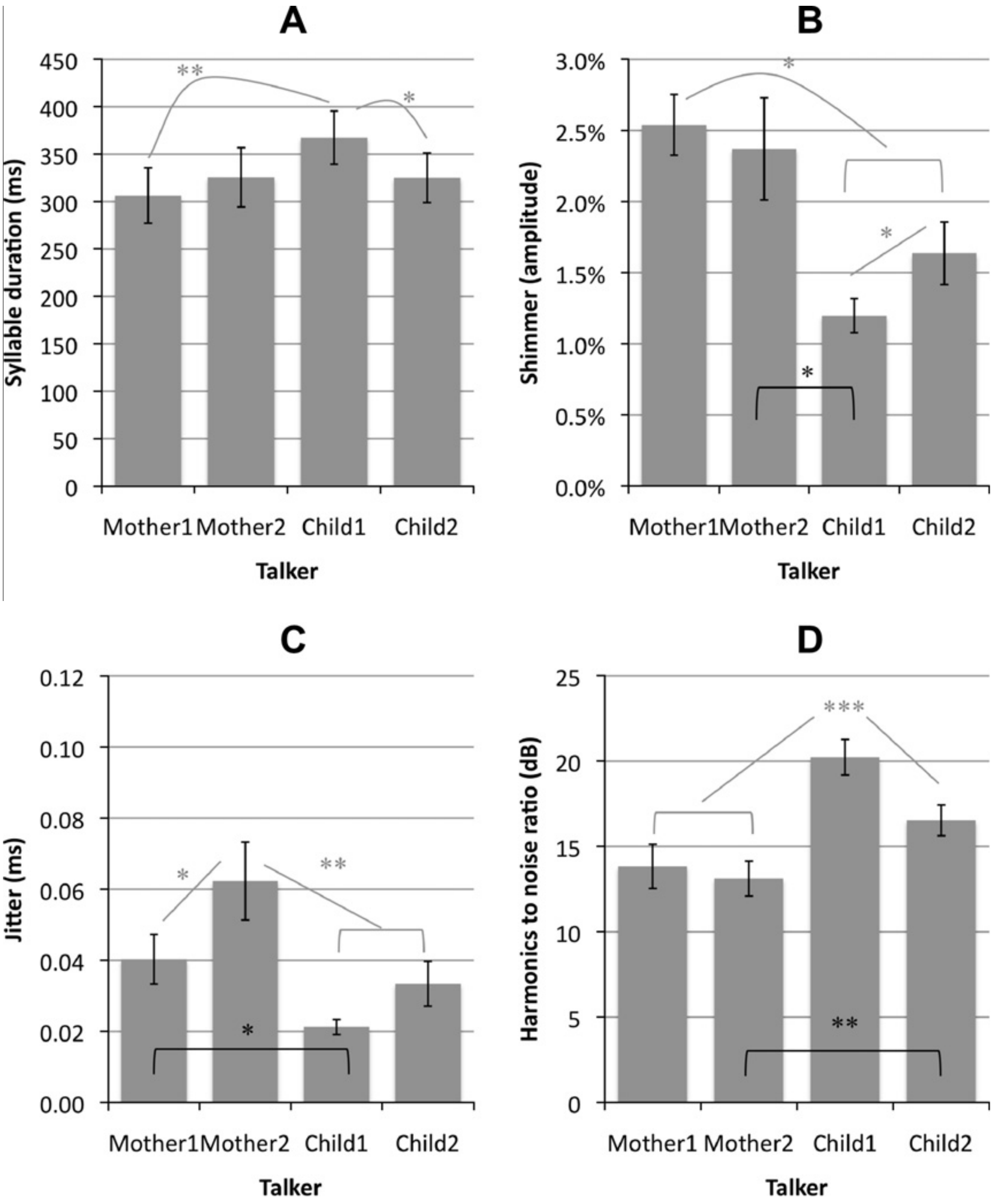
Additional voice measures conducted in Praat for talkers in Experiments 1 to 5: (A) syllable duration (utterance duration/number of syllables in ms); (B) shimmer (average difference between amplitude of an f0 period and the average of its two neighbors); (C) jitter (average difference between successive period durations in ms); (D) harmonics-to-noise ratio (in dB; related to hoarseness)



*p < .05.
**p < .01.

Appendix B

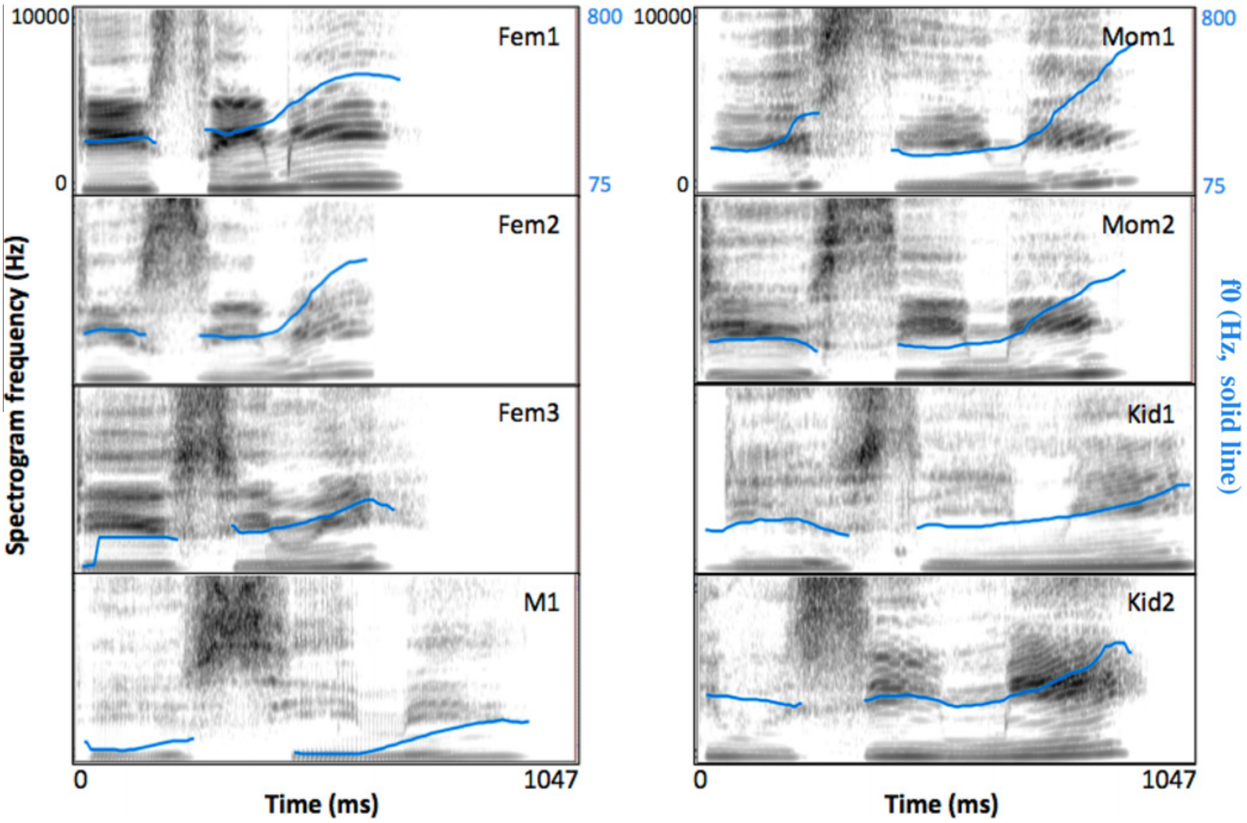
Additional voice measures for talkers in Experiment 6: (A) syllable duration (in ms); (B) shimmer (amplitude variability); (C) jitter (variability in period duration); (D) harmonics-to-noise ratio (in dB).



$^{*}p < .05$.
$^{**}p < .01$.
$^{***}p < .001$.

Appendix C

Spectrograms (plus f0 track in solid line) of talkers from all experiments saying, ‘‘Do you see me?’’



Note. Fundamental frequency (f0) and spectrum use different *y*-axes.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset time. *Journal of the Acoustical Society of America, 113*, 544–552.

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science, 35*, 1105–1138.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104*, 54–75.

Barker, B. A., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition, 94*, 45–53.

Bartholomeus, B. (1973). Voice identification by nursery school children. *Canadian Journal of Psychology, 27*, 464–472.

Berman, J. M. J., Chambers, C. G., & Graham, S. A. (2010). Preschoolers' appreciation of speaker vocal affect as a cue to referential intent. *Journal of Experimental Child Psychology, 107*, 87–99.

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native- and Jamaican-accented words. *Psychological Science, 20*, 539–542.

Boersma, P., & Weenink, D. (2010). Praat: Doing phonetics by computer (Version 5.1.44) [computer program]. Retrieved 4 October 2010 from <http://www.praat.org>.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40*, 1441–1449.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*, 975–979.

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

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience. The role of talker variation in lexical access. *Cognition, 106*, 633–664.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*, 264–285.

Cutler, A., & Swinney, D. A. (1987). Prosody and the development of comprehension. *Journal of Child Language, 14*, 145–167.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208*, 1174–1176.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-IV: Peabody picture vocabulary test* (4th ed.). Circle Pines, MN: American Guidance Service.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2208–2213). Amsterdam, Netherlands.

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends in Cognitive Sciences, 9*, 481–488.

Friederici, A. D. (2006). The neural basis of language development and its impairment. *Neuron, 52*, 941–952.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*, 448–458.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Hayes, B. K., & Taplin, J. E. (1993). Developmental differences in the use of prototype and exemplar-specific information. *Journal of Experimental Child Psychology, 55*, 329–352.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099–3111.

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

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1570–1582.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1143–1154.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science, 14*, 1005–1011.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359–384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252–1263.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32*, 402–420.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 12577–12580.

Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H. H., et al (2003). Effects of experience on fetal voice recognition. *Psychological Science, 14*, 220–224.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics, 15*, 431–463.

Lange-Küttner, C. (2010). Discrimination of sea-bird sounds vs. garden-bird songs: Do Scottish and German-Saxon infants show the same preferential looking behaviour as adults? *European Journal of Developmental Psychology, 7*, 578–602.

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition, 122*, 12–36.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242–1255.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /1/: A first report for publication. *Journal of the Acoustical Society of America, 89*, 874–886.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 391–409.

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

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*, 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101–B111.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12*, 369–378.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*, 219–246.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143–178.

Melinder, A., Gredebäck, G., Westerlund, A., & Nelson, C. A. (2010). Brain activation during upright and inverted encoding of own- and other-age faces: ERP evidence for an own-age bias. *Developmental Science, 13*, 588–598.

Moher, M., Feigenson, L., & Halberda, J. (2010). A one-to-one bias and fast mapping support preschoolers' learning about faces and voices. *Cognitive Science, 34*, 719–751.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72*, 834–843.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

Nelson, N. L., & Russell, J. A. (2011). Preschoolers' use of dynamic facial, bodily, and vocal cues to emotion. *Journal of Experimental Child Psychology, 110*, 52–61.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America, 109*, 1181–1196.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*, 62–85.

Nittrouer, S., Lowenstein, J. H., & Packer, R. R. (2009). Children discover the spectral skeletons in their native language before the amplitude envelopes. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1245–1253.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Tokyo: Ohmasha.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Ohde, R. N., & Haley, K. L. (1997). Stop-consonant and vowel perception in 3- and 4-year-old children. *Journal of the Acoustical Society of America, 102*, 3711–3722.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition, 114*, 42–55.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science, 333*, 595.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Pezdek, K., Blandon-Gitlin, I., & Moore, C. (2003). Children's face recognition memory: More evidence for the cross-race effect. *Journal of Applied Psychology, 88*, 760–763.

Quam, C., & Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Development, 83*, 236–250.

Ramsey, J. L., Langlois, J. H., & Marti, N. C. (2005). Infant categorization of faces: Ladies first. *Developmental Review, 25*, 212–246.

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

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*, 608–635.

Schmale, R., Cristià, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy, 15*, 650–662.

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science, 12*, 583–601.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644–649.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*, 833–870.

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy, 3*, 365–394.

Sloutsky, V. M., & Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15*, 553–558.

Spence, M. J., Rollins, P. R., & Jerger, S. (2002). Children's recognition of cartoon voices. *Journal of Speech, Language, and Hearing Research, 45*, 214–222.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381–382.

Staum Casasanto, L. (2008). Does social information influence sentence processing? In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 799–804). Washington, DC.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 13273–13278.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters: I. Recognition of backward voices. *Journal of Phonetics, 13*, 19–38.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 333–342.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3*, 1–30.

Werker, J. F., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America, 123*, 4524–4538.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition, 113*, 234–243.