Cognition 129 (2013) 362–378

Contents lists available at ScienceDirect

Cognition

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

Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?

Bob McMurray [a,b,c,d,*], Kristine A. Kovack-Lesh [e], Dresden Goodwin [e], William McEchron [a]

[a] Dept. of Psychology, University of Iowa, United States
[b] Dept. of Communication Sciences and Disorders, University of Iowa, United States
[c] Dept. of Linguistics, University of Iowa, United States
[d] The Delta Center, University of Iowa, United States
[e] Dept. of Psychology, Ripon College, United States

ARTICLE INFO

Article history:
Received 20 September 2012
Revised 18 July 2013
Accepted 22 July 2013
Available online 24 August 2013

Keywords:
Infant directed speech
Speech categorization
Statistical learning
Phonetic analysis
Vowels
Voices onset time

ABSTRACT

Infant directed speech (IDS) is a speech register characterized by simpler sentences, a slower rate, and more variable prosody. Recent work has implicated it in more subtle aspects of language development. Kuhl et al. (1997) demonstrated that segmental cues for vowels are affected by IDS in a way that may enhance development: the average locations of the extreme "point" vowels (/a/, /i/ and /u/) are further apart in acoustic space. If infants learn speech categories, in part, from the statistical distributions of such cues, these changes may specifically enhance speech category learning. We revisited this by asking (1) if these findings extend to a new cue (Voice Onset Time, a cue for voicing); (2) whether they extend to the interior vowels which are much harder to learn and/or discriminate; and (3) whether these changes may be an unintended phonetic consequence of factors like speaking rate or prosodic changes associated with IDS. Eighteen caregivers were recorded reading a picture book including minimal pairs for voicing (e.g., *beach*/*peach*) and a variety of vowels to either an adult or their infant. Acoustic measurements suggested that VOT was different in IDS, but not in a way that necessarily supports better development, and that these changes are almost entirely due to slower rate of speech of IDS. Measurements of the vowel suggested that in addition to changes in the mean, there was also an increase in variance, and statistical modeling suggests that this may counteract the benefit of any expansion of the vowel space. As a whole this suggests that changes in segmental cues associated with IDS may be an unintended by-product of the slower rate of speech and different prosodic structure, and do not necessarily derive from a motivation to enhance development.

1. Introduction

During the first year of life, infants' speech perception systems begin to be tuned to the characteristics of their native language (Werker & Curtin, 2005; Werker & Tees, 1984). Over the first 12–18 months, infants show a reduction in their ability to discriminate phonetic contrasts that are not used in their language (Werker & Lalonde, 1988; Werker & Tees, 1984); they gain the ability to discriminate difficult contrasts (Eilers & Minifie, 1975; Eilers, Wilson, & Moore, 1977); and they are continually refining existing categories (Kuhl, Stevens, Deguchi, Kiritani, & Iverson, 2006). A growing number of scholars have posited that this process is guided, in part, by the statistics of acoustic cues in the speech that infants hear (de Boer & Kuhl, 2003;

* Corresponding author. Address: Dept. of Psychology, University of Iowa, E11 SSH, Iowa City, IA 52242, United States. Tel.: +1 319 335 2408 (voice); fax: +1 319 335 0191.
E-mail address: bob-mcmurray@uiowa.edu (B. McMurray).

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.07.015

Guenther & Gjaja, 1996; Maye, Werker, & Gerken, 2003; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; McMurray, Aslin, & Toscano, 2009; Pierrehumbert, 2003; Toscano & McMurray, 2010; Vallabha, McClelland, Pons, Werker, & Amano, 2007), and recent work shows that computational models of this learning mechanism can account for all three patterns of development (McMurray, Aslin, et al., 2009).

Statistical learning is based on the idea that phonological speech contrasts can be described by one or more continuous acoustic cues, which themselves are the product of articulation. For example, voicing (which distinguishes /b, d, g/ from /p, t, k/) is marked primarily by voice onset time (or VOT, the continuous time between the release of the articulators and the onset of voicing) (Lisker & Abramson, 1964). For voiced sounds, like /b,d,g/, the release of the articulators (the lips or tongue) occurs nearly simultaneously with the onset of voicing (in languages like English), resulting in VOTs near 0 ms. For voiceless sounds, like /p, t, k/, the onset of voicing is delayed by about 50 ms after the consonantal release. However, variation across talkers, speaking rates, and the effects of other phonetic properties of the signal creates some variation around these means resulting in statistical clusters (Fig. 1A; Allen & Miller, 1999; Lisker & Abramson, 1964).

Analogously, most vowels can be characterized by the frequency of the first three formants and their duration (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). The vowel /i/ as in *beet*, for example has a low F1 and a high F2; while /ɑ/ as in *Bob* has a high F1 and a low F2. These individual formant frequencies derive in part from the position of the tongue during the articulation of the vowel; as this is variable as a function of talker, coarticulation, etc., those cues also form statistical clusters around the prototypical values for the vowels of the language. Here, however, clusters may only be distinct when examined in two dimensions (Fig. 1B; data from Cole, Linebaugh, Munson, & McMurray, 2010; see also Hillenbrand et al., 1995; Peterson & Barney, 1952).

Given this description of the input, distributional learning posits a fairly simple mechanism for acquiring speech categories. By estimating the mean (or prototypical) cue-value and variance (or extent of allowable variation around this mean) of each cluster, children could arrive at a reasonable set of descriptors for the categories along a dimension or dimensions. There has been an explosion of computational models that show this can be done by a variety of learning mechanisms (de Boer & Kuhl, 2003; Guenther & Gjaja, 1996; McMurray, Aslin, et al., 2009; McMurray & Spivey, 2000; Toscano & McMurray, 2010; Vallabha et al., 2007). These models demonstrate how a variety of [largely] unsupervised clustering approaches can harness the statistical structure of the input to find the relevant categories, and thus establish the computational tractability of this hypothesis.

Evidence for such mechanisms comes from two sources. First, adult perceptual categories show a graded structure (Andruski, Blumstein, & Burton, 1994; Kuhl, 1991; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; McMurray, Tanenhaus, & Aslin, 2002; Miller, 1997; Miller & Volaitis, 1989; Toscano, McMurray, Dennhardt, & Luck, 2010; Utman, Blumstein, & Burton, 2000; Volaitis & Miller, 1992) that matches the graded clusters of speech cues. Infants are also sensitive to such gradations, contra earlier claims of categorical perception (Galle & McMurray, submitted for publication; McMurray & Aslin, 2005; Miller & Eimas, 1996). This correspondence suggests that this gradiency may be a remnant of the statistical learning process that undergirds development (McMurray & Farris-Trimble, 2012; McMurray, Horst, Toscano, & Samuelson, 2009).

Second, laboratory learning studies by Maye and colleagues have documented that distributional learning can occur over a short time span. Maye et al. (2003) exposed infants to a stream of speech sounds in which VOT clustered either bimodally (two categories) or unimodally (one category) and then tested their subsequent discrimination. Eight-month-olds that received bimodally structured input discriminated tokens that straddled the center of the continuum, while those receiving unimodally structured input did not. This suggests that this short (2 min) exposure to statistically structured speech was sufficient to bias discrimination, at least immediately after exposure. Given infants' likely abilities to discriminate these tokens prior to exposure, a unimodal distribution was sufficient to collapse categories. Subsequent work demonstrated the converse, that exposure to a bimodal distribution of speech sounds helps infants separate categories they do not already have (Maye, Weiss, & Aslin, 2008). Moreover, later in development, by 10 months, infants have difficulty using distributional statistics for speech sounds not in their native language (Yoshida, Pons, Maye, & Werker, 2010), suggesting that the perceptual

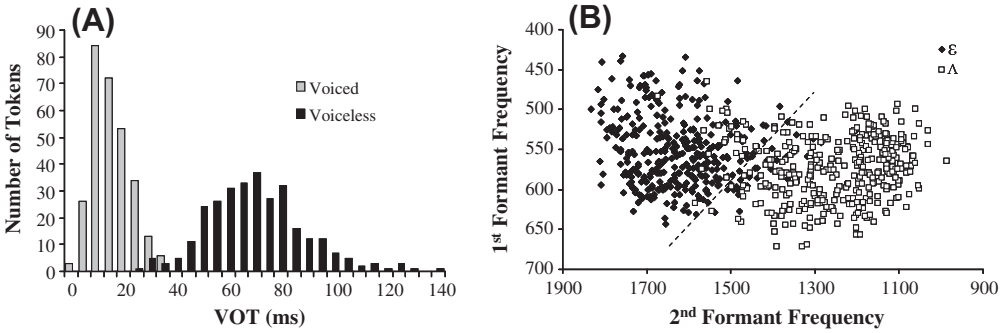

Fig. 1. The statistical distributions of various speech cues. (A) Voice Onset Time (from Allen & Miller, 1999); (B) formant frequencies for two vowels from the male speakers of Cole et al. (2010).