Effect of the relationship between target and masker sex on infants' recognition of speech

Rochelle S. Newman[a] **and Giovanna Morini**
Department of Hearing & Speech Sciences, University of Maryland, 0100 Lefrak Hall,
College Park, Maryland 20742, USA
Rnewman1@umd.edu, gmorini@udel.edu

Abstract: When faced with multiple people speaking simultaneously, adult listeners use the sex of the talkers as a cue for separating competing streams of speech. As a result, adult listeners show better performance when a target and a background voice differ from one another in sex. Recent research suggests that infants under 1 year do not show this advantage. So when do infants begin to use talker-gender cues for stream segregation? These studies find that 16-month-olds do not show an advantage when the masker and target differ in sex. However, by 30 months, toddlers show the more adult-like pattern of performance.

[DDO]
Date Received: September 16, 2016 **Date Accepted:** January 23, 2017

1. Introduction

Infants and toddlers are frequently in environments where multiple people speak simultaneously (e.g., Barker and Newman, 2004; van de Weijer, 1998). In these situations, the ability to separate the competing streams of speech, and attend to a single source, is of critical importance for being able to learn from or recognize the spoken message.

One cue that adult listeners rely on is the sex of the talker. Female talkers typically have both a higher fundamental frequency ($F0$) and generate higher formant frequencies than men (Peterson and Barney, 1952). Adults use these acoustic differences to aid in their separation of different voices/streams of speech, and, as a result, generally show better performance listening to a voice in the presence of an opposite-sex distractor stream than in the presence of a same-sex distractor (Arehart *et al.*, 1997; Assmann and Summerfield, 1990; Brokx and Nooteboom, 1982; Darwin *et al.*, 2003; Darwin and Hukin, 2000; Festen and Plomp, 1990; Mackersie *et al.*, 2011).

Work with infants (Newman, 2009) suggests that they have particular difficulty attending to one voice while another voice speaks in the background. Moreover, one recent study suggests that infants do not show the advantage that adults show when there is a difference in talker sex across streams. Leibold *et al.* (2013) conditioned 7- to 13-month-olds and adults to respond whenever they heard a particular spondee. Listeners were then tested on their ability to identify the target spondee word in the presence of a two-talker masker. Whereas adults showed better performance (a lower threshold) when the target speaker was of the opposite sex than the masker, infants showed no difference in threshold levels across conditions. The authors suggest that "…*infants may derive less benefit than adults from a mismatch in target and masker sex. This would be consistent with an immature ability to use sex-based acoustic differences to segregate auditory streams*" (p. 4).

Thus, whereas a number of studies suggest that differences in talker sex are an important cue for adult listeners, this same cue appears to be of far less benefit for infants (Leibold *et al.*, 2013). When might the ability to use this cue develop? The current study focuses on toddlers, both because they are in a period of rapid lexical growth (where the ability to attend to one talker might be of a particular benefit) and because they are frequently in daycare and public settings, where hearing multiple voices is likely common. In experiment 1, we test 30-month-olds on their ability to recognize speech spoken by a female voice in the presence of either a same-sex or an opposite-sex masker.

2. Experiment 1

We used a variant of the preferential-looking procedure (Golinkoff *et al.*, 1987; Spelke, 1979) to test children's word recognition. Children saw two visual objects

[a] Author to whom correspondence should be addressed.

simultaneously, and heard a female voice telling them (in a child-directed speaking style) which object to look at. This voice was combined with a background stream consisting of a single talker (either male or female) reading a passage in adult-directed speech, at a 5 dB signal-to-noise ratio (SNR). This SNR was chosen because it falls within the range of noise levels typically found in day care settings (Frank and Golden, 1999), and thus is one that children might experience on a regular basis.

2.1 Method

Participants. Eighteen toddlers (12 male, 6 female), aged 30 months (range: 29.4–30.9 months; mean 30.4) participated. Data from two additional children were excluded for excessive fussiness/crying or parental distraction. An additional child was later identified with an Autism Spectrum Disorder, and her data were replaced. The children were assigned to one of six stimulus orders (see Procedure), with three participants per order. Parents reported that their children had normal hearing, and were not currently experiencing symptoms indicative of an ear infection. All heard at least 80% English in the home, with 4 children hearing a second language 5%–15% of the time.

Stimuli. Stimuli consisted of digital images of eight words (*truck, ball, kitty, doggy, blocks, keys, horsie, birdie*) well-known to children of this age (Fenson *et al.*, 1994), and a simultaneous audio signal. Image pairs matched for approximate size and color. The audio recordings were a blend of a target voice stream and a distractor voice stream. The target stream consisted of a single female talker producing four sentences. Three of the sentences ended with a target word while the first sentence was an attention getter ("Hey baby! Look at the ____! Do you see the ____? Where's the ____?"). The initial attention getter was included because studies with adults suggest that stream segregation can take time to build up (Bregman, 1978); as we were unsure what the buildup time might be for children of this age, we felt that including a slightly longer lead in might be advantageous. Baseline trials contained similar sentences that did not indicate any particular object ("Hey baby! Look at that! Do you see that? Look there!"). The two distractor voices (one male, one female) both read the same fluent passage from Cormac McCarthy's *Blood Meridian*. All sentences were recorded in a noise-reducing soundbooth (Shure SM81 microphone, Shure, Inc., Niles, IL, 44.1 kHz sampling rate, 16 bits precision), adjusted to be the same duration and root-mean-square (RMS) amplitude, and were combined at a 5 dB SNR. The distractor passage began first, followed 250 ms later by the onset of the target stream. The target word occurred 2000 ms subsequent to this point; total duration of the combined recordings was 7.25 s.

Procedure. Children sat on their caregiver's lap, facing a widescreen TV. At the start of each of 20 trials, an image of a baby laughing appeared to attract the participant's attention. Subsequently, two images appeared at ∼20° visual angle simultaneously with an auditory stimulus at approximately 65–70 dBA (all items were matched for RMS amplitude prior to presentation).

Each pair of objects occurred a total of 5 times across the 20 trials, twice with the correct answer being the object on the left, twice the object on the right, and once with the voice just instructing the child to look in general (baseline trials). Objects were always presented in the same pairs (truck/ball; keys/blocks, etc.); which member of a pair appeared on the left vs right was counterbalanced across participants. Of the 16 non-baseline trials (test trials), 8 occurred in the presence of the male distractor voice (one per object) and 8 with the female distractor voice. Baseline trials were used to ensure that there were no consistent object biases (across children) or side biases (within a child). Across children, each of the 8 objects was looked at between 30% and 70% of the time on baseline trials, suggesting no strong preference for or against a particular object. All participants looked at least 25% of the time to each side, suggesting no strong side biases.

Participants were presented with one of six different pseudo-randomized trial orders [with the restriction that the correct response did not occur on the same side (left vs right) more than 3 trials in a row]. We examined whether children looked at the appropriate object as a measure of their ability to understand the speech despite the distractor. The caregiver listened to masking music over headphones throughout the study to prevent any biasing of the child's behavior.

Coding. Two experimenters, blind to condition, individually coded each child's looking behaviors on a frame-by-frame basis using Supercoder coding software (Hollich, 2005). The first 2250 ms (68 frames) of each trial occurred prior to the first presentation of the target word. Since children could not know which object to look at until hearing the word, these initial 68 frames were excluded from analyses of target looking. Accuracy was defined as a proportion of time that infants remained fixated on the picture of the target object, rather than the foil, subsequent to the onset of the target word. If the two coders disagreed on any trial by more than 15 frames (0.5 s), a third coder was used. The averages of the two closest codings were used as the final data. This occurred on 34 of the 360 trials (20 test trials per child × 18 children), or just under 10% of the time. The final data were extremely reliable; correlations on the percentage of left (vs right) looking for each participant ranged from 0.9918 to 0.9994 with an average correlation of 0.9971.

There are a number of different measures that have been taken from preferential looking studies; we chose to base all measures on overall proportions of looking to the correct object. We measured this both over the entire trial (starting at the onset of the word, and continuing until the end of the trial) and over a limited 2000-ms window beginning 367 ms post word-onset. Using a shorter window is more common in the literature, because of a concern that participants might look initially toward the correct answer but then scan back and forth for the remainder of the trial, adding noise to any analysis based on an entire trial. However, work using a limited window of analysis has all been conducted in quiet listening conditions, and we were concerned that noise might slow children's processing. In such a case, using a short analysis window might limit the opportunity to see children's actual abilities. We therefore initially conducted analyses both ways; we found a nearly identical pattern regardless of window size, but results were less variable with the longer windows, and we thus report those data below.

We presume that if children can understand the speech, despite the distractor, they will look longer to each image on those trials when it is named (and is thus the target) than on trials when that object was the distractor (and the alternate object was named). This method of comparison has previously been used successfully (Bergelson and Swingley, 2012) and has the advantage of including looking to each image in both the target and distractor conditions. We expect that if children can use talker sex as a cue to separate streams or to attend to the target voice, they should perform better when the distractor voice is male (and easier to distinguish from the female target voice) than when it is female (and thus putatively more similar to the target voice).

2.2 Results and discussion

For each of the two background speech conditions, we calculated the proportion of time the child spent looking at each object when named, and compared this to the amount of time spent looking at that same object when unnamed. We then conducted a 2 (talker sex) × 2 (named vs unnamed) repeated measures analysis of variance (ANOVA). We found a significant effect of naming [$F(1,17) = 96.91$, $p < 0.0001$], and critically, a significant interaction between naming and talker sex [$F(1,17) = 9.89$, $p < 0.01$]. When the background voice was male, children looked at the appropriate image 74% of the time, a significant increase over chance performance [$t(17) = 10.71$, $p < 0.0001$]. When the background voice was female, children looked at the appropriate image 67% of the time, also a significant increase over chance [$t(17) = 7.20$, $p < 0.0001$], but a lesser difference [$t(17) = 3.34$, $p < 0.005$]. (Because we are using proportions of looking to the correct vs incorrect object, there is no possibility of an overall effect of talker sex—on each trial overall looking across the two objects totals 100% of the looking time, so this cannot differ across the different background voices.) Excluding the four children who heard another language besides English in the home did not change the results. Thus, children aged 30 months appear to be able to use the sex of two voices to help them when listening in a multi-talker environment.

3. Experiment 2

The results from experiment 1 suggest that 30-month-old toddlers can use talker sex as a cue to help them distinguish two voices. In contrast, prior results suggest that infants under 1 year of age do not show such an effect (Leibold *et al.*, 2013). In order to better compare recent infant results to the results found here, we opted to use the current method to test younger children.

Since the current study involves a word-recognition task, the youngest age we can reasonably test is the age when these words are typically learned. According to the Macarthur Communicative Development Inventory (Fenson *et al.*, 1994), all of the words in the current study are reported to be known by a majority of children at age 16 months. In contrast, these objects are not all known by a majority of children at 14 months. We therefore opted to test children aged 16–17 months using the same methods as in experiment 1.

3.1 Method

Participants. We anticipated that these younger participants might show a more variable performance. We therefore increased our sample size slightly to compensate; 24 infants (14 male, 10 female), aged 16–17 months (range: 16.0–17.4 months; mean 16.8) participated. An additional 8 children participated, but their data were excluded for excessive fussiness/crying or for parental intervention ($n = 6$), equipment failure ($n = 2$). Parents reported that their children had normal hearing, and were not currently experiencing symptoms indicative of an ear infection. All heard at least 90% English in the home, with 5 hearing another spoken language between 5% and 10% of the time, and one being exposed to sign language simultaneously with spoken English.

Stimuli, Procedure, and Coding. Stimuli, procedures, and coding were identical to those in experiment 1, except that pilot testing suggested the 5 dB SNR level was too difficult for these younger children (performance was not clearly above chance, making it impossible to determine if talker sex was a useful cue); we therefore increased the SNR to 10 dB. A third coder was used on 9.6% of the trials (46 out of 480). Each of the objects was looked at between 35% and 65% of the time on baseline trials (across participants), and all participants looked at least 20% of the time to each side across baseline trials, suggesting no strong side or object biases.

3.2 Results and discussion

As before, we conducted a 2 (talker sex) × 2 (named vs unnamed) repeated measures ANOVA. We found a significant effect of naming [$F(1,23) = 29.80$, $p < 0.0001$], but no interaction between naming and sex [$F(1,23) = 0.26$, $p = 0.62$]. When the background voice was male, children looked at the appropriate image 59% of the time, a significant increase over chance performance [$t(23) = 4.46$, $p < 0.03$]. Although this is not as strong of a looking preference as that from the older infants, the above-chance looking indicates that the children do know these words, and are able to recognize them in the presence of a male voice distractor. When the background voice was female, children also looked at the appropriate image at above chance levels: 58% of the time [$t(23) = 5.27$, $p < 0.02$]. These looking proportions are nearly identical, and there was no significant difference between these two conditions [$t(23) = 0.50$, $p = 0.62$], suggesting that 16-month-old children do not use the acoustic correlates of the sex of two voices as a cue to help them when listening in a multi-talker environment. These 16-month-olds appear to be behaving more like infants in prior studies (Leibold *et al.*, 2013) than like the older toddlers tested in experiment 1.

We then compared across the two experiments, using a 2 (naming: named vs unnamed) × 2 (talker sex) × 2 (age group) repeated measures ANOVA. As expected, we found a significant effect of naming [$F(1,40) = 128.38$, $p < 0.0001$], which interacted with participant age [$F(1,40) = 21.38$, $p < 0.0001$], indicating that older children showed a stronger effect of the named object in general. We also found a significant naming × sex interaction [$F(1,40) = 7.84$, $p = 0.008$], and, critically, a significant 3-way interaction [$F(1,40) = 4.52$, $p < 0.05$]. Thus, the sex of the background talker had different impacts on children at the two ages, as can be seen in Fig. 1.

The overall effect of age may be an indication that younger infants are less adept at listening in multi-talker environments, but could simply be an indication that they know the target words less well; we cannot distinguish between these two possibilities. But because the distractor voices were presented in a within-subjects design, using the same target words, lexical knowledge cannot explain the difference between the male- and female-voice distractor conditions. This difference can only be explained by children's use of sex-related acoustic cues either to separate speech streams or to choose which stream to attend to after they are separated.

One possibility is that the female voice's higher $F0$ captured younger children's attention, making it harder to ignore. Although we did not measure preferences for the background voices in isolation, we can partially address this by examining looking time on baseline trials; if the female distractor is attracting attention, we might expect longer overall attention (to either object) on these trials. However, this was not the case: 16-month-olds averaged 131 frames looking on male-distractor trials, 132 on female-distractor trials; 30-month-olds averaged 122 frames on male trials, 111 on female trials, $p > 0.10$ for both age groups. Nor was performance simply at ceiling, since looking times could have been as high as 150 frames. Thus, it does not appear that younger children's attention was simply captured by the female background voice.

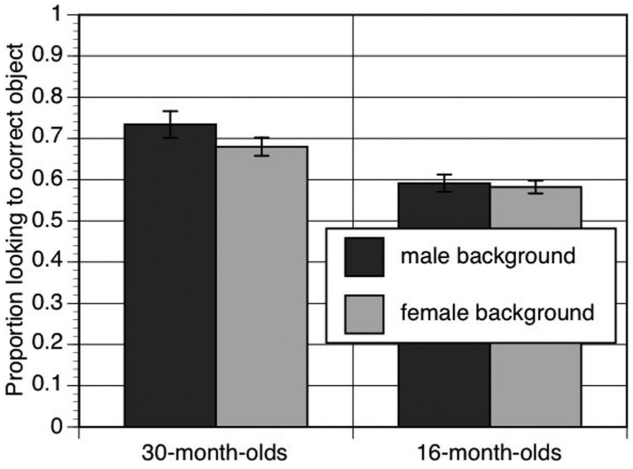

Fig. 1. Proportion of looking time to the correct object (where 50% = chance) for both male and female background voices at 30 months of age (left) and 16 months of age (right).

4. Final discussion

When faced with a multi-talker situation, adult listeners regularly use the sex of the talkers as a cue for separating streams of speech and attending to one stream. Recent research has suggested that young infants (under 1 year of age) do not show the same pattern—their ability to use the acoustic differences that correlate with a change in sex appears to be immature (Leibold *et al.*, 2013). The current study demonstrates that 30-month-old toddlers, but not 16-month-olds infants, do show this ability. Thus, there appears to be some developmental change that is occurring between 16 and 30 months of age.

It is not clear what might be driving this change. These differences might result from maturation in the ability to choose a source for attention. Prior studies have demonstrated that young children have difficulty focusing attention on particular components of a signal or particular frequency ranges (Bargones and Werner, 1994; Polka *et al.*, 2008). Perhaps the ability to benefit from talker sex requires a similar type of attentional focus.

Perhaps the difficulty young infants face is tied to a lack of linguistic experience. Children aged 16 vs 30 months have vastly different amounts of linguistic experience (as indexed in part by a tenfold increase in average vocabulary size; Fenson *et al.*, 1994). Perhaps greater linguistic experience with words not only makes it easier to identify those words in a noisy/degraded signal in general, but also makes it easier to take advantage of available cues that would distinguish those words from background interference.

Another possibility is that the younger infants were not able to hear the difference between the male and female voices. This is possible, but seems unlikely, because even neonates can discriminate male and female voices (Lecanuet *et al.*, 1993). Thus, the ability to distinguish the pitch or timbre differences between genders is in place well before the age tested here.

Finally, it is possible that young children simply do not choose to attend to the target voice, even though they are able to do so—perhaps results would have been different had the target voice been one that was personally important to them (e.g., their mother). Even newborns prefer their mother's voice over other voices (DeCasper and Fifer, 1980) and infants can use this familiarity to help them attend to their mother's voice in the presence of distractors by as young as 7 months (Barker and Newman, 2004). Had the target voice in the current study been highly familiar, children may have had greater motivation to attempt to distinguish the target and background streams, and might have shown an effect of gender differences at a younger age.

Regardless of the underlying cause, the present findings support prior work suggesting that young infants do not use sex-based acoustic differences to aid in listening in the presence of distractors. The ability to use such cues appears to undergo further development between the age of 16 and 30 months.

Acknowledgments

This work was supported in part by NSF and NIH grants to the University of Maryland. We thank George Hollich for providing us with the coding software, and David Rossell, Sean Hendricks, Judy Segal, Whitney Goodrich-Smith, and TeHsin Wu for stimulus creation/recording. We also thank the following for assistance in scheduling, testing, and coding the infants and toddlers: K. Ablorh, A. Abrams, M. Abrams, C. Ali, A. Ambielli, A. Arnold, A. Urbanus, A. Bail, C. Bender, J. Berges, T. Bipat, A. Bloomfield, T. Booth, D. Brunson, K. Cavanaugh, R. Childress, J. Chinnaya, A. Cook, J. Coon, B. Darr, J. Dombroski, S. Dougherty, S. Edelberg, L. Evans, A. Fisher, L. Fischer, A. Gandee, G. Giangrasso, A. Goron, L. Goudreau, C. Hoff, L. Horowitz, M. Janssen, M. Javid, A. Jensen, C. Kettl, E. Kim, P. Kozlofsky, A. Layman, S. Lee, R. Lieberman, P. Lieberman, S. McCruden, E. McLaughlin, K. McPherson, V. Moussavi, M. Nasuta, C. O'Connor, S. Panza, A. Pasquarella, C. Penny, E. Perkins, C. Pontell, K. Puyear, R. Rhodes, A. Rodriguez, H. Saffeir, J. Segal, R. Sherman, E. Shroads, S. Smith, L. Steedman, M. Steinberg, A. Stone, A. Thomas, K. Voelmle, C. Vogel, A. Wildman, C. Wilson, K. Wilson, E. Younkin, and D. Zack-Williams.

References and links

Arehart, K. H., King, C. A., and McLean-Mudgett, K. S. (**1997**). "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss," J. Speech, Lang., Hear. Res. **40**, 1434–1444.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**(2), 680–696.

Bargones, J. Y., and Werner, L. A. (**1994**). "Adults listen selectively; infants do not," Psychol. Sci. **5**(3), 170–174.

Barker, B. A., and Newman, R. S. (**2004**). "Listen to your mother! The role of talker familiarity in infant streaming," Cognition **94**(2), B45–B53.

Bergelson, E., and Swingley, D. (**2012**). "At 6-9 months, human infants know the meanings of many common nouns," Proc. Natl. Acad. Sci. U.S.A. **109**(9), 3253–3258.

Bregman, A. S. (**1978**). "Auditory streaming is cumulative," J. Exp. Psychol. **4**, 380–387.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**(5), 2913–2922.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

DeCasper, A. J., and Fifer, W. P. (**1980**). "Of human bonding: Newborns prefer their mothers' voices," Science **208**(4448), 1174–1176.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., and Pethick, S. J. (**1994**). "Variability in early communicative development," Monographs Soc. Res. Child Develop. Serial 242, **59**(5), 1–173.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736.

Frank, T., and Golden, M. V. (**1999**). "Acoustical analysis of infant/toddler rooms in daycare centers," J. Acoust. Soc. Am. **106**(4), 2172.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (**1987**). "The eyes have it: Lexical and syntactic comprehension in a new paradigm," J. Child Lang. **14**, 23–45.

Hollich, G. (**2005**). "Supercoder: A program for coding preferential looking" (Version 1.5). [Computer Software]. (Purdue University, West Lafayette, IN).

Lecanuet, J.-P., Granier-Deferre, C., Jacquet, A.-Y., Capponi, I., and Ledru, L. (**1993**). "Prenatal discrimination of a male and a female voice uttering the same sentence," Infant Child Develop. **2**(4), 217–228.

Leibold, L. J., Taylor, C. N., Hillock-Dunn, A., and Buss, E. (**2013**). "Effect of talker sex on infants' detection of spondee words in a two-talker or a speech-shaped noise masker," Proc. Meet. Acoust. **19**, 060074.

Mackersie, C. L., Dewey, J., and Guthrie, L. A. (**2011**). "Effects of fundamental frequency and vocal-tract length cues on sentence segregation by listeners with hearing loss," J. Acoust. Soc. Am. **130**(2), 1006–1019.

Newman, R. S. (**2009**). "Infants' listening in multitalker environments: Effect of the number of background talkers," Attn., Percept., Psychophys. **71**(4), 822–836.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–189.

Polka, L., Rvachew, S., and Molnar, M. (**2008**). "Speech perception by 6- to 8-month-olds in the presence of distracting sounds," Infancy **13**(5), 421–439.

Spelke, E. S. (**1979**). "Perceiving bimodally specified events in infancy," Develop. Psychol. **15**, 626–636.

van de Weijer, J. (**1998**). *Language Input for Word Discovery* (Ponsen and Looijen, BV, Wageningen, the Netherlands).