Journal of Memory and Language 66 (2012) 509–529

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

How does context play *a part* in splitting words *apart*? Production and perception of word boundaries in casual speech

Dahee Kim [a,*], Joseph D.W. Stephens [b], Mark A. Pitt [c]

[a] Department of Linguistics, Ohio State University, Columbus, OH, United States
[b] Department of Psychology, North Carolina A&T State University, Greensboro, NC, United States
[c] Department of Psychology, Ohio State University, Columbus, OH, United States

ARTICLE INFO

Article history:
Received 30 April 2010
revision received 10 December 2011
Available online 26 January 2012

Keywords:
Word segmentation
Casual speech
Speech production
Speech perception

ABSTRACT

Four experiments examined listeners' segmentation of ambiguous schwa-initial sequences (e.g., *a long* vs. *along*) in casual speech, where acoustic cues can be unclear, possibly increasing reliance on contextual information to resolve the ambiguity. In Experiment 1, acoustic analyses of talkers' productions showed that the one-word and two-word versions were produced almost identically, regardless of the preceding sentential context (biased or neutral). These tokens were then used in three listening experiments, whose results confirmed the lack of local acoustic cues for disambiguating the interpretation, and the dominance of sentential context in parsing. Findings speak to the H&H theory of speech production (Lindblom, 1990), demonstrate that context alone guides parsing when acoustic cues to word boundaries are absent, and demonstrate how knowledge of how talkers speak can contribute to an understanding of how words are segmented.

Introduction

Spoken language is often a continuous stream of speech. For comprehension to succeed, the listener must segment this stream into a sequence of individual words. A substantial literature has been devoted to determining the degree to which information about word-boundary locations is present in the acoustics of speech (Lehiste, 1960), or dependent upon higher-order contextual factors such as the listener's interpretation of word meaning and sentence structure (Cole, Jakimik, & Cooper, 1980). By identifying the different sources of information, their relative importance, and how they are used in combination (Mattys, White, & Melhorn, 2005; Norris, McQueen, Cutler, & Butterfield, 1997), it is thought that a comprehensive theory of word segmentation can be constructed.

The purpose of the current study is to suggest that a consideration of the talker can inform our understanding of how spoken words are segmented. Successful communication requires the listener to have learned how to segment speech from a variety of talkers (e.g., native, foreign-accented) speaking in a variety of styles (e.g., careful vs. casual speech), and for talkers to have learned to speak with enough clarity so that listeners can parse the speech stream and comprehend the message. An understanding of how talkers actually speak is thus a potentially useful source of information for describing the segmentation problem and for addressing theoretical issues about segmentation in both production and perception. We pursue this idea by examining which acoustic cues to segmentation talkers provide in high-frequency, casually-produced utterances, and how this knowledge can influence thinking about solutions to the segmentation problem.

Previous studies have shown that spoken language is rich in acoustic cues to word boundaries. Major findings from the literature include lengthening of word-initial and word final segments and syllables (Beckman & Edwards, 1990; Lehiste, 1960) and shortening of segments and syllables that are not adjacent to a word boundary (Harris & Umeda, 1974; Klatt, 1976; Lehiste, 1972; Oller,

* Corresponding author. Fax: +1 614 292 8833.
E-mail addresses: Kim.2245@osu.edu (D. Kim), pitt.2@osu.edu (M.A. Pitt).
URL: http://ling.osu.edu/~daheekim (D. Kim).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.12.007

1973). For instance, Lehiste (1960) had speakers produce English words and word sequences with boundary ambiguities (e.g., *Grade A* vs. *gray day*; *nitrate* vs. *night-rate* vs. *Nye trait*) in isolation and embedded in short sentences. Duration and spectrographic analyses revealed a consistent and considerable durational difference among word-initial, word-medial, and word-final segments, in that word-initial or word-final segments were longer than word-medial segments. Related to the lengthening of word-initial segments, articulatory strengthening that occurs at the initial position of a word or a prosodic boundary has also been extensively studied (Byrd & Saltzman, 1998; Cooper, 1991). Byrd and Saltzman (1998), for instance, compared lip movements of syllable onset /m/ in different boundary conditions (e.g., word medial *mommamia* vs. word initial *Momma–Mimi* vs. across a word boundary *Momma, Mimi*, among others), and reported that lip movements become slower when /m/ is adjacent to a word or prosodic boundary than when it is not adjacent to any boundaries. Byrd, Kaun, Narayanan, and Saltzman (2000) extended this finding by confirming that articulatory gestures "get larger, longer, and further apart" when adjacent to a boundary. Additional phonetic properties that have been reported to correlate with word boundaries include amplitude contour, allophonic realizations of segments (e.g., clear /l/ vs. dark /l/; Umeda & Coker, 1975), spectral differences of vowels (Hoard, 1966; Lehiste, 1960), and degree of coarticulation (Krakow, 1989; Redford, Davis, & Miikkulainen, 2004). Taken together, the presence of such systematic variation in speech production demonstrates that talkers readily produce word boundary cues for listeners as they speak.

Although it is uncontroversial that acoustic cues to word boundaries can be found in the speech signal, relatively little is known about whether talkers produce these cues in utterances that are likely to occur in everyday speech, and whether talkers adjust their production of these cues based on the presence or absence of linguistic context making the message more ambiguous (e.g., *The baker looked at the drawing of a plump eye* vs. *plum pie*; Lieberman, 1963; Mattys & Melhorn, 2007).

In the Hypo- and Hyper-articulation (H&H) theory of speech production, Lindblom (1990) argues that talkers adapt their speaking style to the communicative demands of the listeners. More specifically, the theory states that talkers adjust their articulatory effort and its corresponding clarity of speech according to their estimate of how difficult it is for the listener to comprehend the message. If talkers estimate that comprehension would be difficult, due to the lack of strong contextual cues or background noise for instance, they would produce hyper-articulated speech to ensure that listeners comprehend the speech. If talkers believe the message is clear and unambiguous, they would default to produce hypo-articulated speech. The consequences for word segmentation are that talkers may produce clear acoustic cues to word boundaries when they estimate that listeners require clarity, but provide far weaker boundary cues in other communicative situations.

The preceding discussion makes it clear that consideration of the talker has implications for theorizing about how listeners segment words. Most notably, if the talker is responsible for producing speech at a minimally sufficient level of clarity, as suggested by H&H theory, the communication system could be placing an undue burden on the talker. A talker may cause communication to break down, for instance, by misestimating the level of clarity the listener requires. Such failures may be prevented if the perceptual system is robust against a high degree of uncertainty and ambiguity in the acoustic signal, by being less reliant on the talker speaking clearly. To ensure successful communication, a flexible and lenient model of speech perception may be preferred. However, such a model risks making the talker irrelevant. A theory of speech perception and word segmentation can be informed by knowledge about how talkers speak, including what acoustic cues are consistently present in the speech produced by the talker. For example, if aspiration in voiceless stops is shown to occur only at word onsets, there is a good reason for listeners to segment upon hearing it. Such reliable behavior of the talker can justify its specification in a model of the listener.

Studies of how listeners segment words demonstrate that they are efficient in exploiting all available information, acoustic and contextual. Acoustic cues that have been shown to affect listeners' segmentation include the acoustics of word-onset segments (Nakatani & Dukes, 1977), allophonic variation (Christie, 1974; Gow & Gordon, 1995), fundamental frequency contour (Ladd & Schepman, 2003), amplitude contour (Lehiste, 1960), and durational patterns corresponding to prosodic boundaries, including word boundaries (Cho, McQueen, & Cox, 2007; Salverda et al., 2007).

For example, Gow and Gordon (1995) showed that listeners are sensitive to the phonetic details distinguishing phonemically identical segments at different locations within a word. Using cross-modal semantic priming, they found that listeners were faster in deciding that *kiss* is a word after hearing the phrase *two lips* than after hearing the word *tulips*, suggesting that listeners are sensitive to the phonetic difference between the two primes. Studies focusing on the processing of embedded words have reported similar findings. Davis, Marslen-Wilson, and Gaskell (2002) found greater lexical activation for monosyllabic words such as *cap* when the syllable *cap* was originally produced as a whole word (e.g., in the phrase *cap tucked*) rather than as the initial syllable of a longer word (e.g., *captain*), which later was also confirmed by eye-tracking experiments (Salverda, Dahan, & McQueen, 2003).

Contextual cues that have been shown to affect listeners' segmentation include the lexical status of words in the stimuli (Mattys et al., 2005), the plausibility of possible interpretations of ambiguous word sequences (Mattys & Melhorn, 2007; e.g., *plump eye* or *plum pie* after hearing *The baker looked at the drawing of a* ... vs. *The surgeon looked at the drawing of a* ...), phonotactic constraints (McQueen, 1998), phonological properties of the language (Cutler & Norris, 1988; Norris et al., 1997), and the preceding sentential context (Cole et al., 1980). For example, Cole et al. (1980) showed that the semantic content of a preceding narrative influences segmentation. They had listeners detect a mispronunciation as they heard the sentence *they saw the carko on the ferry*. The sentence occurred

at the end of a story that was either about an automobile boarding a ferry or about a shipment of cargo. Listeners were faster in detecting the mispronounced syllable (*ko* instead of *go*) when they heard the story about a cargo shipment than the story about an automobile, reflecting their interpretation of *carko* as a two-word phrase (*car go*) in the automobile story and as a single word (*cargo*) in the ferry story.

More recently, Mattys et al. (2005) examined how acoustic cues and contextual cues to word boundaries are integrated by placing the two in direct opposition to each other. In their Experiment 6B, listeners performed a lexical decision task by responding to either *cremate* or *mate* after hearing "An alternative to traditional burial is to *cremate* the dead." The speaker produced a pause between the two syllables of the target word (e.g., *cremate*) such that the initial phoneme of the second syllable (e.g., /m/ in *cremate*) contained cues that would lead listeners to favor interpreting the second syllable as a single word (e.g., *mate*). This syllable then replaced its corresponding token in the two-syllable target word. Although one might expect the acoustic cues to cause mis-segmentation (e.g., hearing *mate* instead of *cremate*), significant priming of the two-syllable target word was found, suggesting that contextual information, when available, overrides local cues in word segmentation. Results from subsequent studies (Mattys & Melhorn, 2007; Mattys, Melhorn, & White, 2007) have softened this conclusion somewhat by finding that the influence of context can be diminished depending on the relative strength of the local acoustics and contextual information such as sentential or lexical bias.

Returning to the theoretical questions raised above, it may seem from the literature that much of the responsibility for word segmentation does rest on the shoulders of the talker. The growing list of acoustic cues produced by talkers and used by listeners in segmentation suggests that acoustic cues to word boundaries are plentiful in speech and may be the primary source of information for segmentation. In those rare instances where acoustic cues are not available, other cues such as lexical bias can resolve an ambiguity. Indeed, models of spoken word recognition (e.g., TRACE: McClelland & Elman, 1986; Shortlist: Norris & McQueen, 2008) have shown how competition between candidate lexical items can provide correct segmentations in cases where local acoustic cues do not clearly specify word boundaries. From this perspective, context – especially broader semantic and syntactic context – serves a secondary, subordinate role in segmentation.

We wondered whether this is an accurate description of how words are segmented in everyday speech communication. Given that producing casual or hypo-speech is representative of many verbal exchanges, we hypothesized that word boundary cues in these registers might be acoustically weak or unclear. This, in turn, would call for an increased reliance on the surrounding context to ensure successful segmentation, perhaps to a degree that has not been previously realized.

The current study was a first step in exploring this idea. We approached the problem by seeking out cases of word-boundary ambiguity that are most frequently encountered in everyday speech, and then examining tokens of these utterances when produced in a casual speaking style while still maintaining experimental control. We considered both production and perception. We first analyzed talkers' productions of local ambiguities to determine what acoustic cues were produced. Then we measured listeners' segmentations of these productions to determine the contribution of acoustic and contextual information necessary to segment the sequence as intended by the talker.

Among the various contexts where word boundary ambiguities can arise, we focused on ambiguous sequences beginning with a schwa, such as *along* vs. *a long*, where the vowel [ə] could be interpreted either as the initial syllable of a disyllabic word (e.g., *along*) or as the indefinite article in a two-word phrase (e.g., *a long*). As mentioned above, our choice of utterances was motivated by the desire to study segmentation in an environment that is common in the language so as to maximize the generality of the findings. Schwa-initial utterances are arguably the most frequently encountered cases of word-boundary ambiguity in the English language. In the Buckeye corpus of conversational speech (Pitt et al., 2007), 6.6% of the words begin with [ə]. In 42% of these cases, [ə] was the realization of a monosyllabic word (e.g., *a*, *I*, *the*, *of*; in many such cases, it resulted from extreme reduction of a function word during rapid speech), whereas the remainder (58%) were cases in which [ə] was the first syllable of a multi-syllabic word. Taking the analysis a step further, in 11.8% of the schwa-initial words (0.8% of the corpus) there was a local ambiguity, in the sense that the sequence by itself could be interpreted as a single word or a two-word phrase. An important feature of these items is that they are not easily disambiguated through lexical competition because both one-word and two-word versions are lexically acceptable, though they may differ in lexical frequency.

Stephens and Pitt (2007) conducted a perception experiment using a subset of these schwa-initial tokens from the corpus. Listeners judged whether the talker said the one-word (e.g., *along*) or two-word (e.g., *a long*) version of the ambiguous sequence. When presented in isolation, listeners judged the intended two-word tokens (i.e., those for which there should be acoustic cues signaling a word boundary) as two words only slightly more often than the intended one-word tokens (40% vs. 36%). In contrast, when a portion of the original context surrounding the ambiguous sequence was provided (e.g., "*...try to get along with...*", "*...and it takes a long time...*"), the tokens were disambiguated, with intended two-word tokens being judged as such much more often than one-word tokens (74% vs. 21%).

In contrast to the primacy of acoustic cues implied by prior studies, these results suggest that context is primarily responsible for enabling listeners to segment function words like *a* in casual speech. However, because the stimuli came from a corpus of spontaneous speech, they were uncontrolled on too many dimensions to generalize the results with confidence. The current study extended this preliminary work to a more controlled environment. Experiment 1 focused on the talker. We recorded utterances containing lexically ambiguous items and measured the acoustic properties of the regions containing these items. In one condition, the preceding sentential context

biased one interpretation of the word string (e.g., *along*). In another condition the prior context was neutral. This manipulation enabled us to ascertain whether talkers traded off acoustic clarity for contextual predictability (cf. Lindblom, 1990), producing clearer cues to the presence or absence of a boundary when the context is neutral than when it is biased.

The subsequent experiments focused on the listener. In Experiment 2, we asked if the ambiguous sequences, when presented alone, could be segmented as intended by the talker. Because listeners have been shown to capitalize on a wide range of acoustic cues to segment words, we expected them to take full advantage of such cues to the extent that they were available. Experiments 3 and 4 examined which properties—semantic and syntactic biases, and intonational and rhythmic continuity, respectively—of the preceding sentential context influence segmentation of the ambiguous sequences. If our concerns regarding the word-boundary ambiguity in casual speech are borne out in the data, we expected listeners to rely heavily on contextual cues to segment the ambiguous sequences.

Experiment 1: production and acoustic analysis of ambiguous sequences

Experiment 1 investigated the production of local acoustic cues to word boundaries. Participants produced the schwa-initial ambiguous sequences (e.g., *along/a long, away/a way, apart/a part*, etc.) in two types of sentence frames, one in which both the one-word and two-word interpretations were permissible (neutral context), and one in which the sentence was biased towards one interpretation (biasing context). Productions were analyzed to determine how the one-word and two-word realizations differed, and how contextual bias affects the manifestation of cues differentiating the two versions. If talkers modulate the clarity of speech as they speak, considering the needs of the listener (Lindblom, 1990), they might produce stronger cues to the intended segmentation in a neutral context compared to these same ambiguous sequences following a context that contains strong biases toward one interpretation. Greater predictability of the sequence in the biasing context might result in the production of weaker segmentation cues.

Method

Participants

Twenty native speakers of American English (12 male, 8 female) participated; none reported speech or hearing difficulties. Each completed two 1-h testing sessions and received a 20.00 USD honorarium for participating.

Stimulus materials

Twenty schwa-initial sequences that can either be segmented as a single word (e.g., *along*) or a two-word phrase (e.g., *a long*) were used. We refer to this two-level variable as the *version* of the sequence. Sequences were selected so that both versions would be familiar to participants. In addition, using frequency counts from the New York Times annotated corpus (Sandhaus, 2008), the sequences were chosen so that the relative frequencies of the one- and two-word versions varied. For some sequences, the one-word version was more frequent; for other sequences the two-word version was more frequent.

Ambiguous sequences were embedded in two types of carrier sentences (another two-level variable referred to as *context*), yielding four sentences per sequence. In the neutral context condition, there was a single preceding context that permitted both the one-word (e.g., *adore*) and two-word (e.g., *a door*) interpretations of the schwa-initial sequence prior to being disambiguated at the end of the sentence (e.g., *The servant came to adore every puppy, The servant came to a door in the basement*). In the biasing context, the sentence frame was created so that only one interpretation was semantically and syntactically permissible prior to the onset of the schwa-initial sequence (e.g., *Lovers are meant to adore each other*, *The hallway leads to a door at the end*). Thus, the neutral and biasing conditions differed in whether the one- or two-word interpretation could be predicted on the basis of preceding context. Note also that the biasing context always predicted the *correct* interpretation of the sequence. Across the four sentences, the word immediately preceding it was held constant. This was also true of the immediately following segment whenever possible. Otherwise, sounds from the same major sound class followed the ambiguous sequence. The length (i.e., number of words and syllables) of the precursor as well as total sentence length was equated as closely as possible across the four conditions. Sentences are listed in Appendix A.

Two stimulus lists were created, each with a random ordering of 40 targets and 80 filler sentences. Fillers, which differed across lists, were similar to the target sentences in every way except that they did not contain an ambiguous schwa-initial sequence. They were included to mask the repetition of the ambiguous sequences in the lists, which was necessary because the design was completely within-subjects. Each ambiguous sequence appeared twice in a list, but in a different version-bias pairing. For example, list 1 contained *along* in a biasing prior context and *a long* in a neutral prior context. List 2 contained the other two combinations.

Procedure

Participants were tested individually in a sound-attenuated room. They sat in front of a computer monitor and wore a head-mounted microphone. Presentation software (Version 12.0, www.neurobs.com) controlled sentence presentation and recording.

On each trial, participants saw a sentence on a computer screen. One word in the sentence was replaced by Xs (e.g., *He chose to go along with the XXXXXX*). In target sentences, Xs replaced the last word, which was always after the ambiguous sequence. In filler trials, Xs replaced a word in the beginning or middle of the sentence. Participants were instructed to read and remember the sentence verbatim. After 3 s, the sentence disappeared and the Xs remained on the screen for 750 ms. Then, a word (e.g., *plan*) replaced the Xs at the same time a beep played over a loudspeaker. The word was displayed for 1000 ms, and

the beep was participants' cue to mentally replace the Xs in the sentence with the new word, and then speak the complete sentence into the microphone. Participants were given 6 s in which to repeat the sentence, and to do so before four exclamation marks appeared on the screen, which signaled the end of the trial. To approximate a communicative exchange, participants were instructed to say the complete sentence as if they were talking to a friend. Recordings were saved in individual 16-bit wav-format sound files (22 kHz sampling rate).

This procedure is a variant of one used previously to elicit casual speech (Dilley & Pitt, 2010). Because it requires most of the sentence to be held in memory, participants assume that the experiment is a test of memory, and tend not to monitor the clarity of their speech.

The two testing sessions were spaced 1 week apart to further minimize the possibility that participants would notice repetition of the ambiguous sequences. (In a post-experiment questionnaire, none of the participants reported noticing the repetitions, and none correctly guessed the purpose of the experiment.) One list was presented in each session, with the order of list presentation counterbalanced across participants. Three practice trials began each session.

Results

Speakers' sound files were annotated using the Praat speech analysis software (Boersma & Weenink, 2008). Two phonetically trained labelers marked the beginning and end of each syllable in the ambiguous sequences (e.g., syllable boundaries in [ə.lɔŋ]) as well as the beginning and end of the initial consonant of the second syllable (e.g., the onset and the offset of [l] in [ə.lɔŋ]). If the second syllable of the ambiguous sequence began with a stop, release burst onset and voice onset were marked separately to provide a complete description of durational patterns involving word junctures.

Labeling consistency was assessed between the two labelers by comparing the temporal locations of their labels in a random 10% of the tokens. Cronbach's α was .99, indicating a high degree of inter-labeler consistency. Utterances containing disfluencies (171 tokens; 10.7% of the data) were not included in the analysis, leaving 1429 usable tokens.

All acoustic measurements were made in Praat using custom scripts that detected the boundaries described above and automatically extracted acoustic measures. Thirty-four acoustic measures were made of each token (listed in Appendix B), which can be classified into three groups: duration, amplitude, and fundamental frequency.

Duration

Phonemically identical segments and syllables may show different durational patterns depending on where they occur within a word (Klatt, 1975; Lehiste, 1973). More specifically, segments and syllables at word boundaries—at the beginning or at the end of a word—tend to be longer than those that occur in the middle of a word (Turk & Shattuck-Hufnagel, 2000). Based on this finding, one might expect that the schwa in two-word versions (e.g., [ə] in *a long*), which begins and ends a word, would be longer than the schwa in one-word versions (e.g., [ə] in *along*), which does not end the word.

Similarly, Cooper (1991) reported that the application of word-initial lengthening is not limited to syllables and that word-initial segments tend to be longer than word-medial segments. Therefore, we might expect the initial segment of the second syllable to be longer when it begins a word (e.g., [l] in *a long*) than when it is in the middle of a word (e.g., [l] in *a long*). One might further expect that two-word versions, overall, would be longer than one-word versions.

We tested these hypotheses by measuring the duration of the following strings: entire ambiguous sequence, schwa, second syllable, and the initial segment of the second syllable. The acoustic measurements were then fitted using a linear mixed-effects model in the lme4 package in R (Baayen, Davidson, & Bates, 2008; Bates & Maechler, 2009; R Development Core Team, 2009). Version and bias were fixed factors and talker and sequence (item) were random factors. Statistical significance was assessed using pMCMC in the languageR package in R (Baayen, 2008).

Shown in the left graph of Fig. 1 are the mean durations of the two versions (collapsed over the context conditions) in all four analyses. In all cases, measurements of the two-word versions were slightly longer than the one-word versions, with the largest difference being no greater than 10 ms. Nevertheless, as the 95% confidence intervals suggest, the distributions were sufficiently narrow to yield three statistically significant differences. There were significant effects of version for the entire sequence duration ($\beta = 9.23$, pMCMC = 0.01), schwa duration ($\beta = 3.45$, pMCMC = 0.01), and duration of the initial segment of the second syllable ($\beta = 6.25$, pMCMC = 0.00). The duration of the second syllable did not differ reliably across versions ($\beta = 5.38$, pMCMC = 0.11).

The preceding comparisons were performed using a measure of absolute duration. It is possible that the small trends could be altered, possibly amplified, by using a relative measure of duration, which would control for differences in speaking rate. Three of the four analyses, except overall duration, were rerun using proportional measurements in which the millisecond duration of the subsequence (e.g., schwa) was divided by the duration of the entire ambiguous sequence. The duration of the second syllable ($\beta = -0.89$, pMCMC = 0.03) and the duration of the initial segment of the second syllable ($\beta = 1.29$, pMCMC = 0.00) yielded reliable differences across versions. The effects were again exceedingly small. The duration of the second syllable was, proportionally, 0.26% longer in the one-word productions (77.59%) than in two-word productions (77.33%). The duration of the initial segment of the second syllable was 1.29% longer when it began a new word (two-word versions, 23.92%) than when it was in the middle of a word (one-word versions, 22.63%).

For completeness, two additional relative duration measures were made, the ratio of schwa duration to the duration of the second syllable, and the ratio of schwa duration to the duration of the initial segment of the second syllable (Christie, 1977). Analyses revealed no significant effect of version.

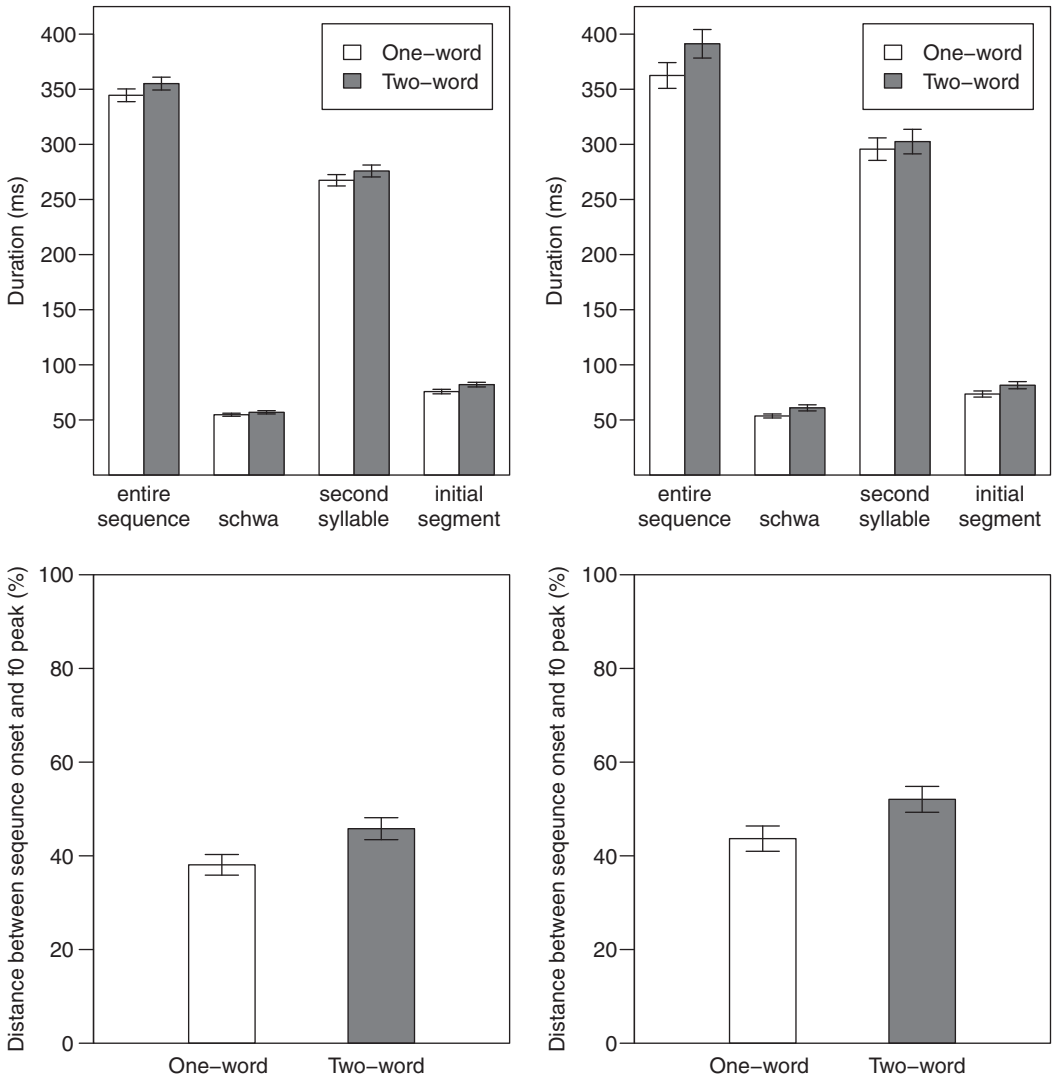

Fig. 1. Graphs of duration and f0 measures in Experiment 1 (left side) and in speech corpora (right side). The top row contains four mean duration measures of the one-word and two-word versions of the ambiguous sequences. The bottom row contains the temporal location of f0 peak measured from word onset. Error bars represent 95% confidence intervals.

Overall, extensive duration analyses identified only puny differences in duration measures between the one-word and two-word versions. The most reliable finding was a tendency for the initial segment of the second syllable to be longer when it began a word, which is an instance of domain-initial strengthening (Cooper, 1991; Son & Van Santen, 2005).[1] Prior context had virtually no effect on duration measures. Only the analysis on the duration of the second syllable in milliseconds yielded a significant effect of context ($\beta = -8.05$, pMCMC = 0.02), with duration of the second syllable being longer when the ambiguous sequence was produced after a neutral precursor. No analyses yielded a reliable main effect of context or an interaction of version and context.

These results generalize those of Turk and Shattuck-Hufnagel (2000), who in their duration analysis of word boundaries, showed that a boundary between a function word and a content word is weaker than a boundary between two content words, in the sense that durational adjustments at word boundaries, such as boundary-final lengthening, are only occasionally manifested at the boundaries adjacent to a function word. The lack of clear durational cues differentiating one-word and two-word productions may be accounted for by the syntactic properties of the stimuli.

Amplitude and fundamental frequency

The presence or absence of a word boundary can be signaled by non-durational cues, which include metrical stress pattern (Cutler & Butterfield, 1992; Cutler & Carter, 1987), fundamental frequency (Davis et al., 2002), and

[1] Another way in which durational differences might emerge is by the speaker shortening the closure duration or voice onset time (VOT) when producing intervocalic stops (Redford, 2007). To test if this occurred, closure duration and VOT were measured after the schwa in each token containing a stop (e.g., *apart*) across all four sentences. Statistical analyses yielded no reliable differences.

fundamental frequency contour—temporal alignment of f0 maxima and minima (Dilley, Ladd, & Schepman, 2005; Ladd & Schepman, 2003; Welby, 2007), among others. To determine whether speakers differentiated the one-word and two-word versions along these dimensions, amplitude and fundamental frequency measures were examined.

Eight amplitude measures were extracted from each token: root mean square (RMS) amplitude of ambiguous sequence, RMS amplitude of the first syllable, RMS amplitude of the second syllable, amplitude minimum and maximum of each syllable in the ambiguous sequence (four measures), and the amplitude of the initial segment of the second syllable. Again, mixed models were used to evaluate differences in measurement across version and bias conditions. No effects of version or interactions of version and bias were reliable for any of the measures.

In contrast, effects of context were statistically significant on seven out of eight measures (individual analyses are in Appendix C in Supplementary data). Across the board, the results show that amplitude measures were higher in the ambiguous sequence when it followed the neutral than biasing context. For example, mean amplitude of schwa was 76.97 dB in the neutral condition and 75.52 dB in the biasing condition ($\beta = -1.27$, pMCMC = 0.00). As in the duration analyses, differences between conditions were small and variability was impressively low. The only measure in which the effect of context was non-significant was minimum amplitude.

Five fundamental frequency (f0) measures were extracted from each token: mean and maximum f0 of the ambiguous sequence, mean f0 of each syllable, and maximum f0 of the second syllable, which bears primary stress. Across all analyses, only that on the mean f0 of the second syllable yielded a significant effect of version ($\beta = 4.33$, pMCMC = 0.01). The second syllable had a higher mean f0 when it was an independent word (144.01 Hz, *long* in *a long*) than when it was part of a longer word (140.35 Hz, *along*). Although version did not interact with context, effects of context were obtained in four of the f0 measures except for the mean f0 of the second syllable. Ambiguous sequences following a neutral precursor had higher mean and maximum f0s than sequences following a biasing context. An additional measure, the ratio of the maximum f0 to the mean f0 of the entire ambiguous sequence, was not differentiated by version or bias.[2]

Two additional measures of f0 were made. They were prompted by the results of Ladd and Schepman (2003), who found that English speaking listeners used the temporal location of pitch minimum and maximum to resolve word-boundary ambiguities: sequences with an earlier pitch peak were more likely to be interpreted as one word than two words. The two measures were quantified as the temporal location of f0 maximum relative to the onset of the sequence, and the temporal location of f0 maximum relative to the onset of the second syllable. Both absolute and relative measures were made, yielding four comparisons. All yielded significant effects of version and context, with intended one-word sequences having an f0 peak earlier in the word than intended two-word sequences (lower left graph in Fig. 1; absolute: $\beta = 28.91$, pMCMC = 0.00; relative: $\beta = 9.05$, pMCMC = 0.00). Sequences following a neutral precursor had an f0 peak earlier in the sequence (39.67%) than sequences following a biasing precursor (44.26%; absolute: $\beta = 16.35$, pMCMC = 0.03; relative: $\beta = 6.01$, pMCMC = 0.00). No interactions were significant.

As with the duration analyses, only a few small differences were found between versions in amplitude and frequency. In contrast, context yielded robust effects across multiple amplitude and frequency measures, all in the direction showing that contextual predictability leads to reduced articulatory effort or hypo-articulation (Lindblom, 1990) as indicated by lower amplitude and fundamental frequency measures.

Discussion

The results of Experiment 1 show that productions of the one-word and two-word versions of the ambiguous sequences are highly similar. Of the 34 acoustic measures, only a few yielded reliable effects.[3] Two stand out by being reliable by both absolute and relative measures. One is the duration of the initial segment of the second syllable. Although small in size, it was very consistent. The other is the temporal location of the f0 peak, which shifted noticeably later in time in the two-word realization compared to the one-word realization.

In the preceding analyses, the contribution of each acoustic cue in distinguishing versions was examined individually. Multiple cues working in concert might be able to disambiguate the versions more decisively. That is, if many cues co-varied across versions, the two realizations could be much more distinct than the above analyses suggest. We tested this possibility by performing a logistic regression analysis on the data, treating the acoustic measures as variables to predict the classification of a token as either one word or two words. Talker was treated as a random variable. The best-fitting model yielded five reliable predictor variables: temporal location of f0 maximum relative to the onset of the sequence (measured in percentage), temporal location of f0 maximum in milliseconds, temporal location of f0 maximum relative to the onset of the second syllable, schwa duration in milliseconds, and schwa duration relative to the entire sequence duration. Among these five predictors, three accounted for slightly more

[2] Additional acoustic measures, formant values and perceptual cues to glottalization, were made on each token. Formants were also compared across the one-word and two-word versions. In one analysis, *F*1 and *F*2 were measured at the midpoint of sequence-initial schwa. In another, *F*1, *F*2 and *F*3 were measured at the midpoint of the segment at the beginning of the second syllable. No analyses reached statistical significance. In addition, we examined whether there were perceptual cues to glottalization (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Redi & Shattuck-Hufnagel, 2001), and found no reliable differences across version or context conditions, probably because glottalization of schwa would mark a boundary between a word preceding the ambiguous sequence (e.g., to in *the man was sent to accompany/a company*) and the schwa. In addition, the stimuli were short, and not likely to contain syntactic or prosodic boundaries that are large enough to induce glottalization between the precursor and the ambiguous sequences.

[3] Figures of the measures that were not reliably distinguished by version (e.g., amplitude) can be found on the first author's webpage: ling.osu.edu/~daheekim/a_segmentation.

than 1% of the variance: temporal location of f0 maximum relative to the onset of the sequence in percentage ($r^2 = 0.02$), temporal location of f0 maximum in milliseconds ($r^2 = 0.01$), temporal location of f0 maximum relative to the onset of the second syllable ($r^2 = 0.01$). Only 5.67% of the variance was accounted for by the best-fitting model ($r^2 = 0.06$), indicating that even when linearly combined, acoustic differences alone cannot effectively distinguish one-word from two-word realizations.

These results help inform the H&H theory (Lindblom, 1990). Although context clearly affected the realization of the ambiguous sequence, altering amplitude and pitch, it did not differentially affect realization of word boundary cues in the two versions. For example, *along* was pronounced no differently than *a long* regardless of whether it was embedded in a neutral or a biasing sentence. This lack of a difference could be due to the fact that an indefinite article and a word following it often belong to the same prosodic and syntactic unit, making the word boundary between them very weak (Turk & Shattuck-Hufnagel, 2000). It could also be attributable to speakers failing to notice the ambiguity, and thus not taking steps to clarify the intended segmentation, which they could easily have done. A related possibility is that the lack of a real listener reduced either talkers' sensitivity to the ambiguity or the apparent necessity of reducing the ambiguity.

To address the latter possibility, the 34 acoustic properties that were measured on the production data were measured on ambiguous schwa-initial tokens from the Switchboard corpus of telephone conversation (Godfrey & Holliman, 1997) and the Buckeye Corpus of Conversational Speech (Pitt et al., 2007). These corpora were selected because the speech materials were generated from an actual interaction involving at least two interlocutors speaking American English. A total of 829 tokens containing the same 20 schwa-initial ambiguous sequences were sampled from 581 talkers. On average, there were 23 one-word and 22 two-word tokens for each ambiguous sequence in this new dataset, although some were much more frequent that others.

Statistical analyses of the acoustic data yielded results much like those of Experiment 1. One-word and two-word versions differed on some of the same measures. These included duration measures (e.g., schwa duration, segment duration) and f0 contour measures (e.g., temporal location of f0 peak relative to the onset of the sequence or the onset of the second-syllable). More generally, when measurement differences were found between versions, they tended to be small and fairly similar in magnitude to those obtained in the production experiment. Two representative examples are shown in the right side of Fig. 1. In the top graph are the four measures of absolute duration. They are strikingly similar to those in the left graph (production experiment) in both overall magnitude and differences as a function of version. As in Experiment 1, the duration of the entire sequence was longer when it was spoken as two words than as one word, although the difference only approached statistical significance in a linear regression analysis ($\beta = 16.18$; $p < 0.07$), probably because of the enormous variability introduced by so many talkers and contexts. In the bottom graphs are the mean temporal locations of peak f0 measured from word onset. As in the production experiment, the peak occurs later in time when the sequence was produced as two words, and the difference was similarly marginal ($\beta = 9.43$; $p < 0.07$).

If the presence of an interlocutor significantly influenced how talkers produced the ambiguous sequences, a logistic regression analysis using acoustic measures with the largest differences between one-word and two-word versions, as in Experiment 1, should predict with high accuracy the talkers' intended realization of each token. Unfortunately, classification accuracy was just as poor as that found in the production experiment, yielding an almost identical *r*-squared value (0.07 vs. 0.06 in Experiment 1).

These results suggest that the presence of an interlocutor minimally influenced how the ambiguous tokens were realized (cf. Scarborough, Brenier, Zhao, Hall-Lew, & Dmitrieva, 2007 for a comparison between the acoustics of speech directed to an imagined vs. a real listener). Trading one limitation (experimental control in Experiment 1) for another (truly conversational speech), the lack of reliable acoustic cues in both data sets provides converging evidence that talkers do not typically disambiguate these sequences in a casual speaking environment, at least not using only acoustic cues. The fact that speech clarity was not affected by this situational factor (i.e., presence or absence or an interlocutor and communicative exchange) identifies a potentially informative boundary condition for H&H theory. We return to this issue in the General Discussion.

Experiment 2: perception in isolation

The results of Experiment 1 suggest that speakers in our production task only minimally differentiated the intended segmentations of the ambiguous sequences. It may be that the small differences observed are nonetheless sufficient for listeners to perceive the intended segmentation. There could also be other, more salient cues that were overlooked and not measured. We addressed these issues in Experiment 2 by excising the ambiguous sequences from their sentences and having listeners judge whether they were one word or two words. Accuracy should be high if either of these possibilities is correct.

Method

Participants

Forty new (i.e., did not participate in Experiment 1) native speakers of American English, none of whom reported speech or hearing difficulties, participated for partial course credit.

Stimulus materials

The ambiguous sequences were excised from the 1429 sentences in Experiment 1. Onsets and offsets of the ambiguous sequences, identified spectrographically and auditorily, were spliced at zero crossings, and all tokens were equated for loudness. Fig. 2 contains example waveforms and spectrograms of tokens of one-word and two-word versions of an ambiguous sequence.

Design

As in Experiment 1, version was crossed with context to yield the same four conditions: neutral one-word, neutral

two words, biased one-word, and biased two words. Stimuli were divided as equally as possible among four lists using the following criteria. All four versions of each sequence occurred with equal frequency in each list. Each sequence was spoken by the same talker only once, and all talkers occurred similarly often. Imbalances across list, which were minor, were caused by the exclusion of sequences with disfluencies. There were an average of 357 trials per list, and 10 participants heard each list.

Procedure

Participants were tested in groups of up to four at a time in sound-attenuated rooms. They wore headphones and sat in front of a computer screen and a seven-button response box. Stimulus presentation, which was randomized in each testing session, and response collection were controlled by personal computer.

On each trial, participants heard an ambiguous sequence. Immediately after the sound file ended, the two interpretations of the sequence were printed on the screen, the one-word version on the left and the two-word version on the right. The left-most button on the response box was labeled "one word" and the right-most was labeled "two words." Participants had 2.5 s to press the button corresponding to their rating. The experiment began with a 12-trial practice session.

Results and discussion

Participants' responses were coded using the numbers one through seven, with one indicating high certainty that the token was one word and seven indicating high certainty that the token was two words. Of interest was whether there were differences as a function of version and the context from which the versions were extracted, which might indicate the presence of perceptual cues for disambiguating the sequences.

Performance in the four conditions is shown in the upper left graph of Fig. 3. Listeners were exceedingly poor at judging the intended segmentation of the ambiguous sequences, with all four means hovering just below the midpoint of the rating scale, indicating a slight bias to respond *one word*. Two-way repeated measures ANOVAs were performed on the subject and item means, and yielded no reliable effects (version: $F1(1,39) = 1.34$, *ns*; $F2(1,19) = 0.71$, *ns*; context: $F1(1,39) = 1.22$, *ns*; $F2(1,19) = 3.39$, *ns*; interaction of version and context: $F1(1,39) = 0.28$, *ns*; $F2(1,19) = 0.15$, *ns*).[4]

[4] Because the few cues that differentiated the one-word and two-word versions tended to be small in magnitude, we wondered whether listeners could detect these subtle differences after significant practice, as listeners can learn to attend to talker-specific word boundary cues (Smith, 2004). A training version of Experiment 2 was therefore carried out with a subset of the stimuli, in which listeners judged whether an ambiguous sequence was one or two words, with feedback on the accuracy of their performance provided after every trial. Seven listeners were tested for 1 h per day for four consecutive days. There were 400 trials per day (5 ambiguous sequences × 4 versions of each sequence × 4 talkers × 5 repetitions of each token). Mean classification accuracy on day 4 (52.32%) was no better than on day 1 (52.95%), reinforcing the conclusion that acoustic differences in the one-word and two-word realizations were too slight to aid segmentation.

Even though listeners could not differentiate intended one-word and two-word realizations, it is still possible that local acoustic cues informed their judgments. We addressed this possibility by performing a linear regression analysis on the data, treating the acoustic measures from Experiment 1 as variables to predict the mean rating score of each of the 1429 tokens. The best-fitting model accounted for 16.12% of the variance. Only three variables accounted for more than 3% of the variance: entire sequence duration ($r^2 = 0.11$); schwa duration ($r^2 = 0.03$); and the duration of the second syllable ($r^2 = 0.03$), all of which indicate that longer sequences tended to be rated as two words. Related to this trend is a one-word response bias. Perhaps because a majority of the stimuli were perceived as being short, listeners exhibited a slight bias to respond one-word more often than two words: listeners were 5.77% more likely to respond using buttons 1 and 2 (35.34%) than 6 and 7 (29.58%). Cutler and Butterfield (1990b) reported a similar one-word preference and suggested that presenting an item in isolation would lead listeners to hear it as a single word rather than as a two-word phrase.

The purpose of Experiments 1 and 2 was to determine whether local acoustic cues were present *and* perceptible in ambiguous schwa-initial sequences. Extensive analyses of duration, amplitude, and frequency measures across one-word and two-word versions yielded few differences, suggesting the realizations would be difficult to tell apart. The results of the rating experiment confirm this hypothesis, as listeners' ratings differed little regardless of their intended pronunciation. Together, the data show that casual productions of ambiguous schwa-initial sequences are indistinguishable on the basis of local acoustic information.

Experiment 3: integration of sentential context and local acoustic cues

It is reasonable to infer from the results of Experiment 2 that listeners rely on non-local cues (i.e., those not between *a* and the following word) to resolve the word boundary ambiguity posed by the schwa-initial sequences. Semantic and syntactic cues are likely ones, but it is also possible that interpretation of local acoustic cues is influenced by other acoustic cues (e.g., intonation) in the preceding context. For example, Pollack and Pickett (1964) found that the ability to identify a short spoken word improved when it was presented along with subsequent words from the sentence from which it was excised, even though listeners knew the identity of the subsequent words. Durational and intonational properties of the following words aided identification of a preceding word, with those immediately adjacent having the strongest influence. We addressed both types of contextual influences in Experiments 3 and 4.

The purpose of Experiment 3 was to explore the interaction of contextual information and local acoustics in the interpretation of the ambiguous sequences. Mattys et al. (2007) did just this when they investigated how listeners integrate local acoustics and contextual information to segment words. In their Experiment 1, participants listened to sentences that ended with a near-homophonous

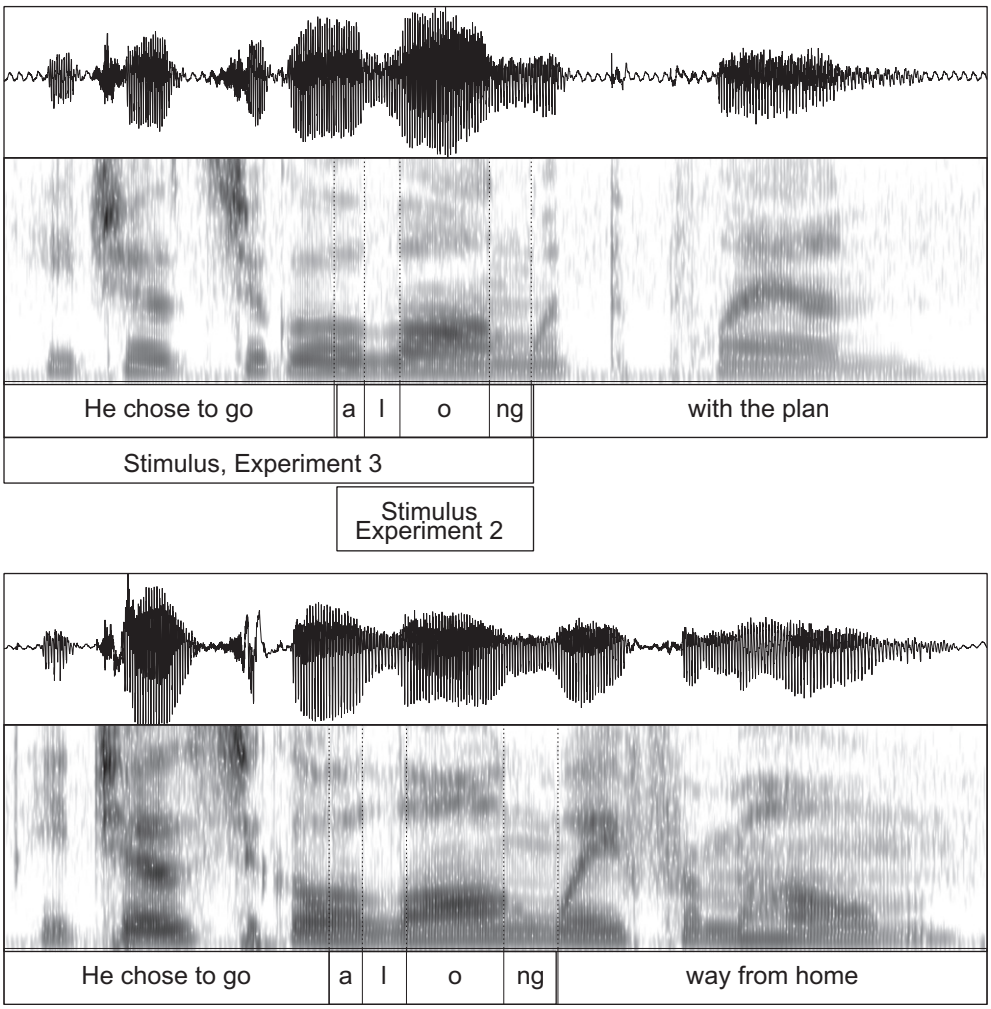

Fig. 2. Waveforms and spectrograms of "along" and "a long" in the neutral prior context conditions from one talker.

phrase (e.g., *plump eye* or *plum pie*) and indicated which word (e.g., *eye* or *pie*) they heard at the end of the sentence. The stimuli were created by concatenating different sentential contexts with different renditions of homophonous phrases that differed in the strength (mild vs. strong) of the acoustic cues signaling the word boundary. Results suggested that listeners' word segmentation was a function of both the strength of local acoustics and the sentential context. A semantically clear sentential context biased listeners to segment the phrase as the alternative consistent with the context, but only when acoustic cues were weak. When they were strong, sentential influences were much weaker.

Extrapolating these results to the present study, one would predict that the type of precursor (neutral or biased) would have a strong influence in how listeners interpret an ambiguous schwa-initial sequence. When the precursor is biased towards one of the two interpretations of a sequence, context might dominate segmentation no matter what version the talker produced (one-word or two-word). Given the acoustic and perceptual similarity of the one-word and two-word versions, this outcome would not be surprising. In the absence of strong sentential constraints (e.g., neutral precursor), ambiguity could again persist, but only if interpretation of the ambiguous sequence is not influenced by information in the surrounding sentence in other ways. For example, relative syllable timing and duration, or other cues such as fundamental frequency or amplitude contour, could partially disambiguate interpretation of the sequence.

We tested these hypotheses by examining listeners' interpretation of the ambiguous sequences with their original precursors and when cross-spliced into the context favoring the alternative interpretation of the sequence (e.g., replacing *a door* with *adore* after the precursor *The hallway leads to*...). Listeners again rated whether the sequence was one word or two. When the sequences were appended to a biasing precursor (the following context was removed), original or cross-spliced, we expected mean ratings to be similar and show a strong bias for the contextually appropriate interpretation. With a neutral precursor, the outcomes were more difficult to predict. If any perceptual differences in the realizations of the sequences are enhanced when embedded in their original contexts, rating

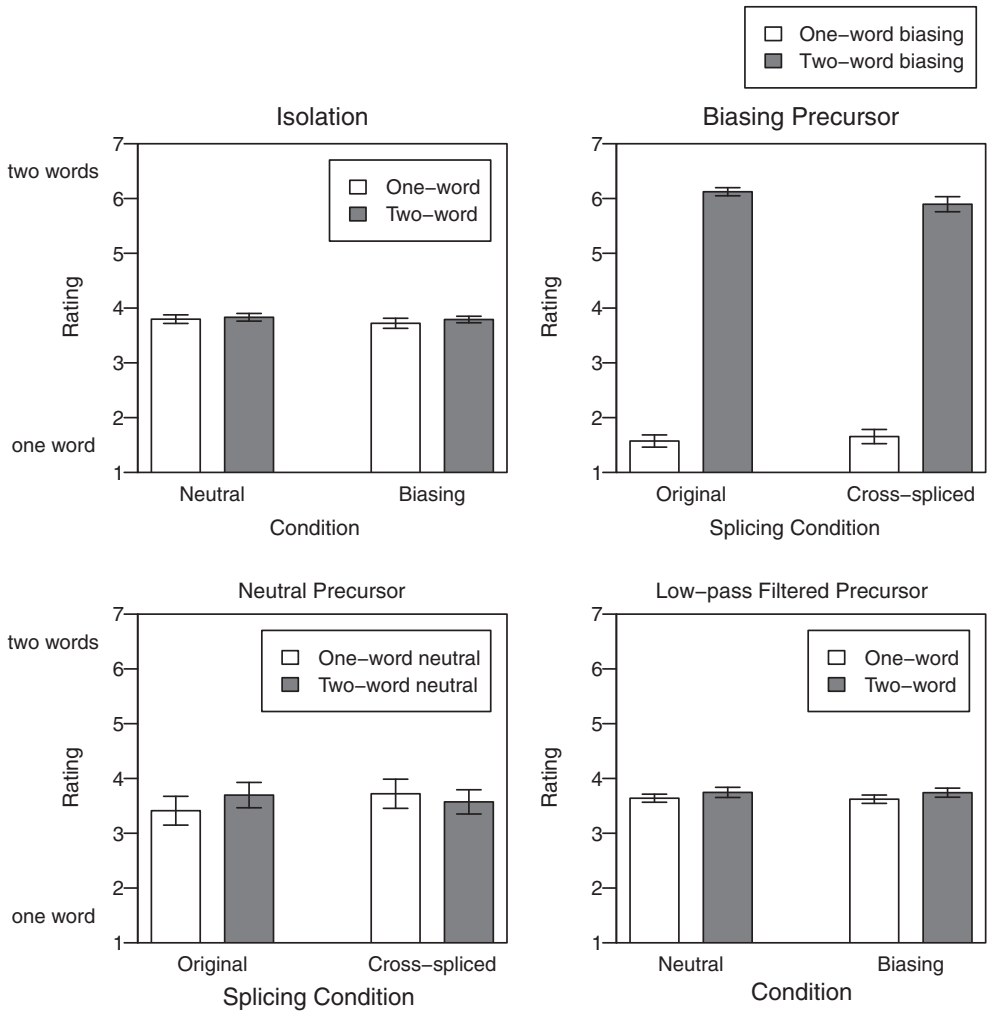

Fig. 3. Graphs of mean ratings in Experiments 2–4. Data from Experiment 2 are in the top left graph. The data of Experiment 3 are in the top right (biased precursor) and bottom left (neutral precursor) graphs. The bottom right graph contains the data of Experiment 4. Bars represent Item errors.

differences between them could emerge. Differences could also emerge in the cross-spliced contexts if contextual cues and local acoustic cues clash perceptually.

Method

The methodology was the same as in Experiment 2 except for the following changes.

Participants

Eighty new participants were drawn from the same pool as Experiment 2. Half heard the original stimuli and the other half the cross-spliced stimuli.

Stimulus materials

Productions from the beginning of the sentence to the end of the ambiguous sequence were excised from the 1429 sentences in Experiment 1 and served as stimuli for the original context condition. To create cross-spliced stimuli, the ambiguous sequence (e.g., one-word intended) and the corresponding prior context (sentence onset to onset of the ambiguous sequence) were excised from their original sentences. Excised ambiguous sequences were appended to the precursor in which the alternative version (e.g., two-word) was intended. Sequences were swapped within the same context conditions. For example, the neutral precursor produced with the one-word version was combined with the two-word version spoken with the neutral precursor. Three listeners rated the naturalness of each cross-spliced sound file to identify cases with discontinuities. Only cross-spliced sound files that were rated as acceptable by all raters served as stimuli (1118 sentences). For all tokens, precursors and the ambiguous sequences were spliced at zero crossings. These items were also equated for loudness.

Procedure

On each trial, participants heard a single phrase and rated the phrase-final ambiguous sequence as either one word or two on the same seven-point rating scale used in Experiment 2.

Results and discussion

The data were analyzed as in Experiment 2. Only responses to the 1118 stimuli that were common to the original and cross-spliced conditions were analyzed. Mean ratings when the ambiguous sequence occurred after the biasing precursor are shown in the upper right graph of Fig. 3, with the data from the original (unspliced) context on the left and the cross-spliced context on the right. The biased precursor completely dictated how the a-initial sequences were interpreted. When the context biased the one-word interpretation (white bars), it did not matter whether the sequence was the original one-word version or the cross-spliced two-word version. In both conditions, responses were strongly one-word biased, with mean ratings near 1.6 and at least 77% of responses being ratings of 1 or 2. When the context biased the two-word interpretation (gray bars), responses were strongly two-word biased, with mean ratings near 6.0. Statistical analyses confirmed what is evident in the graph, yielding a reliable main effect of biasing context (one-word vs two-word), $F1(1,79) = 767.81$, $p < 0.001$; $F2(1,19) = 531.10$; $p < 0.001$. The main effect of splicing (original vs. cross-spliced) was not reliable ($F1(1,79) = 2.08$, *ns*; $F2(1,19) = 1.44$, *ns*).

If non-local acoustic cues affected responding, contextual biases should be weaker in the cross-spliced condition because local and non-local acoustic information should clash. A comparison of the magnitude of the biasing effect (gray bar–white bar) across the two splicing conditions trended in this direction, with the difference being smaller in the cross-spliced than original condition (4.55 vs. 4.25). However, the interaction of biasing context and splicing was reliable only by items, $F1(1,79) = 2.05$, $p = 0.16$; $F2(1,19) = 6.43$; $p < 0.05$. The reason for the discrepant statistical outcomes is that subject variability is much greater than item variability. Listeners differed in how they used the entire rating scale, with some willing to make end-point responses frequently whereas others were hesitant to select an extreme value as often. Across splicing conditions, the changes in mean ratings, a .08 increase in the one-word biasing condition and a .22 decrease in the two-word biasing condition, were not great enough to reach statistical significance (one-word biasing: subjects: $t(79) = 0.73$, *ns*; items: $t(19) = 1.75$, *ns*; two-word biasing: subjects $t(79) = -1.69$, *ns*; items: $t(19) = -1.99$, *ns*).

When the precursor was neutral (lower left graph in Fig. 3), the results also hint at subtle perceptual differences between the two versions of the ambiguous sequence. When the sequences were embedded in their original contexts, the data pattern was qualitatively the same as that found in the corresponding condition when the precursors were biased (upper right graph), but much smaller in magnitude, with ratings being 0.29 higher when a two-word token occurred with its original two-word precursor than when a one-word token occurred with its original one-word precursor. In contrast, a 0.13 difference in the opposite direction was present in the cross-spliced condition, where the mean rating for tokens appended to a one-word neutral precursor was larger than for those appended to a two-word precursor (i.e., ratings for the intended two-word items were still higher than for intended one-word items despite the switch in context). Statistical analyses showed that the interaction of context (one-word vs. two-word) and splicing was marginally reliable, $F1(1,79) = 3.74$, $p < .06$; $F2(1,19) = 3.88$; $p < .06$. Tests of simple effects on the two means in each splicing condition approached reliability only in the original condition, due in part to the large amount of variability (subjects: $t(39) = 1.66$, $p < 0.10$; items: $t(19) = 1.79$, $p < 0.08$; cross-spliced condition: subjects: $t(39) = 1.05$, $p < .30$; items: $t(19) = 1.19$, $p < .25$. In the overall ANOVA, neither the main effect of version nor the main effect of splicing were reliable (version: $F1(1,79) = 0.25$, *ns*; $F2(1,19) = 0.21$, *ns*; splicing: $F1(1,79) = 1.17$, *ns*; $F2(1,19) = 0.21$, *ns*).[5,6]

The fact that appending the ambiguous sequences to their original precursors had the effect of disambiguating them even slightly suggests that non-local acoustic cues influenced segmentation. The failure to find the same effect in the cross-spliced condition is more difficult to interpret. The mismatch between local and non-local acoustic cues caused by cross-splicing could cause fragile contextual influences to disappear, but it could also be a sign that a neutral precursor does not contain cues to disambiguate the sequences, yielding results that are similar to what was found when the sequences were presented in isolation (Experiment 2).

To summarize, context, whether it was neutral or biasing, original or cross-spliced, affected segmentation of the one-word and two-word versions of the ambiguous sequences in two ways. A biasing sentential context completely determined how listeners interpreted the ambiguous sequences, making intended one-word and intended two-word versions virtually interchangeable. In addition, for both the biasing and neutral precursors, slight differences were found between the one-word and two-word versions when in their original context, suggesting that non-local acoustic cues influenced segmentation. In the neutral context condition, we found that the intended one-word and two-word sequence ratings were reliably different from each other only after the original (unspliced) context but not after the cross-spliced context, though intended two-word sequences appended after a one-word precursor was rated slightly higher than intended one-word sequences appended after a two-word

[5] Readers might wonder whether response speed differed across the original and cross-spliced conditions. Mean RTs, measured from offset of the ambiguous sequence, were surprisingly close in the neutral context (original: 1599 ms; cross-spliced: 1594 ms) and in the biasing context (original: 1240 ms; cross-spliced: 1278 ms). Statistical comparisons indicated the effects of splicing were not reliable (neutral context: $t(74.95) = -0.24$, $p = 0.81$; biasing context: $t(76.92) = 0.45$; $p = 0.66$).

[6] A reviewer expressed concern that mixing the biasing and neutral stimuli in Experiment 3 might have caused listeners to pay less attention to speech acoustics and more to meaning, and thereby mitigated any effects of acoustic cues on segmentation. To address this concern, we ran a version of Experiment 3 in which participants ($N = 24$) responded to stimuli in the neutral precursor condition only. Performance patterned just like that in the lower left graph of Fig. 3, and the statistical outcomes were also similar in suggesting that the broader acoustic context minimally aided segmentation. The main effect of version was reliable only by items, $F2(1,19) = 5.01$, $p < .04$, and the interaction of splicing and version reached significance only by items, $F2(1,19) = 7.94$, $p < .01$. There was no main effect of splicing.

neutral precursor (see Footnote 5). Because the same strings of words were in the one-word and two-word neutral precursors, these small effects might be due to intonational or rhythmic continuity between the original precursor and the ambiguous sequence. These results suggest that non-segmental cues such as pitch contour, amplitude contour, or speech rate might be present in the original (i.e., unspliced) context to assist listeners in partially distinguishing the two versions. However, their small size and weak statistical support makes this interpretation rather tenuous. It therefore seemed prudent to test this hypothesis directly, which was the purpose of Experiment 4. The original precursors were low-pass filtered to remove segmental and semantic information about the words in the precursor but retain intonation and rhythmic information. If these nonlocal acoustic cues assist in disambiguation (cf. Pollack & Pickett, 1964), listeners' ratings should differ depending on what version was intended.

Experiment 4: intonational cues to segmentation

Method

The methodology was the same as in Experiment 3 except for the following changes.

Participants

Forty new participants served.

Stimulus materials

The stimuli were those used in the original, unspliced condition of Experiment 3. Only the unspliced stimuli were used because our goal was to assess whether contextual cues that were originally produced with the corresponding ambiguous sequences would aid segmentation. A cross-spliced context would provide cues that were incongruent with the segmentation of the ambiguous sequence. Listeners heard the one-word and two-word versions in the neutral and biasing contexts. Precursors were low-pass filtered at five times the mean f0 of each sound file. Average cut-off point for the low-pass filtering was 693 Hz. The amplitude of the ambiguous sequence was attenuated by 65–80% to equate the amplitude of the low-pass filtered context with that of the unfiltered ambiguous sequence. To ensure that segmental information was not intelligible in the low-pass filtered precursor, 24 listeners transcribed the words in the contexts. None identified more than the occasional word.

Results

The data are plotted in the lower right graph in Fig. 3. If nonlocal intonational and rhythmic cues influenced segmentation in Experiment 3, there should be a main effect of version, with ratings being lower in the one-word than two-word versions. Although the means are quite close across conditions, this data pattern is present and of comparable magnitude in both the neutral (.11) and biasing contexts (.12). Although the differences in means across conditions were quite small, less than half the size of that found with the neutral precursor in Experiment 3, the variability is small enough and the effect of version is consistent enough across participants and items to just reach statistical significance in both analyses, $F1(1,39) = 6.51$, $p < 0.05$; $F2(1,19) = 5.98$, $p < 0.05$. There were no other reliable outcomes (context main effect: $F1(1,39) = 0.02$, *ns*; $F2(1,19) = 0.05$, *ns*; version by context interaction: $F1(1,39) = 0.00$, *ns*; $F2(1,19) = 0.01$, *ns*).

Discussion

The results of Experiments 3 and 4 indicate that segmentation of the ambiguous schwa-initial sequences is determined almost exclusively by the semantic and syntactic information in the precursor. In Experiment 3, the biasing context overrode any local acoustic cues signaling a word boundary, causing listeners to perceive two-word versions as one word and vice versa. In the neutral context, small and occasional differences emerged between the one-word and two-word versions, suggesting that other properties of the context aided segmentation. These trends prompted us to examine whether intonational or rhythmic information in the context influenced interpretation of the ambiguous sequence. Although miniscule, that a similar data pattern emerged after low-pass filtering the precursor indicates that intonational information in the precursor influenced segmentation of the ambiguous sequences.

The large effects of sentential bias were expected on the basis of prior work. The small but significant main effect of version in Experiment 4 reinforces recent findings demonstrating the role of non-segmental properties of prior context on segmentation. For example, Dilley and McAuley (2008) showed that listeners' interpretation of ambiguous two-word phrases (e.g., *note bookworm* or *notebook worm*) differed as a function of the metrical pattern (e.g., alternation between high and low tone) of the syllables preceding the ambiguous sequences. Results of their experiments suggest that prosodic characteristics of the preceding materials allow listeners to group syllables into words, thus affecting how they segmented phrases into words. Results of Experiment 4 appear to confirm the role of non-segmental, intonational and rhythmic contour even when the prior context contained no segmental information, and thus was unintelligible to listeners.

These results lead to the question of what acoustic properties in the filtered precursor listeners used to interpret the ambiguous sequences. We addressed this question by performing a linear regression analysis on the data, treating acoustic measures from the precursors as predictors of mean rating scores. The measures included the duration and f0 of the word preceding the ambiguous sequence (e.g., *go* in *He chose to go along…*), duration and f0 of the entire precursor (e.g., *He chose to* go in *He chose to go along…*), and duration and f0 of the ambiguous sequences in proportion to the duration of the word preceding the ambiguous sequence. Only two variables accounted for more than 3% of the variance. The first of these was the duration of the ambiguous sequence divided by the dura-

tion of the word preceding it ($r^2 = 0.03$). Listeners rated ambiguous sequences that were relatively longer than the duration of the preceding word as more two-word like. The other factor was the duration of the ambiguous sequence ($r^2 = 0.07$). As in Experiment 2, listeners had a tendency to report longer ambiguous sequences as two words.

General discussion

Verbal communication involves talkers and listeners. Knowledge of how talkers speak can help define the problem that a model of the listener must solve. We applied this perspective to obtain a better understanding of how listeners segment spoken words. We examined high-frequency cases of word-boundary ambiguity produced in a casual style, in order to observe how acoustic and contextual cues contribute to speech segmentation. Knowing that a casual pronunciation style can blur word boundaries (Cutler & Butterfield, 1990a, 1990b), it was of interest to determine the relative contribution and reliability of both types of cues in determining word boundaries, a problem that was magnified by an accompanying lexical ambiguity. The specific context we examined was the case of schwa being interpreted either as an indefinite article or as the initial syllable of a two-syllable word. We studied these ambiguities within existing data sets of conversational speech (Switchboard Corpus and Buckeye Corpus) as well as a novel set of utterances elicited in Experiment 1, which, although not truly conversational because a listener was not physically present, nonetheless approximated well a casual and arguably naturalistic speaking style.

Experiment 1 provided a controlled set of utterances that formed the basis for our acoustic analyses and perceptual experiments. It was important that these productions be representative of utterances that would be spontaneously produced in casual conversation. Experiment 1 elicited casual productions from talkers in a paradigm that focused attention away from pronunciation by superficially appearing to test working memory. Despite the absence of a real listener in this task, we expected that speakers' sensitivity to word-boundary ambiguities might still be reflected in their productions and modulated by context. Surprisingly, across 20 stimuli produced by 20 talkers in both neutral and biasing contexts, acoustic differentiation of the ambiguous sequences was minimal. Talkers did not adjust their productions of schwa-initial sequences as a function of version or contextual predictability. Of the few differences that were found, they were so small that even when combined, the cues were poor predictors of the version intended by the talker. The same tendencies were found in an analysis of one-word and two-word tokens in the Switchboard and Buckeye corpora, again, suggesting that the presence or absence of a listener does not necessarily alter how a person produces ambiguous sequences.

The absence of any effective local acoustic cues to segmenting the ambiguous sequences was confirmed in Experiment 2, where listeners provided identical perceptual ratings of the one-word and two-word sequences in isolation.[7] The results of Experiments 3 and 4 showed that the primary determiner of segmentation was the syntactic and semantic bias in the precursor. The two versions of an ambiguous sequence were interchangeable when preceded by a biasing context, producing highly similar outcomes when cross-spliced or embedded in their original contexts. Slight differences were found between versions when preceded by a neutral context. Although the source of some of these differences appears to be the prosody beyond the local environment, the contribution of prosody to disambiguating the sequence was small.

Taken together, the findings suggest that the burden of disambiguating schwa-initial utterances frequently rests with the listener and can depend almost exclusively on context. Talkers rarely provided usable acoustic cues to word juncture. Across talkers and word sequences, the stability and consistency in pronouncing the two versions of each ambiguous sequence were remarkable. There was little talker or item variability, nor were there any signs of cue trading relations, whereby a low value of one cue (e.g., duration) might be compensated for by a high value on another cue (e.g., later f0 peak). In the absence of reliable acoustic cues, listeners resolved the ambiguities in the only way possible, by relying on context. In the paragraphs that follow, we briefly explore the reasons for this very lopsided state of affairs, and the implications of the results more generally.

Availability of acoustic cues for segmentation

The presence of only minute acoustic differences between one-word and two-word versions of schwa-initial utterances contrasts with some previous studies that found reliable acoustic differences between alternative pronunciations (Davis et al., 2002; Nakatani & Dukes, 1977; Salverda et al., 2007). Two reasons for the discrepancy are that the present study focused on casual speech and used utterances involving function words.

As mentioned previously, we made a significant effort in Experiment 1 to elicit speech that approximated casual speech as closely as possible while obtaining specific, repeated utterances from talkers. This approach has not been emphasized in many other studies, where the experimental setup was not conducive to eliciting casual speech, either because the talker was knowledgeable about the purpose of the experiment or the materials or instructions would likely prompt talkers to speak clearly (e.g., reading aloud). The acoustic consequences of such decisions should not be underestimated, as acoustic (particularly durational) cues to word boundaries are more reliable in deliberately clear speech than in more casually-produced speech (Cutler & Butterfield, 1990a, 1990b; White, Wiget, Rauch, & Mattys, 2010). Further, our extensive analyses of conversational

[7] Similar production results in Experiment 1 and in the speech corpora might lead readers to wonder whether listeners could distinguish one-word and two-word realizations taken from the corpora. Recall that multiple pilot experiments using tokens from the Buckeye corpus (see Introduction) prompted this study. To recapitulate, when presented in isolation, listeners did not distinguish the two versions. Sentential context helped listeners to disambiguate and interpret the ambiguous sequence as intended by talkers.

corpora buttress the claim that casual speech fails to include reliable acoustic cues for word boundaries. We do not intend to claim that the speech elicited in Experiment 1 is a definitive model for all casual productions, given that substantial variation in hypo- and hyperarticulation is certain to occur within casually-produced utterances. Likewise, the schwa-initial utterances examined here comprise only a fraction of the potential word-boundary ambiguities or lexical embeddings that could occur in conversational English. Nonetheless, we argue that the utterances produced in Experiment 1 represent a form of word-boundary ambiguity that is likely to be encountered more frequently in everyday environments than those examined in previous studies.

Within this context, it is interesting to consider that talkers in the current study were not inclined to compensate for the informativeness of the prior context (i.e., by producing cues that disambiguated the two alternatives in the neutral context but not in the biased context). Productions of the ambiguous sequences were comparable in both contexts. Thus, the results of the present study suggest talkers do not adjust production of word-boundary cues based on the communication demands associated with semantic or syntactic ambiguity. Although the current production task incorporated no listener and therefore no real communicative exchange, it would still be reasonable to predict that talkers' sensitivity to semantic and syntactic ambiguity should extend to these utterances. Thus, the results of Experiment 1 do not suggest that talkers are unable to provide acoustic cues to resolve ambiguities. However, to the extent that talkers actually make such adjustments in everyday speech, their tendency to do so may be driven more by situational factors (e.g., background noise, feedback from the listener, etc.).

A second, production-oriented, reason for the relative lack of acoustic cues that differentiated the alternatives is the presence of a function word (the indefinite article *a*) in the two-word utterances. A production study that examined similar phrases to those used here (Turk & Shattuck-Hufnagel, 2000) found much weaker acoustic markers for boundaries between function and content words than for boundaries between two content words (e.g., *tune a choir* vs. *tuna choir*; although, again, it may be noted that their production task was not as oriented toward casual productions as Experiment 1). Thus, it might be argued that indefinite articles and subsequent content words (e.g., nouns and adjectives in the current study) form functional units, and that talkers thus make little effort to separate them. In this case, the perceptual task of disambiguating schwa-initial sequences (e.g., *appear* vs. *a pier*) would be more similar to the task of disambiguating homophones (e.g., *pier* vs. *peer*) than to that of locating word boundaries. Whether or not the lack of acoustic segmentation cues found in the present study is specific to function words (or indefinite articles more specifically), this lack of local information still has important implications for speech perception in everyday environments given the frequency of *a* and other function words that can be highly reduced in casual speech (in some cases also reduced to schwa; Bell et al., 2003).

Implications for H&H theory

Recall that in Experiment 1, the effects of sentential bias on production of the ambiguous sequences manifested themselves in predictable ways in amplitude and frequency measures. Sequences produced after a neutral precursor had higher amplitude and higher f0 than sequences produced after a biasing precursor. This outcome is consistent with H&H theory (Lindblom, 1990), which suggests that speakers adapt to communicative demands and adjust their pronunciation accordingly. In this case, talkers spoke with less effort given the higher predictability of the ambiguous sequence. By this same reasoning, one would have expected talkers to have generated distinct realizations of the two versions when the sequences were embedded in neutral sentences. That this did not occur suggests that there are additional factors influencing the production and clarity of speech other than potential for misinterpretation of the utterance (Bard et al., 2000).

One may hypothesize that talkers in Experiment 1 did not compensate for acoustic cues to schwa segmentation because they were not aware of the ambiguity. It may be unreasonable to expect talkers to monitor for all cases of potential ambiguity that listeners might encounter. To do so could be highly resource demanding, with talkers having to analyze their speech thoroughly at every linguistic boundary in the utterance. Talkers may therefore not rectify ambiguities that they are not aware of (Allbritton, McKoon, & Ratcliff, 1996; Cutler, 1987; Dell & Brown, 1991). This does not mean that cues to word juncture would not be produced when speaking ambiguous sequences, but rather that their clarity could vary greatly, being dependent on other properties of the communicative situation, such as sentential focus or external factors prompting the talker to articulate clearly (e.g., noisy surroundings).

An alternative interpretation of the findings is that the two versions were not differentiated because, in the eyes of the talker, subsequent context disambiguated them. By this account, the current data do not qualify H&H theory, but instead are another example of the talker minimizing articulatory effort in communication. One concern with this explanation is that it begins to make the talker omniscient with respect to signal clarity, knowing at an astonishing level of detail, and monitoring on the fly, when more or less acoustic information must be provided to ensure accurate comprehension by the listener. If this is true, responsibility for successful communication rests squarely with the talker.

Regardless of the accuracy of these two viewpoints, the results of this study are intended to push H&H theory forward by focusing on implementational details. The production data, both from Experiment 1 and the corpora analysis, are in good agreement in showing that talkers minimally differentiate the two realizations. That said, given the many talker, listener, and environmental factors that can influence a communicative exchange, it would be a gross overextension of the current findings to suggest that talkers never do so. Clearly they can and will when the situation demands it. However, it may be that in most exchanges and in light of the current perception results,

talkers simply do not need to adjust their articulations because, given the context, listeners successfully disambiguate the alternatives in the vast majority of situations. In short, the equilibrium in speech clarity that is established between talker and listener is modulated by listeners' extensive use of contextual information.

Implications for speech perception

The results of Experiments 2–4 indicate that correct interpretation of the ambiguous sequences depends almost entirely on semantic and syntactic context, and not on acoustic cues present within the tokens themselves. Even prosodic information carried by the sentences in which they were embedded played an exceedingly limited role. Such an outcome can give the impression that listeners were completely insensitive to acoustic differences across tokens, but this was not the case. For example, regression analyses on the results of Experiment 2 revealed a reliable influence of overall stimulus duration on participants' responses, such that longer sequences were more likely to be classified as containing two words. However, to the extent that listeners relied on overall duration, this cue did not assist them in correctly interpreting the sequences (Experiment 2). Further, when sentence contexts were provided in Experiment 3, this reliance on overall duration was abandoned in favor of higher-order information, even when the preceding sentence context did not suggest a particular interpretation of the ambiguous sequence (neutral context).

Additional evidence demonstrating that listeners used whatever contextual information was available to interpret the ambiguous sequences comes from a regression analysis of responses to the items in neutral sentence contexts in Experiment 3, using as predictors a number of properties of the phrases, such as which alternative (*along* vs. *a long*) is more frequent. The results showed that participants' responses were influenced by only two factors, and then only weakly: first, participants sometimes favored interpretations in which the phrase ending in the ambiguous sequence formed a complete sentence (e.g., *He chose to go along* vs. *He chose to go a long*; $\beta = .66$, $p < .01$); and second, participants favored interpretations in which the ambiguous sequence plus the preceding word was a more frequent phrase in spoken English, $\beta = 1.35$, $p < .01$ (e.g., *to acquire* is more frequent than *to a choir*).

Results such as these suggest that like acoustic cues, listeners use contextual information very efficiently. Because talkers rarely speak without any linguistic or situational contexts, context (i.e., successfully segmented words) is a highly dependable and useful source of information with which to locate word boundaries, so much so that the recognition system capitalizes on it fully. The results of the current study take this idea a step further and suggest that context can be relied on exclusively in segmentation. It is sensible that the recognition system is designed to be so flexible, given that speech clarity can be highly variable and unpredictable. It is not just the clarity of the speaker that the recognition system must contend with, but also the distortions caused by competing auditory events in the environment (e.g., noise, music). Successful segmentation means not just exploiting all available information, but being sufficiently adaptable to make accurate inferences based on partial or ambiguous information. To put this point another way, listeners rarely ask talkers to speak more clearly. Why? One answer can appeal to talkers providing sufficient acoustic cues to locate word boundaries. The current data demonstrate an alternative answer, which is that context can suffice in the absence of such cues.

The strength of these conclusions depends on the representativeness of the current data, so it is reasonable to wonder whether the experiment would replicate with other short function or content words. This is an open question, as our knowledge of speech acoustics is limited largely to clear speech. We do not know how often acoustic cues are sufficient for correct segmentation, let alone what all of the cues are. A casual speaking style has the potential to place a considerable burden on the listener, but this assumption needs to be validated by measuring the strength of cues at word junctures and then measuring the degree to which they clarify word boundaries, just as we did in the current study. Given that *a* is arguably the most frequent and one of the shortest embedded words in English, there is little reason to doubt the representativeness of the findings.

That said, it is also an open question whether the current results generalize to other types of word boundary ambiguities, such as when a phoneme is perceived as word-initial instead of word-final, changing the words that were heard (*plum pie* vs. *plump eye*). It will undoubtedly depend on the clarity with which acoustic cues signal the segment's position in the word. Segments whose realizations are highly contextually conditioned (e.g., stops) probably demarcate word boundaries more reliably than segments that are realized similarly across contexts. Our stimuli are clearly most similar to the latter type. Another consideration is that our experimental results are based on a very specific situational context (i.e., the production task). Although the findings are buttressed by analyses of spoken language corpora, it is still reasonable to assume that acoustic realizations of word boundaries will vary systematically with situational demands. Such differences could even be found within the corpora if the appropriate factors could be identified and analyzed.

An implication of the current findings that we wish to emphasize is not that acoustic cues to word boundaries are less important than previously realized, but rather that context is likely more influential than previously realized. Models of spoken word recognition (Grossberg, Boardman, & Cohen, 1997; McClelland & Elman, 1986; Norris & McQueen, 2008) rightly give priority to bottom-up cues in processing; only lexical entries that match the speech input are activated. The clarity of speech cues is assumed to vary, and contextual information (i.e., activated lexical candidates) biases processing of the input. The current results suggest that when it comes to segmentation, models may need to be more flexible in terms of using contextual information that is more distantly removed from the word boundary than the current lexical items. This is not necessarily difficult to achieve in many model architectures (e.g.,

connectionist, Bayesian), but the challenge lies in implementing the ability to use broader windows of lexical, semantic, and syntactic context so as to ensure successful segmentation. As described in the Introduction, segmentation in the absence of acoustic word-boundary cues can often be achieved in models via lexical competition; however, lexical competition may not suffice in situations such as the one examined here. For example, although not a model of segmentation, the behavior of TRACE (McClelland & Elman, 1986) nicely illustrates the difficulty of segmenting words using only lexical knowledge, including lexical competition. Phonemes are position-independent representations in the TRACE model, activated by a set of static phonetic features distributed over time. Activated phonemes excite matching lexical entries, including competitors. By presenting pairs of short words to the model as an uninterrupted string of phoneme inputs (i.e., without the model's word-boundary symbol), the segmentation behavior of the model can be seen (Frauenfelder & Peeters, 1990). For example, lexical candidates that match continuously from the start of the speech input are favored, making the model strongly biased to yield one-word responses. Given the ambiguous sequence *apart*, just as listeners were in Experiment 2, the one-word version will receive the greatest amount of activation, although *a* and *part* are activated at different points in time because of their complete overlap with the input. Given the two-word version *a part*, model behavior can be a bit variable. The two words will receive the greatest amount of activation most of the time, but situations can be identified in which the one-word version does so at some point during the simulation. Thus, in cases where more than one lexical interpretation is possible, a broader range of contextual information must be considered. Of course, competition from other activated entries can significantly modulate this process. The use of lexical knowledge can be quite useful, but the complexity of the segmentation problem likely requires more than this much of the time. The right mix of contextual cues is crucial, and those for one situation probably generalize only partially to another.

The current findings certainly do not negate the results of prior studies that have found acoustic cues for segmentation are produced by speakers and that listeners are sensitive to them. Clearly listeners do exploit fine-grained phonetic cues that signal a word boundary. Listeners' ability to take advantage of subtle acoustic cues has been shown across languages and paradigms (Cho et al., 2007; Salverda et al., 2003; Shatzman & McQueen, 2006; Spinelli, McQueen, & Cutler, 2003). In contrast to time-sensitive measures, such as eye tracking, the rating task used in the present study may lack the sensitivity necessary to observe subtle perceptual influences of any systematic acoustic differences that might exist between the one-word and two-word realizations. Thus, without such fine-grained data in the current study, it would be incorrect to conclude that contextual biases completely obliterate the influence of acoustics in the cognitive processes that underlie word segmentation. Nonetheless, the lack of acoustic influences on more coarse measures of perception is in many ways the crux of the current findings, because any acoustic differences did not translate in a systematic way.[8] Still, we acknowledge that a complete account of word segmentation must include specification of the time-dependent interplay between acoustics and context in the processing that lead up to the decision of where to place a word boundary.

Finally, our results might seem as though they are in conflict with the metrical segmentation strategy (MSS; e.g., Cutler & Norris, 1988), in which strong syllables are assumed to be word-initial given their predominance in this position in English. Because the stress pattern was constant (i.e., weak–strong) across all of the schwa-initial utterances, the MSS would predict a strong two-word bias for these stimuli when presented in isolation, which was not what was found. Despite this incorrect prediction, we are reluctant to argue against the MSS, which was not proposed as a single, comprehensive theoretical account of segmentation, but as one that captures a phonological regularity in language that has been repeatedly shown to bias segmentation (Cutler & Butterfield, 1990a, 1990b, 1992; Cutler, Dahan, & van Donselaar, 1997). Listeners in the present experiment simply did not rely on metrical stress when responding, perhaps because they realized early on it was not a useful cue.

The metrical properties of the stimuli nonetheless have implications for the interpretation of the current findings. The weak-strong pattern seen in frequent words such as *along* and *around* contradicts the dominant English pattern of strong initial syllables in content words; thus, to the extent that listeners use the MSS to segment speech, the correct interpretation of these words requires some other information or process that can override the MSS's tendency to erroneously insert a word boundary before the second syllable (e.g., Mattys & Samuel, 1997). We might have expected acoustic cues to play such a role, and provide bottom-up evidence for a one-word interpretation of these words. Instead, the current study found no reliable acoustic basis for determining whether a word boundary was present, which again raises the issue of the reliability of acoustic cues in segmenting casual speech. More generally, the results may suggest that acoustic cues are equally unimportant in other instances of lexical embedding where the stress pattern is weak-strong (e.g., *tar* in *guitar*).

Conclusion

In the quest to understand how spoken words are segmented, by considering how talkers usually speak, we can define the acoustic landscape with greater precision and thereby better define the problem facing listeners along with possible solutions. In a frequently-occurring situation where listeners encounter a potential word boundary and could benefit from disambiguating acoustics, talkers do not provide reliable cues, even when the preceding context is semantically ambiguous. As a consequence, listeners appear to rely primarily on context to ensure accurate segmentation.

[8] The one exception to this was listeners' tendency to classify the longest-duration stimuli as two words. This response bias was overridden by the presence of a precursor, and more importantly, did not facilitate hearing the ambiguous sequences as intended by the talker.

Acknowledgments

We thank Anouschka Bergmann for significant help in conducting Experiment 1, and Michael Tat, Erin McBurney, Erica Haugtvedt, Victoria Hoover, Morgan Albert, and Emily Vance for help testing participants. We also thank the reviewers for the excellent feedback that greatly improved the manuscript. This work was supported by research Grant DC004330 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Appendix A. Stimuli

Ambiguous sequence	Version	Bias	Sentence
	1-word	Neutral	The man was sent to *accompany* a young lady
accompany	2-word	Neutral	The man was sent to *a company* in New Orleans
a company	1-word	Biasing	The bodyguard had to *accompany* a celebrity
	2-word	Biasing	Employees are attracted to *a company* of high standards
	1-word	Neutral	John was sent to *acquire* new skills
acquire	2-word	Neutral	John was sent to *a choir* near Kentucky
a choir	1-word	Biasing	The collectors wanted to *acquire* new items
	2-word	Biasing	The singers belong to *a choir* near Kentucky
	1-word	Neutral	They said it was *across* the street from here
across	2-word	Neutral	They said it was *a cross* they saw in church
a cross	1-word	Biasing	He took the bridge and was *across* the river quickly
	2-word	Biasing	On the altar there was *a cross* that looked golden
	1-word	Neutral	The young girl had *acute* kidney disease
acute	2-word	Neutral	The young girl had *a cute* kitten in her arms
a cute	1-word	Biasing	The doctor said he had *acute* colon cancer
	2-word	Biasing	The little princess had *a cute* collection of dolls
	1-word	Neutral	The servant came to *adore* every puppy
adore	2-word	Neutral	The servant came to *a door* in the basement
a door	1-word	Biasing	Lovers are meant to *adore* each other
	2-word	Biasing	The hallway leads to *a door* at the end
	1-word	Neutral	She didn't know what *affair* Dave was involved in
affair	2-word	Neutral	She didn't know what *a fair* deal would be
a fair	1-word	Biasing	The adulterer asked what *affair* Diane was talking about
	2-word	Biasing	The judge should know what *a fair* deal would be
	1-word	Neutral	The people thought *ahead* to be prepared
ahead	2-word	Neutral	The people thought *a head* could be buried there
a head	1-word	Biasing	They made plans and thought *ahead* to the future
	2-word	Biasing	That's more hair than I thought *a head* could have
	1-word	Neutral	The man went to *align* the car's wheels
align	2-word	Neutral	The man went to *a line* there on the floor
a line	1-word	Biasing	To straighten the wheels you need to *align* them correctly
	2-word	Biasing	These points belong to *a line* that runs diagonally
	1-word	Neutral	I think Jane has *allowed* Sue to go
allowed	2-word	Neutral	I think Jane has *a loud* singing voice
a loud	1-word	Biasing	The government has *allowed* so much corruption
	2-word	Biasing	The rock singer has *a loud* sound system
	1-word	Neutral	He chose to go *along* with the plan
along	2-word	Neutral	He chose to go *a long* way from home
a long	1-word	Biasing	To fit in you just go *along* with the crowd
	2-word	Biasing	Reliable cars can go *a long* way without fixing
	1-word	Neutral	The teenager came to *amaze* all the spectators
amaze	2-word	Neutral	The teenager came to *a maze* in the corn field

Appendix A. (*continued*)

Ambiguous sequence	Version	Bias	Sentence
a maze	1-word	Biasing	The magician liked to *amaze* all his fans
	2-word	Biasing	Trained rats were taken to *a maze* in the lab
	1-word	Neutral	They have been *apart* ever since the argument
apart	2-word	Neutral	They have been *a part* of our community for years
a part	1-word	Biasing	The divorced couples have been *apart* all these years
	2-word	Biasing	This bolt may have been *a part* of the engine
	1-word	Neutral	He came to *appoint* a successor for himself
appoint	2-word	Neutral	He came to *a point* of frustration with life
a point	1-word	Biasing	The president wanted to *appoint* a good friend
	2-word	Biasing	The sharp knife came to *a point* at its tip
	1-word	Neutral	They were led to *arise* and fight for justice
arise	2-word	Neutral	They were led to *a rise* in altitude
a rise	1-word	Biasing	The sleeping giant could choose to *arise* at any time
	2-word	Biasing	Falling prices could lead to *a rise* in sales
	1-word	Neutral	The thought that *arose* could not have been more brilliant
arose	2-word	Neutral	The thought that *a rose* could bloom here is strange
a rose	1-word	Biasing	The film had zombies that *arose* quickly from the dead
	2-word	Biasing	No flower says the things that *a rose* can say
	1-word	Neutral	Tom was ready for *around* an hour of exercise
around	2-word	Neutral	Tom was ready for *a round* of golf
a round	1-word	Biasing	That house could sell for *around* a hundred thousand
	2-word	Biasing	Tiger Woods was ready for *a round* of golf
	1-word	Neutral	The people came to *arrest* the criminal
arrest	2-word	Neutral	The people came to *a rest* there at the park
a rest	1-word	Biasing	The police wanted to *arrest* the criminal
	2-word	Biasing	The tired athletes were entitled to *a rest* that day
	1-word	Neutral	Meghan took *aside* the children under 10
aside	2-word	Neutral	Meghan took *a side* that the others disagreed with
a side	1-word	Biasing	The trainer took *aside* the boxer and yelled at him
	2-word	Biasing	In debates the politician took *a side* that was popular
	1-word	Neutral	They claim that *attacks* largely happen at night
attacks	2-word	Neutral	They claim that *a tax* levy would help
a tax	1-word	Biasing	It's a vicious animal that *attacks* like lightning
	2-word	Biasing	The IRS told us that *a tax* law had changed
	1-word	Neutral	Steve's dog was *away* 10 days at the kennel
away	2-word	Neutral	Steve's dog was *a way* to avoid loneliness
a way	1-word	Biasing	Last time I traveled I was *away* twice this long
	2-word	Biasing	Being outgoing was *a way* to make friends

Appendix B. Acoustic measures

Group	Measurement
Duration (ms) (5 measures)	Entire sequence duration
	Schwa duration
	Second syllable duration
	Duration of the initial segment of the second syllable (Note: VOT for stops)
	Closure duration of the initial stops of the second syllable

(*continued on next page*)

Appendix B. (*continued*)

Group	Measurement
Duration (%) (6 measures)	Schwa duration / entire sequence duration
	Second syllable duration / entire sequence duration
	Segment duration / entire sequence duration
	Closure duration / entire sequence duration
	schwa duration / 2nd syllable duration
	schwa duration / 2nd-syllable initial segment duration
Amplitude (8 measures)	Mean amplitude of the entire sequence
	Mean amplitude of schwa
	Mean amplitude of the 2nd syllable
	Amplitude of 2nd-syllable initial segment
	Amplitude maximum of schwa
	Amplitude maximum of the 2nd syllable
	Amplitude minimum of schwa
	Amplitude minimum of the 2nd syllable
Pitch contour (10 measures)	Mean f0 of the entire sequence
	Mean f0 of schwa
	Mean f0 of the 2nd syllable
	F0 maximum of the entire sequence
	F0 maximum of the 2nd syllable (i.e., the stressed syllable)
	The ratio between f0 maxima and mean f0 of the entire sequence
	Temporal location of the f0 maximum relative to the onset of the sequence in milliseconds
	Temporal location of the f0 maximum relative to the onset of the sequence in percentage
	Temporal location of the f0 maximum relative to the onset of the second syllable in milliseconds
	Temporal location of the f0 maximum relative to the onset of the second syllable in percentage
Formant (5 measures)	1st formant of schwa at the vowel midpoint
	2nd formant of schwa at the vowel midpoint
	1st formant of 2nd-syllable initial segment at the midpoint
	2nd formant of 2nd-syllable initial segment at the midpoint
	3rd formant of 2nd-syllable initial segment at the midpoint

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jml.2011.12.007.

References

Allbritton, D., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 714–735.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

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*, 390–412.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*(1), 1–22.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31. http://CRAN.R-project.org/package=lme4.

Beckman, M., & Edwards, J. (1990). Lengthening and shortening and the nature of prosodic constituency. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 152–178). New York: Cambridge University Press.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America, 113*, 1001–1024.

Boersma, P., & Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0.43) [Computer program]. Retrieved December 9, 2008, from http://www.praat.org/.

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V. Acquisition and the Lexicon* (pp. 70–87). New York: Cambridge University Press.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics, 26*, 173–199.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics, 35*(2), 210–243.

Christie, W. M. Jr., (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America, 55*, 819–821.

Christie, W. (1977). Some multiple cues for juncture in English. *General Linguistics, 17*, 212–222.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1980). Segmenting speech into words. *Journal of the Acoustical Society of America, 64*, 1323–1332.

Cooper, A. M. (1991). Laryngeal and oral gestures in English /p, t, k/. Proceedings of the XIIth international congress of phonetic sciences, 2, pp. 50-53. Aix-en-Provence.

Cutler, A. (1987). Speaking for listening. In A. Allport, D. G. MacKay, W. Prinz, & E. Sheerer (Eds.), *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing* (pp. 23–40). London: Academic Press Ltd.

Cutler, A., & Butterfield, S. (1990a). Durational cues to word boundaries in clear speech. *Speech Communication, 9*, 485–495.

Cutler, A., & Butterfield, S. (1990b). Syllabic lengthening as a word boundary cue. *Proceedings of the 3rd Australian International Conference on Speech Science and Technology*, 324–328.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*, 218–236.

Cutler, A., & Carter, D. M. (1987). predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*, 133–142.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141–201.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 113–121.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 218–244.

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener.". In L. R. Gleitman, D. J. Napoli, & J. Kegl (Eds.), *Bridges between psychology and linguistics* (pp. 105–129). San Diego: Academic Press.

Dilley, L., Ladd, D. R., & Schepman, A. (2005). Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics, 33*(1), 115–119.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59*, 294–311.

Dilley, L., & Pitt, M. (2010). Altering context speech rate can cause words to appear and disappear. *Psychological Science, 21*, 1664–1670.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of vowel-initial syllables as a function of prosodic structure. *Journal of Phonetics, 24*, 423–444.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 50–86). Cambridge, MA: MIT Press.

Godfrey, J. J., & Holliman, E. (1997). Switchboard-1 Release 2 Linguistic Data Consortium, Philadelphia

Gow, D. W., Jr., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 344–359.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 483–503.

Harris, M. O., & Umeda, N. (1974). Effect of speaking mode on temporal factors in speech. *Journal of the Acoustical Society of America, 56*, 1016–1018.

Hoard, J. E. (1966). Juncture and syllable structure in English. *Phonetica, 15*, 96–109.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Experimental Phonetics, 3*, 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*, 1208–1221.

Krakow, R.A. (1989). The articulatory organization of syllables: a kinematic analysis of labial and velic gestures. Ph.D. dissertation, Yale University.

Ladd, D. R., & Schepman, A. (2003). Sagging transitions between high pitch accents in English: experimental evidence. *Journal of Phonetics, 31*(1), 81–112.

Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. *Phonetica, 5*(Suppl.), 1–54.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America, 51*(6), 2018–2024.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America, 54*(5), 1228–1234.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*, 172–187.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). The Netherlands: Kluwer.

Mattys, S. L., & Melhorn, J. F. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. *Journal of the Acoustical Society of America, 122*, 554–567.

Mattys, S. L., Melhorn, J. F., & White, L. (2007). Effects of syntactic expectations on speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 960–977.

Mattys, S. L., & Samuel, A. G. (1997). How lexical stress affects speech segmentation and interactivity: Evidence from the migration paradigm. *Journal of Memory and Language, 36*, 87–116.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*, 477–500.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*, 21–46.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America, 62*, 714–719.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357–395.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*, 191–243.

Oller, D. K. (1973). The effect of position in utterance on speech-segment duration in English. *Journal of the Acoustical Society of America, 54*, 1235–1247.

Pitt, M. A. et al. (2007). *Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology*. Ohio State University (Distributor).

Pollack, I., & Pickett, J. M. (1964). Intelligibility of excerpts from fluent speech: Auditory vs. structural context. *Journal of Verbal Learning & Verbal Behavior, 3*, 79–84.

R Development Core Team, (2009). R: A language and environment for statistical computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.

Redford, M. A. (2007). Word-internal versus word-peripheral consonantal duration patterns in three languages. *The Journal of the Acoustical Society of America, 121*(3), 1665–1678.

Redford, M. A., Davis, B. L., & Miikkulainen, R. (2004). Phonetic variability and prosodic structure in motherese. *Infant Behavior and Development, 27*, 477–498.

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in realization of glottalization in normal speakers. *Journal of Phonetics, 29*, 407–429.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51–89.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition, 105*, 466–476.

Sandhaus, E. (2008). *The New York Times annotated corpus*. Philadelphia: Linguistic Data Consortium.

Scarborough, R., Brenier, J., Zhao, Y., Hall-Lew, L., & Dmitrieva, O. (2007). An acoustic study of real and imaginary foreigner-directed speech. In *Proceedings of 17th international congress of phonetic sciences, Germany*.

Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science, 17*(5), 372–377.

Smith, R. H. (2004). The role of fine phonetic detail in word segmentation. Unpublished Ph.D. thesis, University of Cambridge.

Son, R. J. J. H., & Van Santen, J. P. H. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication, 47*, 100–123.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language, 48*(2), 233–254.

Stephens, J. D. W., & Pitt, M. A. (2007). Word segmentation in the "real world" of conversational speech. Poster presented at the 48th meeting of the Psychonomic Society, Long Beach, CA.

Turk, A., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*(4), 397–440.

Umeda, N., & Coker, C. H. (1975). Subphonemic details in American English. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception* (pp. 539–564). London: Academic Press.

Welby, P. (2007). The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication, 49*, 28–48.

White, L., Wiget, L., Rauch, O. & Mattys, S. L. (2010). Segmentation cues in spontaneous and read speech. In *Proceedings of the 5th speech prosody conference, USA*.