Disambiguating durational cues for speech segmentation

Padraic Monaghan
Centre for Research in Human Development and Learning, Department of Psychology, Lancaster University, Lancaster LA1 4YF, United Kingdom
p.monaghan@lancaster.ac.uk

Laurence White
Department of Psychology, Plymouth University, Plymouth PL4 8AA, United Kingdom
laurence.white@plymouth.ac.uk

Marjolein M. Merkx
Department of Psychology, Royal Holloway University of London, Egham TW20 0EX, United Kingdom
marjoleinmerkx@hotmail.com

Abstract: Vowels are lengthened in lexically stressed syllables and also in word-final syllables. Both stress and final-syllable lengthening can assist in word segmentation from continuous speech, but in languages like English, with a preponderance of stress-initial words, lengthening cues may conflict for indicating word boundaries. An analysis of a large corpus of English speech demonstrated that speakers provide distributional information sufficient to potentially allow listeners to determine whether vowel lengthening is associated with lexical stress or word finality without relying on a congruence of multiple suprasegmental cues to make the distinction.

PACS numbers: 43.71.Es, 43.70.Mn, 43.66.Ba [AC]
Date Received: January 2, 2013 **Date Accepted:** May 23, 2013

1. Introduction

One of the first steps that a child must take when acquiring their language is the identification of word units by segmenting natural speech. This is not a simple process—pauses between words in continuous speech are rare, and most child-directed speech comprises multi-word utterances (Bernstein Ratner and Rooney, 2001). To assist in this difficult task, there is a large range of cues potentially available to the child in detecting word boundaries, including allophonic variations according to word position, co-articulation within words, and phonotactic constraints (Mattys *et al.*, 2005) as well as statistical information about transitions between syllables (Saffran *et al.*, 1996).

Prosodic cues in speech also provide useful and useable information about word boundaries to the language learner. In English, the final syllables of words are generally longer in duration than initial or medial syllables with increased lengthening effects in word-final syllables that immediately precede phrase boundaries (Wightman *et al.*, 1992). Saffran *et al.* (1996) found that adults were sensitive to vowel lengthening in word-final open syllables for promoting segmentation of an artificial language, and pre-boundary vowel lengthening has also been shown to guide the lexical interpretations of infants (Gout *et al.*, 2004). Infants showed a differential response to *paper* and *pay per*[*suades*] where the sequences were distinguished, among other potential cues, by lengthening of the pre-boundary vowel.

Lexical stress patterns are another potential prosodic source of word boundary information. In English, words tend to have initial stress (Cutler and Carter, 1987), realized as a combination of increased loudness, greater duration, and a pitch excursion—typically a relative increase in pitch—on the stressed syllable. Jusczyk, Houston, and Newsome (1999) found that 7.5-month-old infants preferentially segmented words with trochaic stress (i.e., stress on the first syllable, e.g., *kingdom*) but could not extract words with iambic stress (e.g., *guitar*) from continuous speech.

Thus both lexical stress and final lengthening are used by infants for segmentation. This raises the question of how language learners are able to interpret extended vowel duration on the one hand as a property of lexical stress, and hence suggestive of a preceding word boundary, and on the other hand as final syllable lengthening, indicative of a subsequent boundary. Learning to disambiguate these cues is critically important for infants in implementing prosody-based segmentation strategies. So, the question arises how are such durational cues realized in a manner that promotes learnability?

One possibility for disambiguation is that lengthening as a lexical stress cue can only be distinguished from boundary-related lengthening due to co-occurrence with other cues such as loudness, pitch excursion, and vowel quality, requiring the hearer to integrate durational information with other suprasegmental cues. Thus integration of multiple cues may serve to disambiguate the nature of the two durational cues. We refer to this as the receiver-work hypothesis, as the listener, or receiver, must draw together disparate sources of information (Shannon, 1948).

An alternative possibility is that the distribution of lengthening within the word or phrase is discriminatory in itself, allowing the listener to determine whether the extra duration marks lexical stress or word-/phrase-final position. In this case, the hearer can potentially rely on duration alone and does not have to combine different sources of information to determine the role of lengthening in the speaker's utterance. We refer to this as the transmitter-work hypothesis, as sufficient durational information is provided by the speaker, or transmitter, without requiring integrative work by the hearer.

There are numerous controlled phonetic studies that indicate a number of localized lengthening effects with distinct loci, such as lengthening both of stressed syllables and other syllables in phrasally stressed words (White and Turk, 2010). However, the critical question in terms of the receiver versus transmitter hypotheses is whether these individual effects are additive or interactive as influences on duration of segments. If durational effects are merely additive, then this is consistent with the receiver-work hypothesis—the listener must determine from cues other than duration alone the role of the cue with respect to word boundaries as then determining the cause of a segment's lengthening requires other information to be simultaneously established. In contrast, if durational effects are interactive, then this is consistent with the transmitter hypothesis whereby durational effects can be assigned according to their prosodic role without requiring information from other cues in the speech but can be derived from information on (relative) duration alone.

There is emerging evidence from controlled phonetic studies for the interaction of final-syllable lengthening and stress-related lengthening, as the locus of final lengthening appears to begin with the final stressed syllable in the word, and also includes subsequent unstressed syllables (White and Turk, 2010). Thus durational cues are likely to interact, rather than be additive, according to syllable position and stress position.

The task for the language learner in utilizing durational cues is, however, more complex than that suggested by controlled studies comparing the duration of similar segments in different prosodic configurations. The learner must rather induce differential distributions syntagmatically from an uncontrolled array of full and partial utterances, produced by different speakers and often in rather distinct dialects. To test availability of durational cues under these variable conditions, we analyzed a large multi-speaker multi-accent corpus of American English speech to assess the utility of distributional data for distinguishing lengthening due to lexical stress and word finality.

2. Method

The corpus analyzed was the DARPA TIMIT speech database (Garofolo *et al.*, 1993), which consists of 6300 sentences, 10 sentences each spoken by 630 different speakers from eight different dialect regions of the U.S.A. The database provides automatically tagged duration information for the onset of each segment within each utterance. We selected all disyllabic, trisyllabic, and tetrasyllabic words from the database. Syllabic stress position of the words was taken from the CMU pronouncing dictionary (Carnegie Mellon University, 2012). Words that were transcribed in TIMIT with a different number of syllables to their canonical pronunciation in the dictionary were removed from the analysis; this resulted in 929 word productions being omitted (5.4% of word tokens). Variant pronunciations tended to either omit or add short duration unstressed syllables word-medially, hence including these words would confound the comparison of stressed and unstressed syllables.

There were 11 849 disyllabic, 3809 trisyllabic, and 1443 tetrasyllabic word tokens included in the analysis. We extracted the duration of each vowel in each word. We focused on vowels given the variability in consonant duration due to both intrinsic and structural factors, such as consonant clustering, and the confounding effects of syllabification. We also distinguished words in utterance-initial, utterance-medial, and utterance-final position to separate the potential contribution of higher-level boundaries to segmental lengthening effects from lengthening effects due to syllable position within the word itself.

3. Results

The data for disyllabic, trisyllabic, and tetrasyllabic words were analyzed separately using linear mixed effects models with duration of vowel as the dependent variable. In the models, speaker and segment type were entered as random factors to account for speaker variation and also to ensure that potentially different distributions of particular vowels at certain positions within words did not affect the results. The word's position within the utterance (initial, medial, final), the syllable's position within the word (1, 2, 3, 4), the lexical stress of the syllable (unstressed, primary, secondary), and all two- and three-way interactions among utterance position, syllable position and stress were entered as fixed factors in the model. Parameter estimation was conducted using restricted maximum likelihood method. *Post hoc* tests comparing stress type for each syllable position were performed with Sidak adjustment. Utterance position was included in the analysis to ensure that any observed word-final effects were not due to utterance-level effects. However, to simplify presentation of the results, we do not report utterance position main effects or interactions, but focus on effects of syllable position and lexical stress. Estimated means and standard errors of the mean for all word lengths are shown in Table 1.

3.1 Disyllabic words

As anticipated, each of the main effects for vowels in disyllabic words was significant. For syllable position, $F(1, 22\,519) = 350.88$, $p < 0.001$, final syllable vowels were longer

Table 1. Estimated mean durations in ms (standard error of the mean in parentheses) by syllable position and stress condition (0, unstressed; 1, primary stress; 2, secondary stress) for disyllabic, trisyllabic, and tetrasyllabic words.

	First syllable			Second syllable			Third syllable			Fourth syllable		
Word length	0	1	2	0	1	2	0	1	2	0	1	2
Disyllabic	88 (6)	100 (6)	88 (7)	100 (6)	142 (6)	113 (6)						
Trisllabic	80 (6)	87 (6)	83 (7)	87 (6)	99 (6)	91 (7)	99 (6)	132 (8)	122 (6)			
Tetrasyllabic	80 (6)	84 (6)	74 (6)	86 (6)	89 (6)	82 (14)	94 (6)	98 (6)	91 (6)	101 (6)	132 (19)	129 (8)

than initial syllables. For lexical stress, $F(2,\ 22\,495) = 461.32$, $p < 0.001$, vowels in primary stressed syllables were longer than secondary stress, which were longer than unstressed, all $p < 0.001$.

Consistent with the transmitter-hypothesis, all two- and three-way interactions were also significant. For the syllable position and stress interaction, $F(2,\ 22\,616) = 161.90$, $p < 0.001$, with the difference between first and second syllable greater in primary (difference = 42 ms) and secondary (25 ms) stressed syllables than unstressed syllables (12 ms). Thus the lengthening effects of word finality are greater when the syllable carries stress than otherwise.

3.2 Trisyllabic words

Once again all main effects were significant. For syllable position, $F(2,\ 9962) = 115.24$, $p < 0.001$, final syllable vowels were longer than medial syllables, which were longer than initial syllables, both $p < 0.001$. For stress, $F(2,\ 9974) = 47.48$, $p < 0.001$, vowels in primary stressed syllables were longer than secondary stress, $p = 0.004$, which were in turn longer than unstressed, $p < 0.001$. As for vowels in disyllabic words, the interaction between syllable position and stress was significant, $F(4,\ 10103) = 18.48$, $p < 0.001$, the difference among the syllable durations was greater for primary stressed (45 ms difference from first to third syllable) than secondary (39 ms) or unstressed syllables (19 ms).

The interaction between syllable position and stress is consistent with the hypothesis that the locus of final lengthening is determined with respect to the stressed syllable, specifically, that final lengthening begins on the final stressed vowel and also affects subsequent segments (White and Turk, 2010). The effects of word position and stress location on unstressed syllables in disyllables are necessarily confounded but can be distinguished in trisyllables. To test the locus of lengthening for trisyllabic words, we measured the duration of vowels in unstressed syllables in terms of the stress context in which they occur. We performed three analyses: Comparing durations of unstressed vowels in the second and third syllables of words with primary stress on the first syllable, unstressed vowels in the first and third syllables of words with primary stress on the second syllable, and unstressed vowels in the first and second syllables of words that had primary stress on the third syllable. We predicted that syllable position in words would have an effect on duration only for syllables following a stressed syllable. For each analysis, we tested a linear mixed effects model with speaker and segment type as random effects, and utterance position, syllable position, and the interaction between utterance position and syllable position as fixed factors. Again, for clarity, we do not report the effects of utterance position.

For the primary stress on the first syllable, there was a significant main effect of syllable position, $F(1,\ 2623.92) = 162.31$, $p < 0.001$, with third syllables [111 ms (9 ms)] longer than second [93 ms (9 ms)] syllables. For the primary stress on the second syllable, there was again a significant main effect of syllable position, $F(1,\ 2398.91) = 178.02$, $p < 0.001$, with third syllables (101 ms) longer than first syllables (85 ms). For primary stress on the third syllable, there was no significant main effect of syllable position, $F < 1$, with first and second syllables of duration [90 ms (15 ms) versus 96 ms (9 ms), respectively]. These results are consistent with a locus of final lengthening beginning with a primary stressed syllable and including subsequent unstressed syllables.

3.3 Tetrasyllabic words

For syllable position, $F(3,\ 4605) = 49.01$, $p < 0.001$, fourth syllable vowels were longer than all other syllables, all $p < 0.001$. Third syllables were also longer than first syllables, $p < 0.001$, but other syllables did not differ significantly in length, all $p \geq 0.212$. For stress, $F(2,\ 4673) = 4.16$, $p = 0.016$, vowels in primary stressed syllables were marginally significantly longer than unstressed syllables, $p = 0.060$, but secondary stressed syllables did not differ in length to primary stressed nor unstressed syllables, $p \geq 0.361$.

The interaction between syllable position and stress was not significant, $F(6, 4715) = 1.57$, $p = 0.153$. The lack of significant effects is likely due to relative data sparsity compared with disyllables and trisyllables.

We repeated the follow-up analyses applied to trisyllabic words, investigating the effect of stress position on the distribution of lengthening. We distinguished vowels in unstressed syllables in words with first, second, third, or fourth syllable primary stress and assessed effects of utterance position and syllable position on duration.

For first-syllable primary stressed words, there was a significant main effect of syllable position, $F(2, 445.77) = 16.40$, $p < 0.001$, with unstressed syllables progressively longer through the word when the first syllable was stressed, second syllable [81 ms (8 ms)] was shorter than third syllable [93 ms (9 ms)], $p = 0.021$, and fourth syllable [98 ms (8 ms)], which did not differ from one another, $p = 0.704$. For second-syllable primary stressed words, there was a significant main effect of syllable position, $F(2, 1420.76) = 40.91$, $p < 0.001$, with first syllable [78 ms (6 ms)] shorter than third [89 ms (6 ms)], which was in turn shorter than fourth [103 ms (7 ms)], both $p < 0.001$. For third-syllable primary stressed words, there was a significant main effect of syllable position, $F(2, 439.15) = 7.90$, $p < 0.001$, with fourth syllable position [95 ms (8 ms)] significantly longer than second syllable [86 ms (8 ms)], $p < 0.001$, but neither were significantly different than first syllable position [79 ms (15 ms)], $p \geq 0.514$. For fourth-syllable primary stressed words, there were only two instances in the corpus.

4. Discussion

The results replicated previous separate observations of lengthening effects due to lexical stress and syllable position (e.g., Wightman *et al.*, 1992) with vowels in word-final syllables greater in duration than word-initial or word-medial syllables, and vowels in primary stressed syllables longer than unstressed syllables. Though the analyses assigned variance separately to utterance-final position to distinguish word-level effects of duration, some of the word-final syllables will also have been phrase-final, and so a neat division between the durational effects of phrase boundary and word boundary cannot be firmly established.

However, the critical question was whether these lengthening effects were additive or interactive. If effects were additive then when the hearer is exposed to a segment that is longer than usual, they cannot know without other information whether the lengthening is due to syllable position or stress, consistent with the receiver-hypothesis. If effects were interactive, then the source of lengthening can be ascertained without requiring the listener to integrate multiple cues, consistent with the transmitter-hypothesis.

For disyllabic and trisyllabic words, we found interactive effects of stress and syllable position. Furthermore, our analyses enable us to determine the nature of the interaction. Consistent with phonetically controlled studies (White and Turk, 2010), we demonstrated in a multi-speaker multi-accent corpus of American English that word-/phrase-final lengthening begins with a stressed syllable and also affects subsequent, but not preceding, unstressed syllables within the word. Thus the locus of stressed syllable lengthening and the locus of word-/phrase-final lengthening are distinct, providing the listener with information adequate to assign durational information to lexical stress alone or to the interaction between lexical stress and word /phrase finality.

The separable expression of duration for the cues entails that these multiple cues can be simultaneously available to infants learning to segment speech. This is consistent with other domains where multiple single cues are simultaneously available and used, such as for visual depth perception (Jacobs, 2002). In the case of segmentation, it has been shown that infants, before the age of 1 yr, use both stress placement (Jusczyk *et al.*, 1999) and boundary-related timing cues (Gout *et al.*, 2004) to segment speech. Our findings suggest that parallel durational sources of information may indeed be sufficient to achieve this. Thus for example, the durational profiles of both of the common syllables in …*pay persuades*… versus …*paper*… will be distinct, facilitating identification of boundaries preceding the first stressed syllable and following the first or second syllable.

Christiansen, Allen, and Seidenberg (1998) tested a simple recurrent network model of segmentation, where adding more cues to the input in terms of utterance boundaries, stress, and phonology resulted in a better match between predicting end of utterance and the word boundaries within the speech than using any single cue. The parallel use of multiple cues may be somewhat circumscribed in adult segmentation: For example, stress does not appear to be relied on for segmentation by English listeners where—in clear speech—higher-level cues are available (Mattys *et al.*, 2005). Infants lack well developed higher-level cues, however, and thus the parallel exploitation of multiple signal-derived cues would greatly facilitate segmentation and thus word learning.

The analyses are based upon adult speech produced under laboratory conditions, and it remains an open question as to how expression of the durational cues may vary in child-directed speech. We suggest that the greater variation in prosody evident in child-directed speech is likely to increase the magnitude and preponderance of the durational cues rather than change the nature of their interaction (Papousek *et al.*, 1985). Nevertheless, our corpus analysis suggests that the speaker's transmission of the signal is operationalized to enable multiple cues to be simultaneously available and without requiring additional work by the listener to determine the role of the cues. It remains an intriguing question the extent to which—for other sources of information useful for segmentation—the work is performed by the listener, by the speaker, or how the effort is subdivided, in terms of providing distinguishable multiple cues.

Acknowledgments

This research was supported by ESRC research Grant No. RES-000-22-1538 while the first and third authors were at the University of York. The third author's post at Royal Holloway has now expired. Thanks to Nick Chater and Lou Boves for helpful comments on analysis.

References and links

Bernstein Ratner, N., and Rooney, B. (**2001**). "How accessible is the lexicon in Motherese?," in *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*, edited by J. Weissenborn and B. Höhle (John Benjamins, Amsterdam), Vol. 1.

Carnegie Mellon University (**2012**). "The Carnegie Mellon University pronouncing dictionary v07a," http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Last viewed December 2012).

Christiansen, M. H., Allen, J., and Seidenberg, M. S. (**1998**). "Learning to segment speech using multiple cues: A connectionist model," Lang. Cognit. Processes **13**, 221–268.

Cutler, A., and Carter, D. M. (**1987**). "The predominance of strong initial syllables in the English vocabulary," Comput. Speech Lang. **2**, 133–142.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (**1993**). "DARPA TIMIT acoustic-phonetic continuous speech corpus," CD-ROM, NTIS order number PB91-100354.

Gout, A., Christophe, A., and Morgan, J. L. (**2004**). "Phonological phrase boundaries constrain lexical access. II. Infant data," J. Mem. Lang. **51**, 548–567.

Jacobs, R. A. (**2002**). "What determines visual cue reliability?," Trends Cogn. Sci. **6**, 345–350.

Jusczyk, P. W., Houston, D. M., and Newsome, M. (**1999**). "The beginnings of word segmentation in English-learning infants," Cognit. Psychol. **39**, 159–207.

Mattys, S. L., White, L., and Melhorn, J. F. (**2005**). "Integration of multiple segmentation cues: A hierarchical framework," J. Exp. Psychol. Gen. **134**, 477–500.

Papousek, M., Papousek, H., and Bornstein, M. (**1985**). "The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech," in *Social Perception in Infants*, edited by T. M. Field and N. A. Fox (Ablex, Norwood, NJ).

Saffran, J. R., Newport, E. L., and Aslin, R. N. (**1996**). "Word segmentation: The role of distributional cues," J. Mem. Lang. **35**, 606–621.

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656.

White, L., and Turk, A. E. (**2010**). "English words on the Procrustean bed: Polysyllabic shortening reconsidered," J. Phonetics **38**, 459–471.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717.