Cognition 173 (2018) 16–20

Contents lists available at ScienceDirect

Cognition

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

Brief article

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

ARTICLE INFO

Keywords:
Developmental speech perception
Speech production
Word recognition
Indexical variation
Talker familiarity

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.

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., McLelland & 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, 1997). 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

* Corresponding author.
E-mail addresses: angela.cooper@utoronto.ca (A. Cooper), natalie.fecher@utoronto.ca (N. Fecher), elizabeth.johnson@utoronto.ca (E.K. Johnson).

https://doi.org/10.1016/j.cognition.2017.12.013
Received 14 July 2017; Received in revised form 11 December 2017; Accepted 20 December 2017

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 & 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

2.1. 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).

2.2. 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.

2.3. Procedure

2.3.1. 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.

2.3.2. 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,