Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech

Elizabeth K. Johnson
Department of Psychology, University of Toronto, 3359 Mississauga Road, Mississauga, Ontario, Canada, L5L 1C6
elizabeth.johnson@utoronto.ca

Abstract: The Headturn Preference Paradigm was used to examine infants' use of prosodically conditioned acoustic-phonetic cues to find words in speech. Twelve-month-olds were familiarized to one passage containing an intended target (e.g., *toga* from *toga#lore*) and one passage containing an unintended target (e.g., *dogma* from *dog#maligns*). Infants were tested on the familiarized intended word (e.g., *toga*), familiarized unintended word (e.g., *dogma*), and two unfamiliar words. Infants listened longer to familiar intended words than to familiar unintended or unfamiliar words, demonstrating their use of word-level prosodically conditioned cues to segment words from speech. Implications for models of developmental speech perception are discussed.

PACS numbers: 43.71.Ft, 43.71.Sy, 43.71.Es [JH]
Date Received: December 20, 2007 **Date Accepted:** March 10, 2008

1. Introduction

One of the most fundamental questions in developmental speech perception is how infants learn to attend to the speech signal in an adult-like manner and perceive multiword utterances as strings of discrete recognizable words. This task is more complicated than common intuition suggests because fluent speech does not consist of clearly separated words that can be easily mapped onto lexical representations. On the contrary, reliable cues to word boundaries, such as silences between words, do not exist. And no single word is ever produced identically twice. Despite these difficulties, infants begin mastering the speech signal remarkably early.[1]

Within the first year of life, infants begin using many of the same bottom-up segmentation strategies as adults. For example, in English most content words begin with a stressed syllable, and adult listeners are appropriately biased to perceive stressed syllables as word onsets.[2] By 7.5 months, English learners have detected this pattern and are so focused on stress cues that they appear to perceive all stressed syllables as word onsets.[3] By the end of the first year of life infants' segmentation attempts become more accurate (i.e., more adult-like) as they expand their repertoire of segmentation strategies. For example, the use of phonotactic information (i.e., constraints on which consonant transitions are likely within versus across word boundaries) is thought to help infants overcome their earlier reliance on stress to find word boundaries.[1]

Another strategy that has been proposed to help infants refine their segmentation attempts is use of prosodically conditioned acoustic-phonetic information.[4,5] The prosodic structure of utterances can be characterized as hierarchical.[6] Roughly speaking, utterances contain intonational phrases, intonational phrases contain phonological phrases, phonological phrases contain prosodic words, and prosodic words contain syllables. Prosodic boundaries above the prosodic word level and above always coincide with word boundaries. Acoustic-phonetic cues mark the placement of speech units within the hierarchy. For example, speech units at the end of prosodic constituents tend to be lengthened,[7] and speech units along the onset of prosodic constituents tend to be more forcefully articulated than those situated in the middle.[8]

Adult listeners are sensitive to these acoustic-phonetic cues at all levels. They detect

syllable boundaries distinguishing potentially ambiguous phrases such as *known ocean* and *no notion*,[9] as well as phonological phrase boundaries that mark prosodic boundaries without breaking syllable boundaries.[10,11] Most importantly for the current study, adults also use acoustic-phonetic cues at the word level to infer speaker intent and segment words from speech. For example, the word *ham* is recognized more readily if it is produced as a monosyllabic word rather than as the first syllable of a longer word such as *hamster*.[12] Infants are similar to adults in that they have been shown to use utterance boundaries,[13] prosodic phrase boundaries,[4] and syllable boundaries[1] to find words in speech. However, infants have not yet been shown to use acoustic-phonetic cues at the prosodic word level to segment words from speech. Existing studies examining infants perception of syllable sequences straddling word boundaries (e.g., *taris* in *guitar is*[3]) were not designed to determine if infants are sensitive to acoustic-phonetic cues to word boundaries, and thus used stimuli containing multiple cues to word boundaries. Knowing whether infants use word-level acoustic-phonetic cues to segment words from speech would impact our understanding of developmental speech perception. For example, use of these cues could impose constraints on distributional theories of developmental word segmentation.[3,5,14]

In the current study we use the Headturn Preference Procedure to test 12-month-olds' use of word-level acoustic-phonetic cues to segment words from speech. The experiment familiarizes infants with two types of passages. One contained a reoccurring sequence consisting of a monosyllabic word followed by stress-final bisyllabic word (e.g., *toe#galore*); the other contained a reoccurring sequence consisting of a stress-initial bisyllabic word followed by a monosyllabic word (e.g., *dogma#lines*). Each of these target sequences is potentially ambiguous in the sense that it can be parsed as beginning with either a monosyllabic or bisyllabic word (e.g., *toe#galore* and *toga#lore*). If infants segment intended items more readily than unintended items, then this would suggest that infants use word-level acoustic-phonetic cues to locate word boundaries.

2. Method

Participants: Forty-eight American-English-learning 12-month-olds from the Baltimore-Annapolis region were tested (22 females). The infants had a mean age of 364 days (range: 351 days to 407 days). The data from nine additional infants were excluded for failing to complete the study due to fussiness (7), parental interference (1), and experimenter error (1).

Stimuli: Eight passages containing a reoccurring target syllable sequence were recorded in an infant-directed register by a female speaker naïve to the purpose of the study.

Sample passage containing stress-initial bisyllabic word: *This ruby quest will be more exciting than ever. Ruby quest weddings are frowned upon in Greenwich. The foosball pro was thrilled with that ruby quest. The station's ruby quest will end before ours. The guy who won the ruby quest wore silly goggles. I joined the singing ruby quest.*

Sample passage containing stress-final bisyllabic word: *This rue bequest will surely go down in history. Rue bequest weddings were common last May. The panda was really wild about that rue bequest. The station's rue bequest will make the cat happy. The guy who made the rue bequest wore shiny frog shoes. We sent a singing rue bequest.*

The speaker was asked to imagine that the nonsensical sequences (e.g. *gumbo#teak*) had meaning, and to produce them as an adjective preceding a noun in order to prevent the insertion of a phrase boundary between the syllables. The passages were on average 21.53 s long. The speaker also recorded four test lists (each containing 15 tokens of the same word).

In order to ensure the target syllable sequences in the passages were produced with different intended word boundaries, the duration of the onset and rime of each syllable of the familiarized sequences was measured. Past studies have shown that segments falling along prosodic boundaries tend to be lengthened relative to those not falling along a prosodic boundary. For example, the word *tune* is longer when it is realized as a monosyllabic word (*tune acquire*) than when it is embedded in a bisyllabic word (*tuna choir*).[15,16] In line with predictions from these studies, the first and third syllables or our target sequences were lengthened when produced as monosyllabic as opposed to bisyllabic words (Table 1).

Design: Infants were randomly assigned to familiarization with one of four pairs of

Table 1. Duration measurements of target sequences in the passages (onset/rime boundary in *rue* was difficult to locate). Duration (seconds).

	S1 Onset	S1 Rime	S2 Onset	S2 Rime	S3 Onset	S3 Rime
Ruby#quest	- ----	- ----	0.061 (0.011)	0.158 (0.025)	0.062 (0.011)	0.472 (0.095)
Dogma#lines	0.024 (0.004)	0.25 (0.032)	0.056 (0.014)	0.103 (0.027)	0.059 (0.01)	0.442 (0.252)
Gumbo#teak	0.020 (0.016)	0.208 (0.016)	0.083 (0.011)	0.175 (0.039)	0.083 (0.01)	0.239 (0.07)
Toga#lore	0.063 (0.029)	0.171 (0.024)	0.014 (0.003)	0.127 (0.022)	0.084 (0.02)	0.268 (0.112)
Rue#bequest	- ----	- ----	0.043 (0.009)	0.141 (0.014)	0.047 (0.008)	0.452 (0.145)
Dog#maligns	0.024 (0.01)	0.265 (0.029)	0.065 (0.016)	0.122 (0.039)	0.058 (0.014)	0.387 (0.162)
Gum#boutique	0.021 (0.003)	0.241 (0.025)	0.063 (0.009)	0.192 (0.039)	0.063 (0.01)	0.231 (0.096)
Toe#galore	0.109 (0.049)	0.185 (0.026)	0.028 (0.006)	0.137 (0.03)	0.078 (0.007)	0.282 (0.08)
Mean SW#S	0.035 (0.025)	0.210 (0.04)	0.054 (0.027)	0.141 (0.04)	0.072 (0.018)	0.355 (0.175)
Mean S#WS	0.051 (0.05)	0.230 (0.04)	0.050 (0.019)	0.148 (0.04)	0.061 (0.015)	0.338 (0.147)

passages: (1) *rue#bequest* and *dogma#lines*, (2) *toga#lore* and *gum#boutique*, (3) *ruby#quest* and *dog#maligns*, or (4) *toe#galore* and *gumbo#teak*. All infants were tested on the same four test items: *ruby*, *dogma*, *gumbo*, and *toga*. Two of these test items were familiar, in the sense that the syllable sequence occurred during the familiarization. However, one familiar syllable sequence spanned a word boundary while the other did not. The other two test items were unfamiliar.

Procedure and Apparatus: Infants were tested using a standard variant of the Headturn Preference Procedure.[1,3] The experimenter remained out of view of the infant, recording the direction and duration of the infants' orientation through the use of a button box. The randomization of stimuli and termination of trials was computer controlled. A red light and a speaker were mounted at eye level on the center of each side panel, and a green light was located at eye level on the center of the front panel. During the familiarization phase, the green light flashed at the start of each trial. Once the infant oriented toward the green center light, it stops flashing. One of the two side red lights then immediately began flashing. Once the infant oriented towards the flashing light, a sound file was presented from the speaker hidden behind the flashing light. The sound file continues to play until either the infant looks away for more than two consecutive seconds or the sound file ends. Once the infants accrued 45 s of orientation times towards each passage, the test phase began. Twelve test trials were presented during the test phase (three trials for each of the four test items). Test trials were presented in three blocks, and trial order was randomized within those blocks. Both the experimenter and the caregiver listened to masking music over headphones to prevent them from biasing the study.

3. Results

Mean orientation times to each of the three types of test items (familiar intended, familiar unintended, unfamiliar) were calculated for each of the 48 subjects (see Fig. 1). Thirty-three out of 48 subjects had longer average orientation times to familiar intended test items than unfamiliar test items. In contrast, only 22 out of 48 infants had longer orientation times to familiar unintended test items than unfamiliar test items. And most importantly, 31 out of 48 infants had longer average orientation times to familiar intended test items (*ruby* from *ruby#quest*) than the familiar unintended test items (*ruby* from *rue#bequest*). A mixed design analysis of variance, 3 (test item: intended familiar, unintended familiar, unfamiliar) $\times 4$ (familiarization condition), revealed a significant effect of test item, $F(2,44)=3.83$, $p<0.05$. There was also an effect of familiarization condition, $F(3,44)=31$, $p<0.05$, but importantly, there was no significant interaction between test item and familiarization condition, $F(6,44)=1.48$, $p>0.10$. Planned comparisons revealed a significant difference in orientation times to intended familiar and unfamiliar test items, $F(1,47)=6.12$, $p<0.05$. In addition, there was a significant difference in orientation times to intended and unintended familiar test items, $F(1,47)=4.36$, $p<0.05$. How-

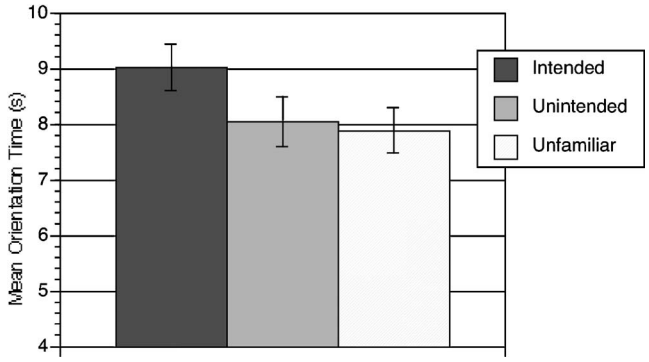

Fig. 1. Mean orientation times in seconds to intended (e.g., *ruby* from *ruby#quest*), unintended (e.g., *ruby* from *rue#bequest*), and unfamiliar test items by 12-month-olds.

ever, there was no significant difference in orientation times to familiar unintended and unfamiliar test items, $F(1,47)=0.11$, $p>0.10$. As Fig. 1 illustrates, these effects were attributable to longer orientation times to familiar intended test items ($M=9.02$ s, $SD=2.9$) than to either familiar unintended test items ($M=8.04$ s, $SD=3.1$) or to unfamiliar test items ($M=7.89$ s, $SD=2.7$).

4. Discussion

Our results demonstrate that 12-month-old infants are like adults in that they use prosodically conditioned acoustic-phonetic cues to segment words from running speech. Note that these cues constrained infants' segmentation behavior in the face of misleading syllable distribution cues to word boundaries, i.e., despite the fact that the conditional probabilities marking the transition between the two syllables of the intended and unintended words were equal. These findings suggest that sensitivity to the prosodic structure of utterances may be fundamental to early word recognition, perhaps even more so than sensitivity to syllable distribution cues. It may also be the case that simple exemplar models cannot explain the results reported in this study. Instead, infants may be like adults in that they compute the prosodic structure of utterances online and use this information to infer word boundaries.[17] These findings contribute to a growing body of evidence demonstrating that acoustic-phonetic variation plays an important role in word recognition.[10,12,17–20]

One limitation of the current study is that we have not identified which acoustic-phonetic cues infants use to detect boundaries at the prosodic word level. Recent studies have suggested that duration cues alone are sufficient for adults to perceive the boundary between prosodic words.[12,17] However, these studies did not rule out the possibility that adults are also sensitive to other prosodically conditioned acoustic-phonetic cues to word boundaries, such as degree of coarticulation or changes in consonant realization due to initial strengthening.[10,11] Thus, there remains the possibility that during the course of early language acquisition infants must learn which acoustic-phonetic cues mark boundaries in their language. It may be the case that infants learn the cues marking smaller junctures such as word boundaries by attending to large junctures such as phrase and utterance boundaries.

Another area for future research will be to identify when infants begin using word-level acoustic-phonetic cues to parse the speech stream. Doing so will be important for integrating our findings with current models of developmental word segmentation. Perhaps sensitivity to word-level acoustic-phonetic cues to word boundaries develops along with sensitivity to phonotactic and allophonic cues at around the end of the first year of life, and in combination these cues help infants overcome their over reliance on lexical stress cues to word boundaries, i.e., integration of word-level acoustic-phonetic cues with other word segmentation cues may help English learners begin extracting noninitially stressed words from speech. Another possibility

is that sensitivity to acoustic-phonetic cues begins developing even earlier, perhaps at the very earliest stages of word segmentation.[19] If this were the case, then this would have important implications for current models of developmental word segmentation. Distributional models posit that infants begin segmenting words from speech by tracking conditional probabilities between syllables, i.e., syllables pairs that are statistically likely to co-occur are likely to be words. Once infants have segmented enough words from speech by tracking conditional probabilities between syllables, then they can notice that most of the words begin with a stressed syllable.[14] Eventually, enough additional segmentation cues are learned for infants to reach an adult-like ability to extract words from speech. If infants were sensitive to prosodically conditioned acoustic-phonetic cues at the onset of word segmentation abilities, then this sensitivity would greatly constrain the use of distributional strategies to find word boundaries. Thus, it is important that future work investigate when and how infants begin detecting prosodically conditioned word boundaries in fluent speech.

Acknowledgments

I thank P. Jusczyk and the Jusczyk Lab Executive Committee for inspiring this work. I thank J. Miller, S. Shattuck-Hufnagel, A. M. Jusczyk, A. Cutler, B. Landua, and A. Seidl for invaluable help leading to the completion of this study.

References and links

[1]P. Jusczyk, *The Discovery of Spoken Language* (MIT Press, Cambridge, MA, 1997).
[2]A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Lang. **31**, 218–236 (1992).
[3]P. W. Jusczyk, D. Houston, and M. Newsome, "The beginnings of word segmentation in English-learning infants," Cogn. Psychol. **39**, 159–207 (1999).
[4]A. Gout, A. Christophe, and J. Morgan, "Phonological phrase boundaries constrain lexical access II. Infant data," J. Mem. Lang. **51**, 548–567 (2004).
[5]M. Shukla, M. Nespor, and J. Mehler, "An interaction between prosody and statistics in the segmentation of fluent speech," Cogn. Psychol. **54**, 1–32 (2007).
[6]M. Nespor and I. Vogel, *Prosodic Phonology* (Foris, Dordrecht, 1986).
[7]C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717 (1992).
[8]C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740 (1997).
[9]L. Nakatani and K. Dukes, "Locus of segmental cues for word juncture," J. Acoust. Soc. Am. **62**, 714–719 (1977).
[10]T. Cho, J. McQueen, and E. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," J. Phonetics **35**, 210–243 (2007).
[11]A. Christophe, S. Peperkamp, C. Pallier, E. Block, and J. Mehler, "Phonological phrase boundaries constrain lexical access I: Adult data," J. Mem. Lang. **51**, 523–547 (2004).
[12]A. P. Salverda, D. Dahan, and J. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**, 51–89 (2003).
[13]A. Seidl and E. K. Johnson, "Infant word segmentation revisited: Edge alignment facilitates target extraction," Dev. Sci. **9**, 566–574 (2006).
[14]E. Thiessen and J. Saffran, "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants," Dev. Psychol. **39**, 706–716 (2003).
[15]A. Turk and S. Shattuck-Hufnagel, "Word-boundary-related durational patterns in English," J. Phonetics **28**, 397–440 (2000).
[16]T. Cho and E. K. Johnson, "Acoustic correlates of phrase-internal lexical boundaries in Dutch," In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, edited by S. H. Kin and M. J. Bae (Sunjin, Jeju, Korea, 2004), pp. 1297–1300.
[17]K. Shatzman and J. McQueen, "Prosodic knowledge affects recognition of newly acquired words," Psychol. Sci. **17**, 372–377 (2006).
[18]B. McMurray and R. N. Aslin, "Infants are sensitive to within-category variation in speech perception," Cognition **95**, B15–B26 (2005).
[19]E. K. Johnson, "Speaker intent influences infants' segmentation of potentially ambiguous utterances, " Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS, 2003), pp. 1995–1998, Barcelona, Causal Productions.
[20]P. Luce and C. McLennan "Spoken word recognition: The challenge of variation," In *Handbook of Speech Perception*, edited by D. B., Pisoni and R. E. Remez, Wiley-Blackwell, Oxford, pp. 591–609 (2005).