Toddlers' comprehension of adult and child talkers:

adult targets versus vocal tract similarity

Angela Cooper*

Natalie Fecher

Elizabeth K. Johnson

Department of Psychology, University of Toronto, 3359 Mississauga Rd., Mississauga, ON, L5L 1C6, CANADA

angela.cooper@utoronto.ca; natalie.fecher@utoronto.ca; elizabeth.johnson@utoronto.ca

*Corresponding author

Word count: 3000

Abstract

How do children represent words? If lexical representations are based on encoding the indexical characteristics of frequently-heard speakers, this predicts that speakers like a child's own mother should be best understood. Alternatively, if they are based on the child's own motor productions, this predicts an own-voice advantage in word recognition. Here, we address this question by presenting 2.5-year-olds with recordings of their own voice, another child's voice, their own mother's voice, and another mother's voice in a child-friendly eye-tracking procedure. No own-voice or own-mother advantage was observed. Rather, children uniformly performed better on adult voices than child voices, even performing better for unfamiliar adult voices than own voices. We conclude that children represent words not in the form of own-voice motor codes or frequently heard speakers, but on the basis of adult speech targets.

Keywords: developmental speech perception, speech production, word recognition, indexical variation, talker familiarity

1. Introduction

Spoken word recognition involves matching the incoming acoustic input onto stored linguistic representations. This mapping process is complicated by a variety of talker-related factors, including differences in vocal tract size, speaking rate, and accent. Understanding the nature of these linguistic representations has been the focus of considerable study, as it provides insight into how listeners extract a stable percept from a continuously varying acoustic signal. While adult native listeners are adept at efficiently and accurately recognizing words despite this variability (e.g., Clarke & Garrett, 2004; Cutler & Broersma, 2005), the problem of variation is compounded for children. They have smaller vocabularies, less robust phonemic categories, and are still learning what variation is phonologically relevant for distinguishing between words and what variation can be ignored (e.g., Best, Tyler, Gooding, Orlando, & Quann, 2009; Schmale, Cristià, Seidl, & Johnson, 2010). In the face of such variation, how do young children mentally represent and access words? The present work examines the nature of children's early lexical representations by investigating the influence of speaker age (child vs. adult) and familiarity (maternal and own voice vs. strangers' voices) on spoken word recognition.

Given that adult listeners are proficient at recognizing speech produced by a range of talkers, the adult system must be sufficiently flexible to adapt to this variation. Previous research has posited that adult listeners accommodate variation by encoding context-specific non-linguistic information alongside linguistic information during speech perception. Adult listeners have been found to be sensitive to indexical variation, such that listeners are slower and less accurate at identifying or recalling words when there is a change in talker and show enhanced word recognition when listening to a familiar talker (e.g., Mullennix & Pisoni, 1990; Nygaard, Sommers, & Pisoni, 1994). Similarly, linguistic processing in young children also appears to be influenced by indexical variation. For example, Jerger et al. (1993) tested both adults and

children in a Garner speeded classification task, examining the extent to which indexical and linguistic dimensions are integrally processed, finding that indexical variation interfered with linguistic processing and that the magnitude of this interference declined with age.

Many models of spoken word recognition consider linguistic representations to contain acoustic information about a given lexical item (e.g., McClelland & Elman, 1986); however, an alternative view involves the representation of motor actions. According to the common coding theory of perception, percept and action codes are stored within a common representational space, and perception is facilitated when the incoming input more closely matches the stored action code (Prinz, 1990). That is, our perception of an event is influenced by its perceived similarity to how we ourselves would produce that same event. This predicts an own-action advantage in perception, as perceiving self-generated actions would be the best match to our stored action codes. Evidence for this has been found in such domains as writing (e.g., Knoblich, Seigerschmidt, Flach & Prinz, 2002), dart-throwing (Knoblich & Flach, 2001), and piano performance (Repp & Knoblich, 2004). With regards to perceiving speech, this would predict that speech recognition would be facilitated when perceiving one's own speech. Because each speaker has a unique vocal tract size and set of motor patterns, there is greater correspondence between perceived and stored speech motor plans when the same person is both the speaker and listener/observer. Visual speech perception findings with adults support this hypothesis (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2013), as participants were more accurate at lipreading their own productions relative to unfamiliar productions. Tye-Murray, Spehar, Myerson, Hale, and Sommers (2014) also reported an own-voice advantage in audio-visual speech recognition in adverse listening conditions.

Schuerman, Meyer, and McQueen (2015) examined whether this own-voice advantage extends to auditory-only word recognition in adults. Participants identified noise-vocoded words that either the listener had produced or that were productions of the statistically-average speaker. However, contrary to the predictions of the common coding theory, results revealed that listeners were more accurate at identifying words produced by the average speaker relative to their own voice. The authors posit that auditory word recognition may not utilize representations shared by production and perception. Given that performance was better on an average speaker, it may be the case that representations used in auditory perception are developed by aggregating and abstracting over the relevant perceptual information from a range of different speakers so as to be able to generalize to novel speakers.

Talker-related variability in lexical productions tend to be greater in children than in adults (Vihman, 1993), making the problem of mapping speech input onto stored representations all the more challenging for listeners. Prior word recognition studies have nearly always tested children on unfamiliar adult voices (e.g., Swingley & Aslin, 2000); however, children's pronunciations can differ dramatically in systematic ways from adult pronunciations, stemming from differences in vocal tract size, articulatory control and linguistic knowledge. Little is known about how young children (or adults) perceive speech produced by other children (e.g., Bernier and White, 2017; Masapollo, Polka, & Ménard, 2016). Pre-babbling infants have been found to prefer listening to speech with infant vocal properties over adult speech; though, the inclusion of infant vowels in a multi-talker set increased processing demands (Polka, Masapollo, & Ménard, 2014). There is evidence suggesting that children do not find the speech of other children easier to understand than adult productions. Hazan and Markham (2004) tested 7- to 8-year-old

children perceiving speech of other children (*M*=13 years old) embedded in noise and did not find evidence that child talkers were more intelligible to children than adults.

The present work sought to better understand how children perceive the range of speech variation they encounter and its implications for the nature of early lexical representations. To that end, we examined the influence of speaker age and familiarity on spoken word recognition. Children and their mothers were recorded producing a set of words and later returned to complete an eye-tracking task, which presented pairs of pictures of familiar objects, named by one of four voices: 1) their own voice, 2) their own mother's voice, 3) an unfamiliar child's voice, or 4) an unfamiliar mother's voice. If representations are based on shared percept and action codes, as posited by the common coding theory, then children should perform best on own-voice productions followed by the unfamiliar child's productions, as the vocal tracts and motor patterns of child speakers are more similar to the child listener than adult speakers (Motor Hypothesis). However, if listeners' representations are based on exemplar traces containing integrated indexical and linguistic information, then children may instead show a maternal-voice advantage, as the frequency bias in the distribution of accrued exemplars over their lifespan would likely favour their mother's voice (Familiarity Hypothesis).

2. Methods

2.2 Participants

Fifty-four normally developing Canadian English-learning 30- to 36- month-olds were tested (age range=941-1113 days; 32 boys). Parents reported no hearing impairments or recent ear infections. Children were exposed to primarily English (*M*=96% English exposure, range=85-100%) and mothers had a North American English accent. An additional 5 toddlers were tested but were excluded due to experimenter error (4) and fussiness (1).

2.3 Stimuli

The materials consisted of 32 words (4 lists of 8 words each; see Appendix) typically known by 30-month-olds, as indexed by an average word production rate of 95% according to Wordbank vocabulary norms (Frank, Braginsky, Yurovsky, & Marchman, 2016). Images representing the target words were selected, matched for approximate size and visual complexity, for use in an eye-tracking task.

All 32 words were produced by each child and their mother. Every child-mother dyad was paired with a gender-matched dyad to ensure that each dyad's productions would be heard by another participant. Within each set of dyads, the 4 word lists were divided between the 4 talkers (2 children, 2 mothers), and accordingly, 8 productions were segmented per person (leaving 24 productions per talker not presented in the eye-tracking task). Only a subset of the productions was used in the experiment due to limitations in toddlers' attention spans. Which list was segmented for a child versus an adult and for a familiar versus an unfamiliar talker was counterbalanced across sets. Recordings were equalized to the same RMS amplitude level. These productions served as the auditory stimuli for the eye-tracking task built specifically for each participant set.

2.4 Procedure

2.4.1 Production

Word productions were elicited in an experimenter-controlled video game. Children were informed they would be teaching an alien English. An image of the referent of a target word was displayed on the screen, and the alien verbally prompted the child to name the picture, at which point the child was expected to produce the word. Following the child's production, the mother

also produced the word. Participants were encouraged to produce the word in citation form and were prompted to repeat the item as necessary.

2.4.2 Eye-tracking task

After at least one week (*M* = 19 days, range = 7-28 days), children returned to complete the eye-tracking task. Children were presented with 32 pairs of images against a white background; one of these images was a named target, the other an unnamed distracter. Each image was presented twice, serving once as a target and once as a distracter. Every child heard 8 object names each with their own voice, their own mother, an unfamiliar mother and an unfamiliar child, for a total of 32 trials. Target images occurred equally often on each side, and presentation side of the target image was counterbalanced across participants.

Each 6000 ms trial began with the presentation of a pair of pictures. After 300 ms, a non-speech auditory attention-getter was presented. 3000 ms after trial onset, the target word was presented. The experimental session was videotaped and subsequently coded frame-by-frame off-line using SuperCoder (Hollich, 2005). Each 33-ms frame was coded as a look to the left, right, or elsewhere. The two coders were not aware of the auditory or visual information of the trials. Inter-coder agreement on fixation durations was high (mean correlation across four participants' data = 0.97). Following prior work (e.g., Delle Luche, Durrant, Poltrock, & Floccia, 2015), the proportion of fixation time on the target picture was calculated (target fixation time/total fixation time to target + distracter).

3. Results

Target fixation proportions for each voice type were analyzed for a one-second window starting from 267 ms post-word onset (e.g., Creel, 2014; Figure 1). Trials where the participant did not fixate on either the target or distracter during the post-naming period were excluded.

Children's proportion of target fixations were well above chance (50%) for all four voice types ($t(53) > 4.64$, $p<0.001$).

To compare effects of speaker age and familiarity, a linear mixed effects regression (LMER) model was conducted (Baayen, Davidson, & Bates, 2008), with the proportion of target fixations as the dependent measure and contrast-coded fixed effects for speaker age (adult, child) and familiarity (familiar, unfamiliar) and their interaction. The maximal random effects structure that would converge was implemented, including random intercepts for participant and item and by-participant random slopes for age and familiarity. Model comparisons were performed to determine whether the inclusion of each fixed factor and the interaction made a significant contribution to the model. A significant main effect of speaker age was found ($\beta = 0.06$, SE $\beta = 0.02$, $\chi^2(1) = 7.76$, $p = 0.005$), with a greater proportion of target fixations for adult voices relative to child voices. No significant effect of familiarity or familiarity x age interaction was obtained ($\chi^2<0.719$, $p>0.4$). To test whether these data support the null hypothesis, we conducted a Bayesian analysis (Rouder, Speckman, Sun, & Morey, 2009). A Bayes factor (BF) larger than 10 constitutes strong evidence for an effect, while a BF < 1/10 is strong evidence in favor of the null hypothesis. Speaker age yielded a BF of 30.9, while familiarity (BF = 0.06) and familiarity x age (BF = 2.6) did not substantively influence the word recognition results.

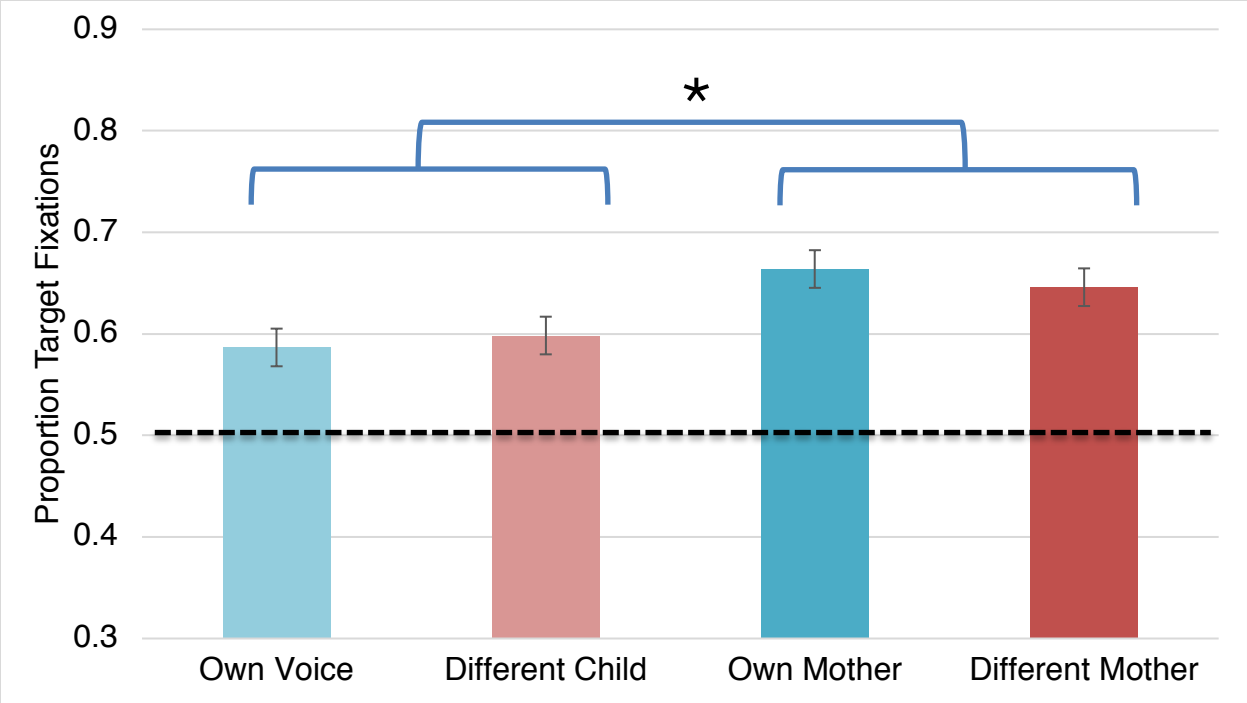

Figure 1 Proportion of fixations to the target image for each voice type during a one-second window starting 250 ms post target word onset. Errors bars indicate +/- 1 standard error. Dashed line denotes chance level.

Additionally, the latency of shift to target was also calculated (Figure 2). This measure determines how long it takes for gaze to shift onto the target image if fixating on the distracter at word onset. Log-transformed latencies were submitted to an LMER model with the same contrast-coded fixed effects and random effects structure as the previous analysis. Consistent with the proportion of target fixation findings, a significant effect of speaker age was obtained (β = -0.15, SE β = 0.06, $\chi^2(1) = 5.74$, $p = 0.017$), but no effect of familiarity or familiarity x age interaction ($\chi^2<1.87$, $p>0.17$). Participants were faster to switch from the distracter to the target image at word onset when the voice was an adult rather than a child. While there was a numerical trend for faster switches to the target for own-mother voices, this did not reach significance.

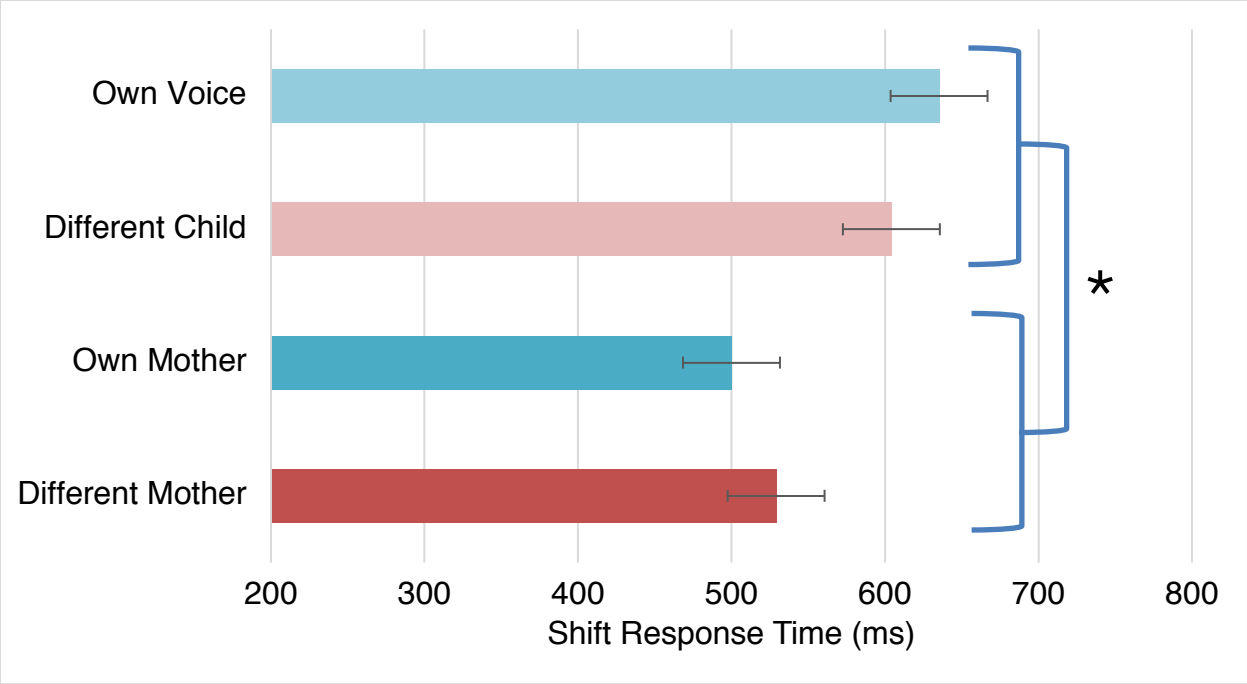

Figure 2 Mean shift response time in milliseconds from distracter to target following word onset for each voice type.

4. Discussion

Both adults and children must represent lexical items in a fashion that allows them to recognize words despite substantial variation in their acoustic-phonetic realization. How children, with smaller vocabularies and less developed phonological systems than adults, accomplish this feat is still a mystery. The current study addressed this question by presenting 2.5-year-olds with recordings of their own voice, another child's voice, their own mother's voice, and another child's mother in a word recognition task. We predicted that if children's representations use shared percept and action codes within a common representational space, then an own-voice advantage should emerge in speech perception, (i.e., Motor Hypothesis; Prinz, 1990). Alternatively, if children's lexical representations are based on frequently heard input, we predicted an own-mother voice advantage (i.e., Familiarity Hypothesis; Nygaard et al., 1994). Contrary to our predictions, however, our results did not strongly support either of these

hypotheses. Rather, toddlers were significantly faster and more accurate at recognizing words produced by adult over child talkers, regardless of familiarity. Indeed, 2.5-year-olds performed better on an unfamiliar adult voice than their own voice. One could argue that the lack of familiarity effect could stem from children's inability to recognize their own voice in a recording; indeed, bone conduction makes one's voice sound different when hearing it while speaking versus in a recording. However, in a separate voice recognition task using these stimuli, children were above chance at identifying both their own mother and own voice, indicating that the absence of a familiarity effect was not due to a lack of awareness of the speaker's identity.

These results suggest that children's lexical representations are not based on their own motor patterns or frequent input, but instead on adult speech targets. That is, children may be aggregating and abstracting over the multitude of adult exemplars in their input, rather than their own productions, from which they form prototypes on which to base their early lexical representations. Convergent evidence for this possibility can be found in adult research, with Schuerman et al. (2015) reporting more accurate word recognition of a statistically-average speaker rather than self-speech. Despite sharing similar vocal tract and motor characteristics with the child speakers, listeners in the present study benefited more from hearing adult productions. This is also in line with findings indicating that overhearing speech early in life (without ever producing it) yields improved productions of sounds in that language (Choi, Broersma & Cutler, 2017; Knightly et al., 2003).

What implications do these findings have for our understanding of the relationship between perception and production? There has been evidence in adult word learning research, where new words that had been produced (versus heard-only) during training were faster to be recognized at test, suggesting a connection between perceptual, production and semantic

representations. However, in contrast to the adults, preliminary work with children suggests that auditory and not production training promoted superior word-learning performance (Zamuner, Morin-Lessard, Strahm, & Page, 2016). This may indicate that the influence of production on speech representations evolves over the lifespan. Given how variable child productions are, it is conceivable that early representations are based on more stable adult speech targets until the speaker's own productions stabilize. Indeed, MacDonald et al. (2012) found, using an altered auditory feedback paradigm, that toddlers did not monitor and self-regulate their speech productions in the same manner as adults and young children. They posited that such compensatory behaviours as a result of auditory feedback discrepancies only develop once internal representations have stabilized.

In sum, to our knowledge, the current study provides the first investigation of own versus other voice perception in children. This work examined two indexical factors, namely speaker age and familiarity, on spoken word recognition in toddlers. Superior performance for adult over child voices was found, irrespective of their familiarity to the listener. Overall, this research supports the notion that early lexical representations are based on adult speech targets rather than own-speech motor codes. It remains for future research to tease apart whether early representations are based on adult productions because those are what they are predominantly exposed to early in life (e.g., caregivers, family members) or whether adult productions are weighted more heavily than child tokens because there is less within- and between-talker variability and are thus deemed more reliable.

5. Acknowledgments

Thanks to Yazad Bhathena and Lisa Hotson, and the members of the Child Language and Speech Studies Lab for their support. This work was supported by grants from the Social Sciences and Humanities Research Council, Natural Sciences and Engineering Research Council, and the Canada Research Chairs program. Portions of this work were presented at the 3rd Workshop for Infant Language Development in Bilbao (June 2016).

6. Appendix

baby
ball
bear
bike
bird
boat
bunny
butterfly
cow
dog
duck
elephant
finger
fish
frog
horse
house
monkey
orange
phone
plane
shoe
spoon
squirrel
strawberry
stroller
swing
toothbrush
train
tree
truck
turtle

7. References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bernier, D., & White, K. S. (2017). What's a foo? Toddlers are not tolerant of other children's mispronunciations. In M. LaMendola & J. Scott (Eds.) *Proceedings of the Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. pp. 88-100.

Best, C., Tyler, M., Gooding, T., Orlando, C., & Quann, C. (2009). Development of phonological constancy. *Psychological Science*, *20*(5), 539–542. https://doi.org/10.1111/j.1467-9280.2009.02327.x

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Creel, S. C. (2014). Preschoolers' flexible use of talker information during word learning. *Journal of Memory and Language*, *73*(1), 81–98. https://doi.org/10.1016/j.jml.2014.03.001

Cutler, A., & Broersma, M. (2005). Phonetic Precision in Listening. In W. J. Hardcastle & J. Mackenzie Beck (Eds.), *A Figure of Speech* (pp. 64–91). Mahwah, NJ: Lawrence Erlabum Associates.

Delle Luche, C., Durrant, S., Poltrock, S., & Floccia, C. (2015). A methodological investigation of the Intermodal Preferential Looking paradigm: Methods of analyses, picture selection and data rejection criteria. *Infant Behavior and Development*, *40*, 151–172. https://doi.org/10.1016/j.infbeh.2015.05.005

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: A Repository For Vocabulary Data. *Journal of Child Language*, *18*, 1–18.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the Acoustical Society of America*, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

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

Jerger, S., Pirozzolo, F., Jerger, J., Elizondo, R., Desai, S., Wright, E., & Reynosa, R. (1993). Developmental trends in the interaction between auditory and linguistic processing. *Perception & Psychophysics*, *54*(3), 310–320. https://doi.org/10.3758/BF03205266

MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology*, *22*(2), 113–117. https://doi.org/10.1016/j.cub.2011.11.052

Masapollo, M., Polka, L., & Ménard, L. (2016). When infants talk, infants listen: Pre-babbling infants prefer listening to speech with infant vocal properties. *Developmental Science*, *19*(2), 318–328. https://doi.org/10.1111/desc.12298

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception As a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Polka, L., Masapollo, M., & Ménard, L. (2014). Who's talking now? Infants' perception of vowels with infant vocal properties. *Psychological Science*, *25*(7), 1448–56.

https://doi.org/10.1177/0956797614533571

Repp, B. H., & Knoblich, G. (2004). Perceiving Action Identity: How Pianists Recognize Their Own Performances. *Psychological Science*, *15*(9), 604–609. https://doi.org/10.1111/j.0956-7976.2004.00727.x

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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*(6), 650–662. https://doi.org/10.1111/j.1532-7078.2010.00032.x

Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS ONE*, *10*(7), 1–18. https://doi.org/10.1371/journal.pone.0129731

Swingley, D. (2016). Two-Year-Olds Interpret Novel Phonological Neighbors as Familiar Words. *Developmental Psychology*, *52*(7), 1011–1023.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2013). Reading your own lips: common-coding theory and visual speech perception. *Psychonomic Bulletin & Review*, *20*(1), 115–9. https://doi.org/10.3758/s13423-012-0328-5

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2014). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise.

Psychonomic Bulletin & Review, 1–6. https://doi.org/10.3758/s13423-014-0774-3

van Heugten, M., & Johnson, E. K. (2015). Toddlers' Word Recognition in an Unfamiliar Regional Accent: The Role of Local Sentence Context and Prior Accent Exposure. *Language and Speech*, 1–11. https://doi.org/10.1177/0023830915600471

Vihman, M. (1993). Variable Paths to Early Word Production. *Journal of Phonetics*, *21*, 61–82.

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. P. A. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, *89*, 55–67. https://doi.org/10.1016/j.jml.2015.10.003