LANGUAGE AND COGNITIVE PROCESSES, 2012, iFirst, 1–28

Psychology Press
Taylor & Francis Group

When cues combine: How distal and proximal acoustic cues are integrated in word segmentation

Christopher C. Heffner[1,2], Laura C. Dilley[1,2,3], J. Devin McAuley[2], and Mark A. Pitt[4]

[1]Department of Linguistics and Germanic, Slavic, Asian and African Languages, Michigan State University, East Lansing, MI, USA
[2]Department of Psychology, Michigan State University, East Lansing, MI, USA
[3]Department of Communicative Sciences and Disorders, Michigan State University, East Lansing, MI, USA
[4]Department of Psychology, The Ohio State University, Columbus, OH, USA

Spoken language contains few reliable acoustic cues to word boundaries, yet listeners readily perceive words as separated in continuous speech. Dilley and Pitt (2010) showed that the rate of nonlocal (i.e., distal) context speech influences word segmentation, but present theories of word segmentation cannot account for whether and how this cue interacts with other acoustic cues proximal to (i.e., in the vicinity of) the word boundary. Four experiments examined the interaction of distal speech rate with four proximal acoustic cues that have been shown to influence segmentation: intensity (Experiment 1), fundamental frequency (Experiment 2), word duration (Experiment 3), and high frequency noise resembling a consonantal onset (Experiment 4). Participants listened to sentence fragments and indicated which of two lexical interpretations they heard, where one interpretation contained more words than the other. Across all four experiments, both distal speech rate and proximal acoustic manipulations affected the reported lexical interpretation, but the two types of cues did not consistently interact. Overall, the results of the set of experiments are inconsistent with a strictly-ranked hierarchy of cues to word boundaries, and instead highlight the necessity of word segmentation and lexical access theories to allow for flexible rankings of cues to word boundary placement.

Keywords: Prosody; Word segmentation; Speech rate; Lexical perception.

Correspondence should be addressed to Laura C. Dilley, Department of Communicative Sciences and Disorders, Room 116 Oyer Hall, East Lansing, MI 48824-1220, USA. E-mail: ldilley@msu.edu

This work was supported by Grant BCS-0847653 to Laura C. Dilley from the National Science Foundation, a Professorial Assistantship award to Christopher C. Heffner from the Michigan State University Honors College, and Andrew Fellowship and Dean's Assistantship awards to Christopher C. Heffner from the Michigan State University College of Social Science. Portions of this research were presented at the 51st Annual Meeting of the Psychonomic Society. We are grateful to Tuuli Morrill, Sven Mattys, and two anonymous reviewers for input on draft versions of this paper. We thank Claire Carpenter, Evamarie Cropsey, Bryan Reynolds, Liz Wieland, and all the members of the MSU Speech Lab for help throughout the project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

http://www.psypress.com/lcp http://dx.doi.org/10.1080/01690965.2012.672229

Understanding spoken language requires segmentation of the continuous acoustical signal into discrete words; however, few acoustic cues have been identified that consistently signal a word boundary (Cole, Jakimik, & Cooper, 1980; Lehiste, 1960). Thus, much remains to be explained regarding what information listeners use to recognise where a word boundary occurs, as well as how many word boundaries—and hence how many words—are present in the speech signal. The present paper focuses on the segmentation of words that begin with a vowel when there is another immediately preceding vowel or sonorant segment. In such contexts, there is often coarticulation across the word boundary such that it is unclear even how many boundaries there are. For example, should [æftəɹɪɪʃ], with a long "er" ([ɚ]) sound, be glossed as *after rich* (which contains one word boundary), *after a rich* (which contains two word boundaries), or some other possibility? And to what extent can various acoustic cues contribute to the perception of a word boundary?

Current theories of word segmentation can be broadly divided into two categories: lexical theories and pre-lexical theories. According to lexical theories of segmentation (e.g., TRACE, McClelland & Elman, 1986; Shortlist, Norris, 1994), word segmentation occurs as a consequence of lexical activation and lexical competition; lexical items are identified from the input sequence of phonemes, and the locations at which one lexical item ends and the next begins are identified as word boundaries. It is unclear, however, how lexical theories would be able to accommodate ambiguity in the number of word boundaries, rather than just in the locations of word boundaries. In lexical theories, the notion of phonemes as discrete inputs to the parser is taken for granted, and, as such, ambiguity concerning the number of phonemes (and, therefore, the number of word boundaries) is rather challenging to accommodate. More recent lexical accounts of word segmentation have taken steps to address the oversimplified account of input to the parser by adding some level of probability to the possible parser inputs (e.g., Shortlist B, Norris & McQueen, 2008). However, it is still unclear how these models would be capable of resolving ambiguity in the number of word boundaries (and words) in the speech signal rather than merely resolving the locations of a fixed number of word boundaries, in cases when multiple potential interpretations lead to perception of valid lexical items.

Pre-lexical accounts of segmentation (e.g., Christiansen, Allen, & Seidenberg, 1998), meanwhile, posit that listeners make use of a variety of sublexical cues in order to segment speech, building on studies of language production. Sublexical cues, which lead to real-time resolution of lexical ambiguity (Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003), can include segmental cues, such as probabilistic phonotactics (McQueen, 1998; Saffran, Newport, & Aslin, 1996), subsegmental cues such as laryngealisation (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996) or allophonic cues (Mattys & Jusczyk, 2001), and suprasegmental cues, such as word stress (Cutler & Norris, 1988) and segmental lengthening (Byrd & Saltzman, 2003; Shatzman & McQueen, 2006; Turk & Shattuck-Hufnagel, 2000). Though pre-lexical theories can accommodate ambiguity in the number of word boundaries more easily, they do not specify which acoustic cues in the speech signal are used to determine the number of word boundaries in any particular sequence, nor to what extent each acoustic cue is utilised. Ascertaining the relative contribution of each acoustic cue to word boundary placement may be difficult, due to uncertainty in the communicative status of each of these cues. Pitch (and its acoustic correlate, fundamental frequency, or F_0) has been considered to be straightforwardly suprasegmental in nature (Lehiste, 1970), yet it is systematically perturbed by segmental boundaries (Hanson, 2009; Pardo & Fowler, 1997), such as those produced

by the potentially subsegmental realisation of glottalised voice quality at the onset of a vowel-initial word (Hillenbrand & Houde, 1996). A similar argument can be constructed for other acoustic cues such as intensity and lengthening.

Despite the predictions of pre-lexical theories, few previous studies have directly manipulated multiple sublexical acoustic cues simultaneously to examine their combined influence on the perception of word boundaries. Repp, Liberman, Eccardt, and Pesetsky (1978) varied the duration of a silent interval at the word boundary within the phrase "grey ship" and the duration of frication noise in "sh" ([ʃ]), thereby generating four possible percepts: "gray ship", "great ship", "gray chip", and "great chip". Across multiple levels of silence duration, lengthening the duration of frication noise triggered more frequent perception of "ship" than "chip". Above a certain silence duration threshold (approximately 20 ms) participants were more likely to hear "great" than "gray", though this varied as a function of frication noise duration. Hillenbrand and Houde (1996) investigated acoustic cues to the perception of intervocalic glottal stops by manipulating amplitude and F_0 contours in synthetic utterances with continuous voicing modelled after the naturally-produced digit sequence "oh-oh" [oʔo]. They found that the presence of a dip in intensity and/or F_0 was almost always sufficient to cause perception of a glottal stop in the middle of the stimulus (i.e., to cause perception of two syllables, rather than one). Shimizu and Dantsuji (1980) found that increasing F_0 in the vowel immediately preceding a candidate word boundary increased the likelihood of perceiving that boundary, but their results depended on the dialect of Japanese spoken by participants, and no statistical analyses were reported. Given the low number of perceptual studies related to potentially suprasegmental cues, and the statistical and ecological limitations in some of the studies that do exist, examining the results of the systematic manipulation of those cues with regard to perception is of paramount importance.

Though the lexical and pre-lexical approaches have some striking differences, recent work has attempted to synthesise the information from the two theoretical perspectives into a unified approach to word segmentation (Mattys, White, & Melhorn, 2005). Mattys et al. (2005) accepted some of the chief arguments of the lexical and pre-lexical approaches, and then set about ranking each of the cues with respect to each other, eventually postulating a three-tier system. Tiers are organised in such a way that cues in higher tiers are assumed to override cues occupying tiers that they dominate. In the top tier (Tier I) are the "Lexical" cues, such as sentential context and lexical knowledge. Next, in Tier II, are the "Segmental" cues, like allophony and phonotactics. And, finally, "Metrical Prosody" constitutes Tier III, featuring cues such as word stress. The system accounts for a number of experimental findings presented in the paper itself as well as a variety of subsequent work (see, e.g., Mattys & Melhorn, 2007).

The Mattys et al. (2005) hierarchy is the most fully elaborated theory proposed yet that attempts to take into account multiple segmentation cues. However, other theoretical approaches have begun to be developed which attempt to account for how multiple sources of information can simultaneously influence phonetic perception. These approaches are based, for instance, on Bayesian statistics (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Toscano & McMurray, 2010), and have begun to be extended to word segmentation (Goldwater, Griffiths, & Johnson, 2009; Norris & McQueen, 2008). Further development of these various models may prove fruitful, as recent research in word segmentation has suggested already that the hierarchy proposed by Mattys et al. (2005) provides an incomplete account for certain experimental findings. For example, transitional probabilities, that

is, the statistical tendency for certain phoneme sequences to straddle word boundaries and others to occur word-internally, are ranked at the same level as phonotactics and coarticulation in Tier II of the proposed hierarchy. Despite their identical position on the hierarchy, phonotactics and coarticulation can both overpower transitional probabilities in cuing word boundary placement, though the dominance of each cue is subject to modulation by signal quality (Fernandes, Ventura, & Kolinsky, 2007; Mersad & Nazzi, 2011). In addition, recent work by Newman, Sawusch, and Wunnenberg (2011) has argued for inclusion of a Possible Word Constraint in the segmentation hierarchy when segmenting fluent speech into individual words (Norris, McQueen, Cutler, & Butterfield, 1997); such a constraint would limit the system to considering only parsings that could conceivably be words in the language, so as not to strand illegal sequences.

A separate line of discrepancies from the Mattys et al. (2005) word segmentation hierarchy has emerged in the form of "distal" prosodic cues that are temporally distant from a potential word boundary. Distal cues contrast with "proximal" prosodic cues that are on a syllable adjacent to a potential word boundary. In one experiment investigating the use of distal prosodic cues, listeners transcribed sequences of syllables such as *down town ship wreck*, which were ambiguous as to whether they ended in a monosyllabic word (as in *down township wreck*), or in a bisyllabic word (as in *downtown shipwreck*). Listeners perceptually grouped the syllables, thereby hearing either a monosyllabic word or a bisyllabic word at the end of the sequence of syllables, in line with patterns established in distal prosodic information (e.g., F_0) even without changing the acoustic information found within the proximal region (Dilley & McAuley, 2008). The location of distal prosodic cues in the Mattys et al. (2005) hierarchy is unclear, a point which was followed up in subsequent work by Dilley, Mattys, and Vinke (2010). Given the nonlexical nature of prosodic cues, distal prosodic cues ought to rank lower than the syntactic and semantic cues situated in Tier I of the hierarchy; however, when distal prosodic and semantic cues are pitted against each other, distal prosodic cues can override semantic cues in determining word boundary placement (Dilley et al., 2010).

Distal speech rate is another distal prosodic cue which has been shown to have robust effects on word segmentation. Dilley and Pitt (2010) investigated the influence of distal speech rate information on word boundary placement by looking at speech fragments showing spectral continuity across the word boundary at the onset of a critical function word (e.g., *or* realised as [ɚ] in the context *Don must see the harbor or boats*...). When the entire fragment was presented unaltered, listeners usually heard a function word. However, when the distal context was slowed down while the proximal context around the function word was presented at the unaltered spoken rate, participants reported hearing a function word much less frequently. In other words, the contrast between the (slowed) distal speech rate and the (relatively fast) proximal speech rate caused them to perceive one less word boundary when the distal speech rate was slowed down. The effects demonstrated by Dilley and Pitt (2010) represent a novel departure from a rich body of previous work on phonetic effects of speech rate on phonemic perception (Miller, 1981; Miller & Volaitis, 1989; Summerfield, 1981) in that distal speech rate manipulations were shown to make entire words appear or disappear perceptually. This illustrates that distal speech rate can affect the number of words (and hence, the number of word boundaries and phonemes) that are heard, not just the location of a boundary along a phonetic continuum for lexical material with a fixed number of phonemes. Recent work has similarly shown that in Dutch, distal

speech rate can influence whether a segment is heard as word-initial or word-final (Reinisch, Jesse, & McQueen, 2011).

However, distal prosodic effects, including the distal speech rate effect, have yet to be situated within the Mattys et al. (2005) hierarchy. More broadly, examining whether distal speech rate trades off against other kinds of segmentation information will help to provide evidence for the idea that word segmentation is a dynamic process that involves flexible and dynamic cue integration. Data which might demonstrate such a flexible and dynamic process could then serve as a starting place for the further development of word segmentation models.

For the present, we focus on the proposal of Mattys et al. (2005), the most fully elaborated word segmentation theory so far that is capable of accounting for the integration of multiple cues. In so doing it is noteworthy that Dilley et al. (2010) found that distal prosodic cues override semantic cues (Tier I) in determining word boundary placement. As such, distal prosodic cues would be situated above semantic cues in a word segmentation hierarchy, a tier above the present "Tier I". Conversely, it is also possible that distal prosodic cues, including the newly-reported distal speech rate cue, are easily outranked by other segmentation cues. Mattys et al. (2005) assigned Tier III, their lowest ranked tier, to be the tier for "metrical prosody". If that description of Tier III is widened to include all prosodic considerations, as seems reasonable, it would be predicted that distal prosodic cues should be easily outranked by Tier II (segmental) and Tier I (knowledge-based) cues.

In the present work, we extended the results of Dilley and Pitt (2010) to determine whether distal speech rate, in particular, is robust in the face of conflicting proximal cues, or whether it is easily overcome by other cues. These experiments further represent some of the first experiments to directly manipulate acoustic cues to test pre-lexical theories' predictions about the use of sublexical cues in word boundary placement. By manipulating both distal and proximal acoustic parameters and assessing their relative strength, this study allowed investigation of the question of whether distal prosodic cues might be integrated into the Mattys et al. (2005) hierarchy in a place that is distinct from that of "metrical" prosodic cues, as suggested by Dilley et al. (2010), while helping to clarify what that place should be. In addition, given the uncertain nature of proximal acoustic cues as segmental, subsegmental, or suprasegmental cues, this work will help clarify the place of various proximal acoustic cues in a segmentation hierarchy, or perhaps whether they can be fit into the hierarchy at all. Most importantly, these experiments were intended to provide a test of whether the weighting of word segmentation cues are evaluated dynamically such that cues can "trade off" in the word segmentation process as a function of their individual strengths.

Four experiments were conducted in which participants listened to sentence fragments and were asked to indicate whether they heard a word, here referred to as a "critical word", in a region of the sentence fragments with acoustic ambiguity to the existence of a word boundary. Each experiment involved manipulating distal speech rate in addition to a different proximal acoustic cue. A "critical word report rate" was computed for each combination of distal speech rate and proximal cue strength, representing how often participants reported hearing a critical word. In Experiment 1, intensity was manipulated; in Experiment 2, F_0; in Experiment 3, proximal word duration; and, in Experiment 4, high-frequency noise. For all experiments, slowing distal speech rate was expected to reduce critical word report rate. For example, a slowed distal speech rate would cause participants to report hearing a word between *after* and *rich* in the phrase *the value went up after her rich* less

often. Additionally, it was hypothesised that strengthening the proximal acoustic cues employed for each experiment would increase critical word report rates. For instance, in Experiment 2, it was predicted that participants would report hearing a word between *after* and *rich* more often with a large change in proximal F_0 than with a small change to this cue.

Several prior studies suggest that an interaction between distal and proximal cues is also likely. First, Dilley et al. (2010) found that orthogonally-manipulated proximal and distal acoustic cues interacted with each other in determining segmentation of possible compound words; the effects of distal prosody were attenuated by particularly strong proximal cues. For the present experiments, it is predicted that strong acoustic manipulations (larger intensity, F_0, word duration, and frication noise) should produce a fairly unambiguous percept of a word boundary, resulting in decreased effectiveness of distal speech rate as a cue compared with the effects of the distal speech rate cue in conditions with a smaller, more ambiguous proximal acoustic discontinuity.

A second reason for expecting an interaction between distal and proximal cues is the robust literature on "cue trading," which reflects interactions between different, often proximal cues to segmental identity in perception of segments (Miller, 1994; Repp, 1982). Some studies have shown that listeners' judgments of locations of boundaries within a continuum of voice onset times (VOTs) for voicing-based phonetic category distinctions are dependent on the speech rate of the containing and immediately preceding syllables (Miller & Volaitis, 1989; Summerfield, 1981). Though most prior studies examining speech rate have primarily investigated this cue at a proximal level, Wayland, Miller, and Volaitis (1994) showed that distal speech rate can demonstrate interactive effects with proximal VOT cues to determine the location of stimuli which are perceived as best exemplars of each voicing category. Prior work such as this showing that speech rate affects phonemic boundaries and locations of exemplars along a phonetic continuum for a fixed number of phonemes is clearly distinct from the findings reported in Dilley and Pitt (2010), where speech rate created percepts of variable numbers of phonemes and word boundaries. Nevertheless, the mechanisms by which speech rate acts may be similar or overlapping in both sets of phenomena, underscoring the likely possibility of interactions between distal speech rate and proximal cues similar to these "cue trading" reports in segmental perception.

EXPERIMENT 1

The first potential proximal acoustic cue to word boundaries we manipulated was intensity. Intensity only rarely has been directly manipulated in studies of word segmentation, even though consonants are often associated with local reductions in intensity (Stevens, 1998). Perception of glottal stops, which systematically fill word-initial onset positions in syllables (Kenstowicz, 1994), can be cued by local intensity decreases within steady-state vocalic regions (Hillenbrand & Houde, 1996). If a local decrease in intensity causes a consonant such as the glottal stop to be perceived within an acoustically ambiguous region that might contain a function word, critical word report rates should increase. Manipulating proximal intensity and distal speech rate simultaneously also allows for a test of whether distal prosodic effects on word boundary placement are sensitive to proximal cues, and whether the cues are involved in any trading relations in determining word boundary placement. If the distal speech rate cue is outranked by the potentially segmental cue of intensity change, it would

provide some evidence that distal speech rate belongs lower than Tier II (segmental cues) in the Mattys et al. (2005) hierarchy.

Method

Participants

Twenty-nine participants (24 female, 5 male) were recruited for research credit at Michigan State University. All were native speakers of English who self-reported normal hearing and were at least 18 years of age ($M = 19.7$ years, Range = 18–28 years).

Stimuli and design

This experiment implemented a 2 (Sentence Fragment) × 4 (Intensity Dip) × 13 (Distal Speech Rate) within-subject design. The stimuli were derived from materials used in Dilley and Pitt (2010), Experiment 1. Speech fragments recorded for that experiment which served as the basis of sentence fragments constructed here were *The value went up after her rich neighbors*... and *People were offended after her rude*... In each case, the word *her* was spoken as "*er*", [ɚ], with no initial /h/ sound, so that spectral blending occurred across the initial word boundary of *her*; see Figure 1. These two sentence fragments formed the basis of what will be referred to here as the "rich" and "rude" sentence fragments, respectively.

Each sentence fragment was divided into a "target" region (the acoustically ambiguous function word, here "*er*" [ɚ]; the preceding syllable, here *-ter* [tɚ]; and the following phoneme, here *r-* [ɹ]) and a "context" region (everything else in the

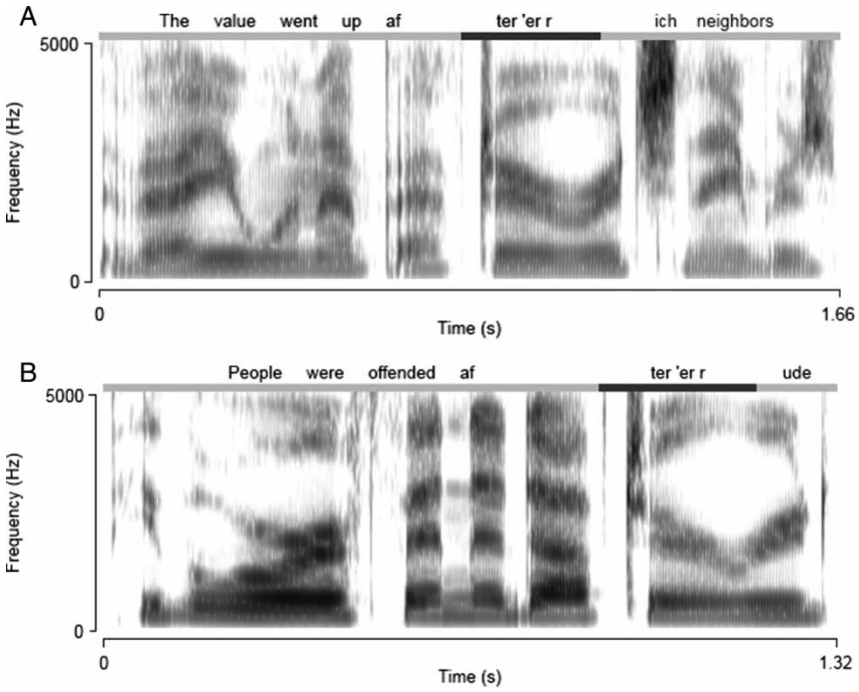

Figure 1. Spectrogram representations of the original recordings of speech materials used for the "rich" (a) and "rude" (b) sentence fragments. The lines above the spectrogram give the approximate durations of the context (light gray) and target (dark gray) regions, with the text above that denoting the orthographic transcriptions of each region.

sentence fragment), following Dilley and Pitt (2010). These fragments were chosen because the stimuli derived from them exhibited a robust distal speech rate effect in Dilley and Pitt's (2010) Experiment 1, permitted a wider range of acoustic manipulation within the target region due to a relatively stable steady-state vocalic region, and allowed investigation of the effect of cues signalling a consonantal onset on word boundary placement. Though these two sentence fragments do have similar target regions ([æft] and [ɹ] bounding a region of [ɚ]), the distal speech rate effect is not unique to the adjacent context they share (Dilley & Pitt, 2010).

Each of the speech fragments was then subjected to parametric manipulations of both distal speech rate and intensity using Praat software (Boersma & Weenink, 2009), as follows. First, 13 different distal speech rates were created by multiplying the duration of the context region by values ranging from 1.25 to 1.85 in steps of 0.05, hereafter "distal duration multipliers". The chosen range of distal duration multipliers was selected to allow for a fine-grained assessment of effects of distal speech rate on critical word reports and potential interactions with the manipulated proximal cue. The specific values for the duration multipliers were selected from pilot work showing that rates of hearing word boundaries varied substantially across this range. The speech fragments were then subject to an intensity manipulation, for which a stylised intensity profile very similar to that of Hillenbrand and Houde (1996) was created in Praat, consisting of a linear, 26.8 ms intensity dip, followed by a constant, 13.8 ms intensity "trough", then a linear, 26.8 ms rise. The durations for the linear intensity fall, trough, and rise were based on values observed in acoustic analysis of glottal stop tokens produced by four speakers from the materials of Dilley and Pitt (2010). For the analysis of glottal stop tokens, the endpoints of the intensity fall and rise were operationally defined as the nearest glottal pulses to the intensity minimum of the glottal segment. The trough was defined as the area immediately adjacent to the intensity minimum with an intensity of less than 0.40 dB greater than that minimum, approximately the just-noticeable difference (JND) for a 70 dB sound (Viemiester & Bacon, 1988). The magnitude of the intensity dip was manipulated in four steps: 0 dB, 6 dB, 12 dB, and 18 dB down from the original stimulus intensity, selected based on the results of Hillenbrand and Houde (1996). These intensity manipulations were performed using Praat to multiply the amplitude of the trough region by scale factors, with troughs centred at the midpoint of the vocalic region of the target (Figure 2).

Apparatus

E-Prime 2.0 Professional software of Psychology Software Tools, Inc. (Sharpsburg, PA) was used to control all aspects of stimulus presentation and response collection. Responses were collected on Lenovo 6209 computers, with participants listening to presentations over Senneheiser HD 280 pro headphones.

Procedure

Participants were told to press a button on a keyboard labelled "yes" or "no" to indicate whether they heard a word between "after" and "rich" (for the "rich" fragment) or between "after" and "rude" (for the "rude" fragment). In particular, for the "rich" fragment, they were told to press "yes" if they heard something like "the value went up after our rich neighbors" (with "a" and "her" also given as examples of a word that might be heard in place of "our"), and "no" if they heard something like "the value went up after rich neighbors". For the "rude" fragment, they were told to

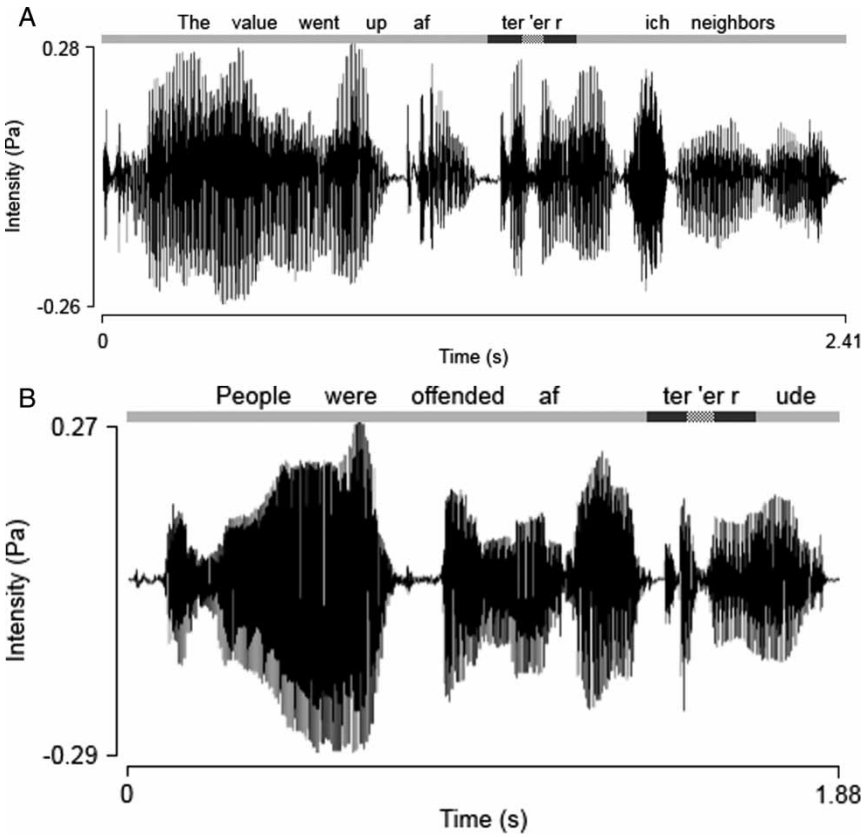

Figure 2. Waveform representations of examples from the "rich" (a) and the "rude" (b) sentence fragments with a distal word rate factor of 1.55 and intensity dip of 18 dB. The lines above the spectrogram give the approximate durations of the context (light gray) and target (dark gray) regions, with the text above that denoting the orthographic transcriptions of each region. The dotted area within the target region of the line reflects the locus of intensity manipulation within the target.

press "yes" if they heard something like "people were offended after a rude" (with "her" also given as an example of a word that might be heard in place of "a"), and "no" if they heard something like "people were offended after rude". Participants completed a short practice session consisting of six trials, representing a variety of stimuli from both ends of the distal speech rate and intensity continua, followed by the experimental trials. Trials across the experiment were blocked by sentence fragment ("rude" or "rich"); there was counterbalancing of block order across participants, such that 15 of participants were randomly assigned to an order in which they heard all "rude" stimuli first, and 14 were assigned to an order in which they heard the "rich" fragment first.

Within each block, a total of 52 trials (4 levels of Intensity Dip $\times$ 13 levels of Distal Speech Rate) cycled through all stimuli for a given sentence fragment in a randomly-generated order; this cycling through of stimuli was repeated six times, such that over the course of the block each stimulus was presented a total of six times. The administration of the instructions for the second sentence fragment occurred midway through the experiment, which also served as a short break for participants. The experiment was self-paced and most participants took between 45 and 55 min to finish it.

Results

Figure 3 shows the effects of Intensity Dip and Distal Speech Rate on the percentage of trials for which participants indicated that a critical word was present (i.e., percentage of "yes" responses) for the "rich" and the "rude" Sentence Fragments, respectively; this will subsequently be referred to as the percentage of critical word reports. Separate lines represent the different intensity levels. Data were analyzed using logit mixed-effect models (Jaeger, 2008) in the lme4 package (Bates, Maechler, Bolker, & Vasishth, 2011) written for the R statistical package (version 2.11.1; the R foundation for statistical computing), with subjects as a random factor and Sentence Fragment, Intensity Dip,

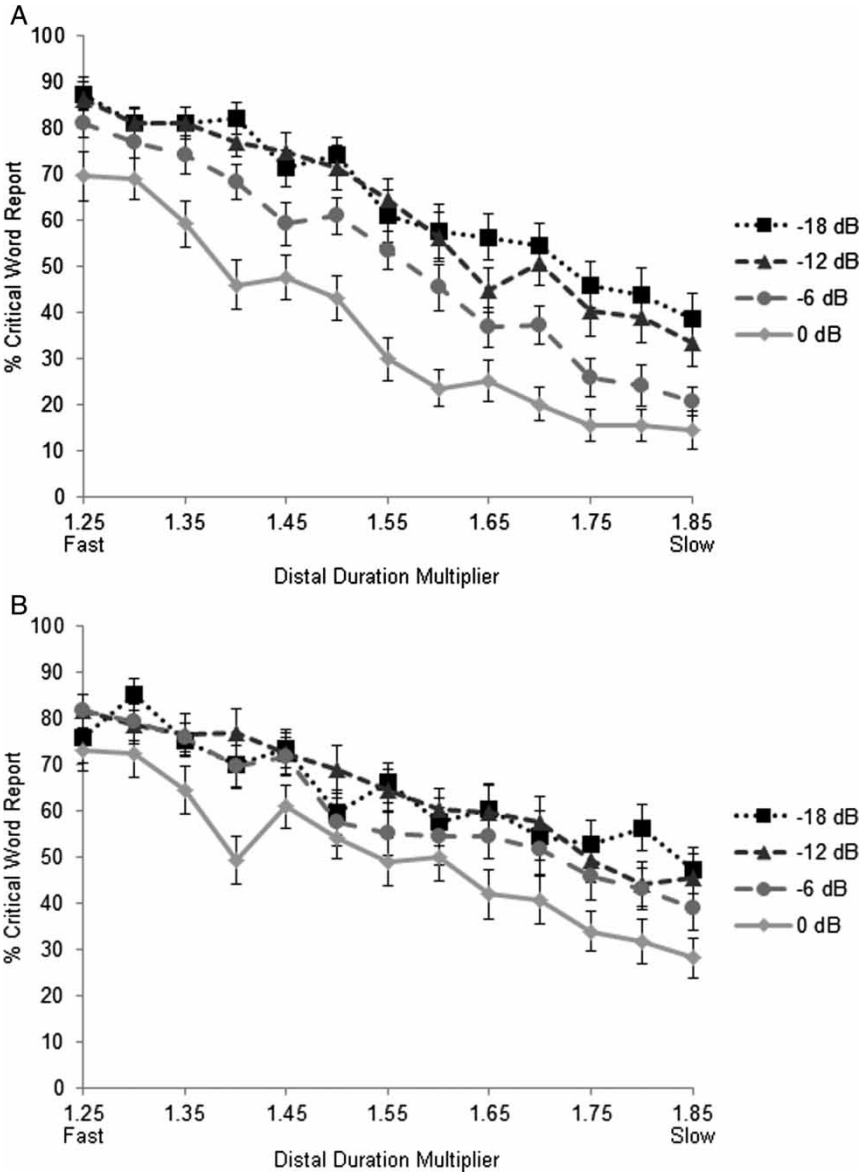

Figure 3. Mean percentage of participant critical word reports as a function of Distal Speech Rate and Intensity Dip, in dB, for the "rich" (a) and "rude" (b) sentence fragments in Experiment 1.

and Distal Speech Rate as fixed factors. Likelihood ratio test results showed that the best-fitting model was the saturated model, with all three fixed factors and their interactions as terms. Slower Distal Speech Rates (corresponding to increasing values of the distal duration multiplier) had lower critical word reports ($b = –0.155$, $p < .01$), whereas increasing the magnitude of the Intensity Dip led to more critical word reports ($b = 0.22$, $p < .01$). In general, there were more critical word reports for the "rude" Sentence Fragment ($M = 60\%$, $SD = 16\%$) than for the "rich" Sentence Fragment ($M = 53\%$, $SD = 12\%$), $b = –0.28$, $p < .01$. There was also a marginally significant interaction between Distal Speech Rate and Intensity Dip ($b = 0.01$, $p < .06$). As the Intensity Dip level increased, there was a tendency of Distal Speech Rate to have less of an effect on the percentage of critical word reports. Sentence Fragment additionally interacted with both Distal Speech Rate ($b = –0.08$, $p < .01$) as well as Intensity Dip ($b = 0.25$, $p < .01$). The interactions involving Sentence Fragment suggest there was a somewhat larger effect of Distal Speech Rate for the "rich" fragment than for the "rude" fragment and that for the "rich" fragment there existed a greater separation between different levels of Intensity Dip than for the "rude" fragment.

Discussion

Results of Experiment 1 replicate and extend the basic findings of Dilley and Pitt (2010). First, consistent with Dilley and Pitt (2010), slowing distal speech rate around an acoustically continuous region of speech decreases the rate of reporting that the speech contains an extra (critical) word. Second, the manipulating the magnitude of a proximal intensity cue (i.e., the magnitude of an intensity dip within a steady-state vocalic region) leads to an increased likelihood of perceiving a word boundary, at least under the present experimental conditions. The latter finding is consistent with previous work on glottal stop perception (Hillenbrand & Houde, 1996), where the authors found that glottal stops were more frequently perceived within a steady-state vocalic region as a function of an intensity dip within that region.

Finally, Experiment 1 reveals mixed support for an interaction between distal and proximal cues. As the proximal acoustic cue of intensity became stronger, there was a tendency for distal speech rate to have less of an effect on critical word reports. However, the interaction between distal and proximal cues was not entirely consistent across sentence fragments; the two cues traded off in a unique way for each sentence fragment rather than being consistently ranked in strength relative to one another.

EXPERIMENT 2

For Experiment 2, proximal F_0 was manipulated to examine its role in word boundary perception. F_0 has been described as a straightforward example of a suprasegmental cue which conveys sentence-level prosodic information (Lehiste, 1970). More recently, however, changes in F_0 have been shown to co-occur with obstruent consonants (Hanson, 2009; Pardo & Fowler, 1997), and F_0 has been implicated in the perception of glottal stops (Hillenbrand & Houde, 1996). Observations of the many functions of F_0 in cueing both segmental and suprasegmental structure led us to hypothesise that a dip in F_0 within the lexically ambiguous target region would potentially induce the perception of the start of a new prosodic unit that began with a glottal stop, leading to an increase in critical word reports. Furthermore, effects of distal speech rate were expected in this experiment as well, with slower distal speech rates reducing critical word reports. Finally, the distal speech rate and proximal F_0 variables were expected to

interact, in line with research on trading relations in segmental identification cues (Miller, 1994; Repp, 1982) and with Experiment 1, with larger dips in F_0 leading to decreased effectiveness of distal speech rate as a cue to word boundaries.

Method

Participants

Twenty participants (14 female, 6 male) were recruited for research credit at Michigan State University. All were native speakers of English who self-reported normal hearing and were at least 18 years of age ($M = 21.5$ years, Range = 18–47 years).

Stimuli and design

The same spoken phrases as used in Experiment 1 formed the basis of stimuli in Experiment 2. Moreover, the distal speech rates used to manipulate context speech rate and trough length for the proximal manipulation were identical to Experiment 1, with a proximal F_0 change replacing the proximal intensity change of Experiment 1, leading to a 2 (Sentence Fragment) $\times$ 4 (F_0 Dip) $\times$ 13 (Distal Speech Rate) within-subject design. The manipulation of distal speech rate was performed before instantiation of the F_0 dip. Data from Hillenbrand and Houde (1996) suggested that F_0 changes greater than about 20 Hz were not differentiated by subjects in determining the existence of glottal stops. The values chosen for this experiment consisted of dips of 0 Hz, 7 Hz, 14 Hz, and 21 Hz in order to elicit glottal stop perceptions variably. The manipulations were performed using Praat, with 26.8 ms F_0 troughs centred at the midpoint of the vocalic region of the target, and 13.8 ms linear interpolations of F_0 between the endpoints of the trough and the surrounding region of the target (Figure 4).

Apparatus

The apparatus used for Experiment 2 were identical to that of Experiment 1, with the sole difference that responses were indicated by the participant pressing buttons labelled "Yes" or "No" on a Psychology Software Tools, Inc. 200A Serial Response Box.

Procedure

The procedure was identical to Experiment 1.

Results

Figure 5 shows the effects of F_0 Dip and Distal Speech Rate on the percentage of critical word reports for the "rude" and the "rich" Sentence Fragments; separate lines show the different values of F_0 Dip. Logit mixed-effect models were again used, with Sentence Fragment, F_0 Dip and Distal Speech Rate as fixed factors and subjects as a random factor. Consistent with Experiment 1, the saturated model, with all three fixed factors and their interactions as terms, was the best-fitting model. Slowing down Distal Speech Rate (by increasing the distal duration multiplier) reduced critical word reports ($b = -0.23$, $p < .01$), and, similar to the Intensity Dip manipulation, increasing values of F_0 Dip increased critical word reports ($b = 0.16$, $p < .01$). There were more critical word reports for the "rude" sentence fragment ($M = 54\%$, $SD = 12\%$) than the "rich" sentence fragment ($M = 51\%$, $SD = 9.6\%$), $b = -0.14$, $p < .01$. The interaction between F_0 Dip and Distal Speech Rate was not reliable, $b = 0.007$, $p = .37$. Again mirroring the results of Experiment 1, Sentence

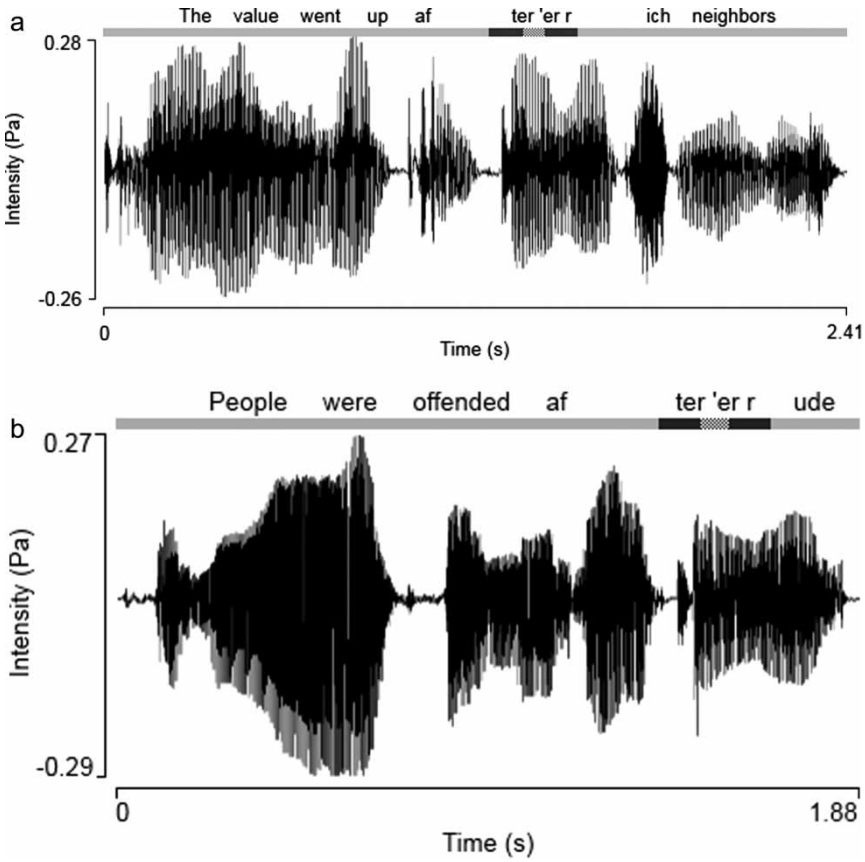

Figure 4. Waveform representations of examples from the "rich" (a) and "rude" (b) sentence fragments with a distal word rate factor of 1.55 and F_0 dip of 21 Hz. The lines above the spectrogram give the approximate durations of the context (light gray) and target (dark gray) regions, with the text above that denoting the orthographic transcriptions of each region. The F_0 dip is visible as a widening of the spacing between pitch periods across the region indicated by the dotted area shown within the target portion of the line.

Fragment additionally interacted with both Distal Speech Rate ($b = –0.08$, $p < .01$) as well as F_0 Dip ($b = 0.38$, $p < .01$). The main effects of Distal Speech Rate and F_0 Dip were replicated in separate analyses by Sentence Fragment, and the interaction between the two was marginally reliable for "rich" ($b = 0.015$, $p < .06$). Note that the effects of F_0 dip persisted across many distal speech rates, particularly for the "rich" fragment.

Discussion

Results of Experiment 2 again confirm the effects of distal speech rate on word boundary placement, with slower Distal Speech Rates leading to fewer critical word reports. Moreover, the proximal acoustic cue of F_0 Dip was effective in modulating critical word reports. The latter result shows that F_0, which is systematically perturbed by consonant segments (Hanson, 2009; Hillenbrand & Houde, 1996; Pardo & Fowler, 1997), is indeed a cue to word boundary presence and placement. These results extend the findings of Hillenbrand and Houde (1996) concerning the influence of F_0 Dip on glottal stop perception to a different set of contexts, suggesting that the F_0-motivated

placement of a glottal stop within the target region led to the perception of an "extra" word within the target region. If, indeed, it is a glottal stop that is perceived within the target region, this could be a form of allophonic variation, given that vowel-initial words are often produced with a glottalised onset (Dilley et al., 1996), and therefore situated within Tier II of the Mattys et al. (2005) hierarchy. Distal speech rate, a prosodic (likely Tier III) cue, here seems to be stronger than the F_0 information, mirroring findings from Experiment 1 and further indicating the difficulty of incorporating this cue into the segmentation hierarchy proposed by Mattys et al. (2005).

Distal Speech Rate and the proximal acoustic cue of F_0 Dip did not interact as predicted for the analysis which also incorporated the effects of each Sentence Fragment. However, an analysis of the "rich" fragment alone show that the interaction between the two variables was marginally significant, such that the stronger the proximal acoustic cue (i.e., the larger the size of the F_0 dip), the smaller the effect of distal speech rate on whether participants heard a critical word. This is particularly apparent in Figure 5(a), where the critical word report rate with an F_0 Dip of 7 Hz was closer to the 14 and 21 Hz F_0 Dips with a faster Distal Speech Rate, but was closer to the 0 Hz F_0 Dip with a slower Distal Speech Rate. The unreliable interaction may simply reflect the lower power to detect significant effects for each sentence fragment considered individually compared with collapsing responses across fragments. Finally, the F_0 cue was found to be more effective in eliciting perceptions of critical words for the "rich" fragment than for the "rude" fragment. One potential explanation for this difference is that the relative magnitudes of F_0 dips for the "rich" fragment were one semitone greater than those for the "rude" fragment.

EXPERIMENT 3

Experiment 3 involved manipulation of distal speech rate in combination with proximal word duration (i.e., the length of the segmental material within the target region). Durational lengthening has been observed to occur at prosodic boundaries including the word boundary (Byrd & Saltzman, 2003; Shatzman & McQueen, 2006; Turk & Shattuck-Hufnagel, 2000). Vowel duration is a clear example of a segmental cue in languages where vowel length is phonemic, such as Finnish or Arabic. In English, however, it is not clear whether vowel length is a segmental, subsegmental, or suprasegmental cue or whether its status may change under some conditions. Here, it was explored whether the distal speech rate would still affect critical word report rates and whether those effects would be independent of a proximal manipulation that was itself temporal in nature. It might be the case, for example, that listeners would interpret the proximal increase in target segmental duration as a decrease in proximal speech rate. They might, therefore, compute the existence of a critical word by way of comparison of relative distal and proximal speech rates. Alternatively, distal speech rate and proximal word duration may be independent of each other, such that the two cues do not interact. Again, by comparing the strength of the distal speech rate and proximal acoustic cues, this will help situate both of them with respect to each other on a word segmentation cue hierarchy.

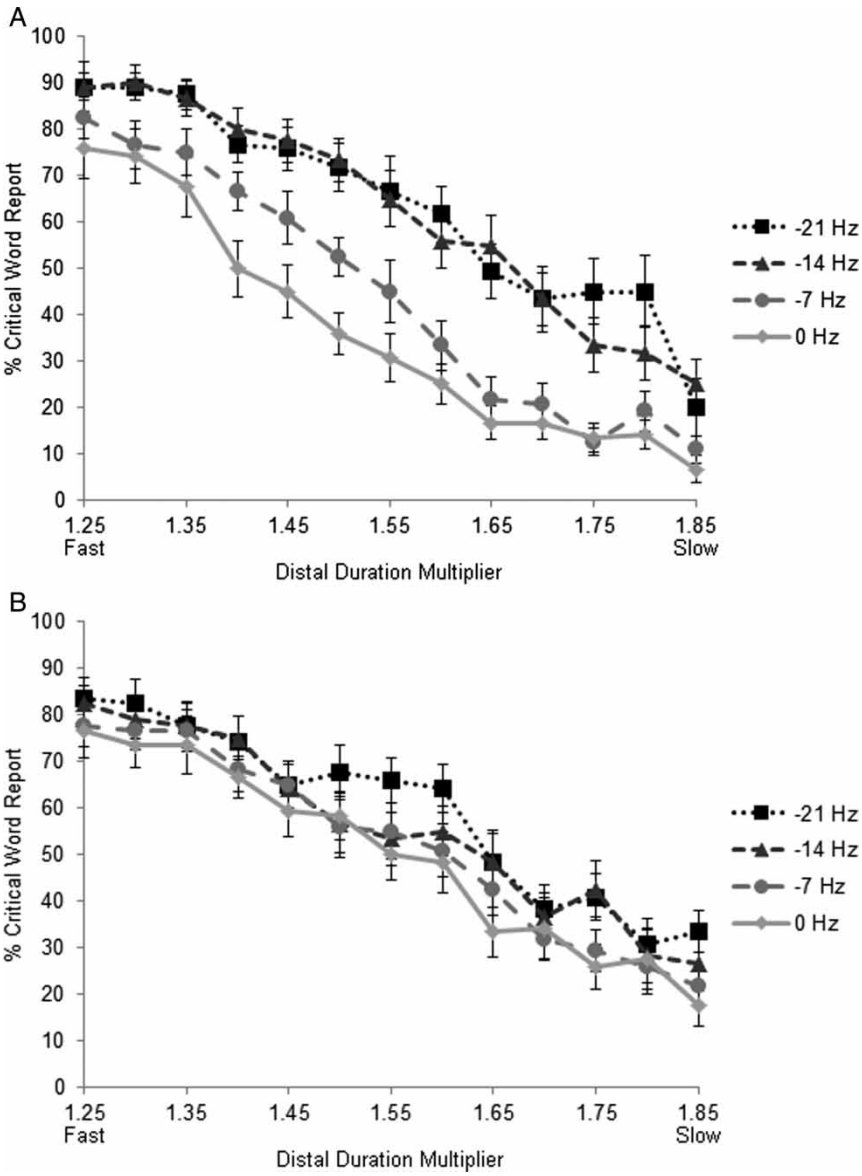

Figure 5. Mean percentage of participant critical word reports as a function of Distal Speech Rate and F_0 Dip, in Hz, for the "rich" (a) and "rude" (b) sentence fragments in Experiment 2.

Method

Participants

Thirty participants (21 female, 9 male) were recruited for research credit at Michigan State University. All were native speakers of English who self-reported normal hearing and were at least 18 years of age ($M = 20.3$ years, Range = 18–29 years).

Stimuli and design

Distal speech rates were identical to Experiments 1 and 2, and four levels of proximal word duration were chosen, again setting up a 2 (Sentence Fragment) $\times$ 4 (Word Duration) $\times$ 13 (Distal Speech Rate) design. Here, distal speech rate and word duration were manipulated simultaneously in Praat. For each combination of sentence fragment and distal duration multiplier, a proximal duration multiplier was assigned, which worked analogously to distal duration multipliers, with target word duration being multiplied by one of four multipliers: 1.00, 1.25, 1.50, and 1.75. These values were selected using two criteria: maximisation of the range of possible percepts, and the creation of a continuum of proximal duration multipliers representative of the approximate span of the distal duration multipliers, to test the hypothesis that the strength of the proximal word duration cue to word boundaries would be relative to the strength of the distal speech rate cue (Figure 6).

Apparatus

The apparatus used was identical to Experiment 2.

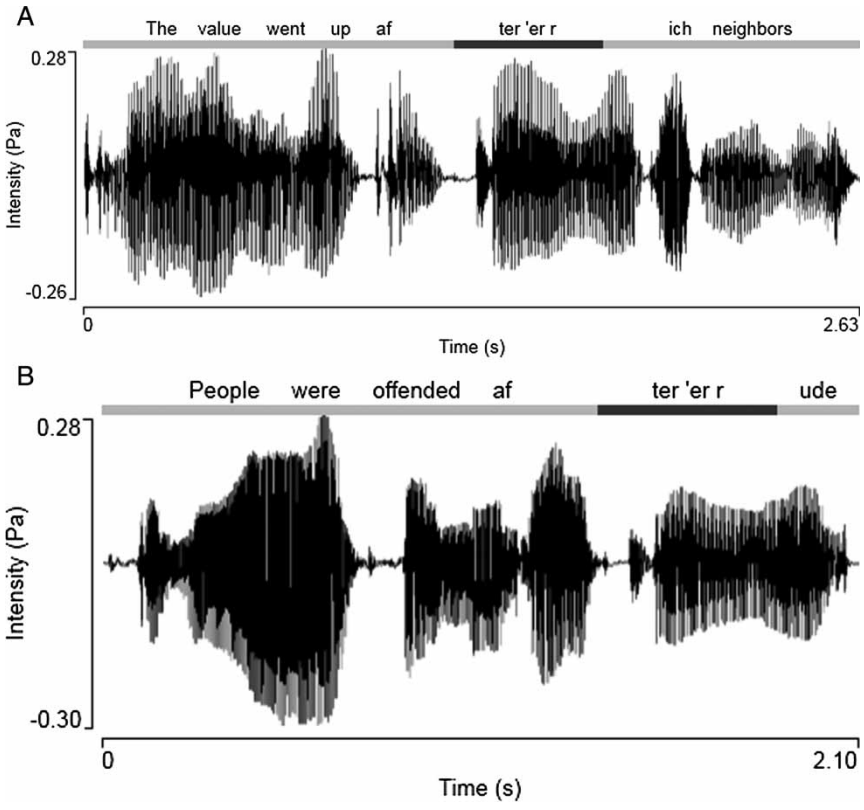

Figure 6. Waveform representations of examples from the "rich" (a) and "rude" (b) sentence fragments with a distal word rate factor of 1.55 and proximal word duration factor of 1.75. The lines above the spectrogram give the approximate durations of the context (light gray) and target (dark gray) regions, with the text above that denoting the orthographic transcriptions of each region.

Procedure

The procedure was the same as in Experiment 2.

Results

Figure 7 shows the effects of Word Duration and Distal Speech Rate on the percentage of critical word reports for the "rich" and "rude" Sentence Fragments. Logit mixed-effects models were used to determine the effects of the fixed factors Word Duration, Distal Speech Rate, and Sentence Fragment, with subjects as a random factor. The saturated model, which incorporated all fixed factors and their

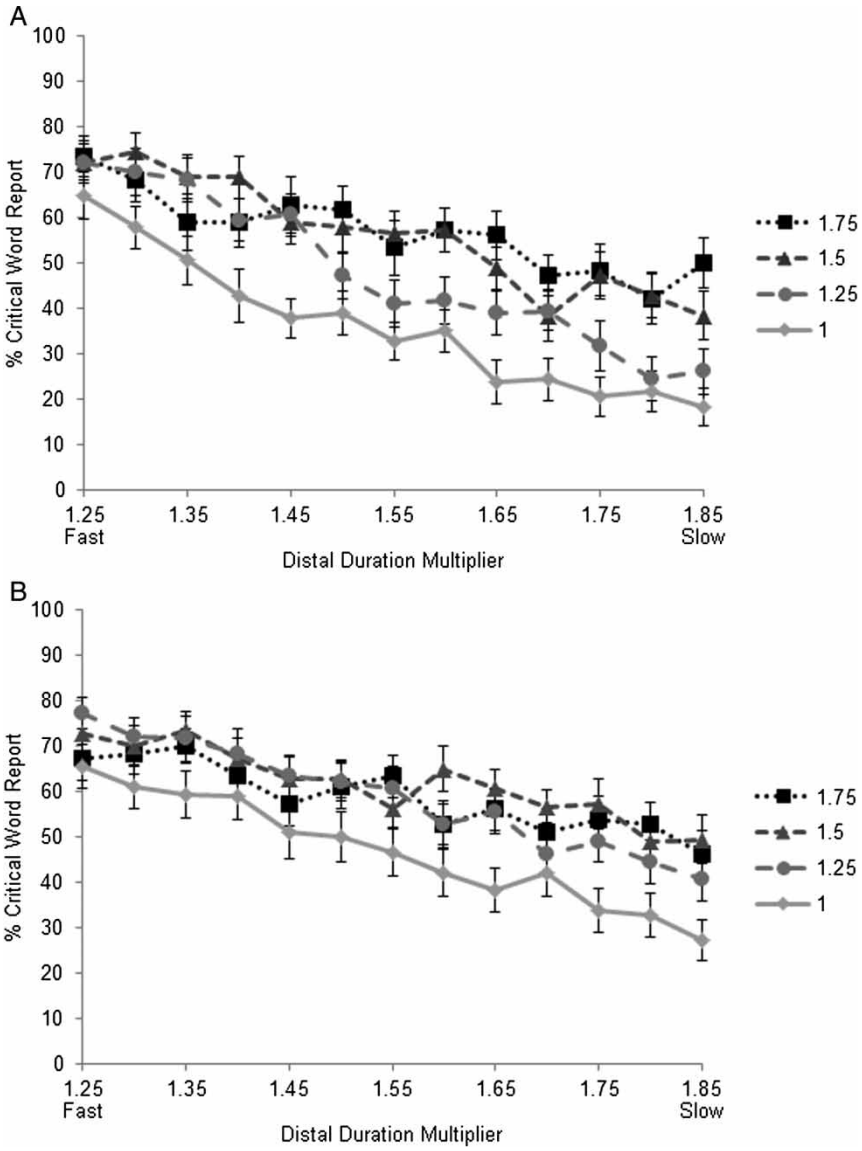

Figure 7. Mean percentage of participant critical word reports as a function of Distal Speech Rate and Word Duration for the "rich" (a) and "rude" (b) sentence fragments in Experiment 3.

interactions, once more provided the best fit to the data. As in the previous two experiments, slowing Distal Speech Rate reduced critical word reports ($b = –0.11$, $p < .01$), whereas increasing the magnitude of the proximal cue, in this case Word Duration, produced more critical word reports ($b = 0.17$, $p < .01$). The "rude" sentence fragment had higher critical word report rates ($M = 57\%$, $SD = 14\%$) than the "rich" fragment ($M = 49\%$, $SD = 16\%$), $b = –0.34$, $p < .01$. The interaction of Distal Speech Rate and Word Duration was reliable ($b = 0.022$, $p < .01$), and notably was reliable for both fragments ("rude": $b = 0.02$, $p < .01$; "rich": $b = 0.03$, $p < .01$). Stimuli with longer Word Duration were less sensitive to the effects of Distal Speech Rate than were stimuli with shorter Word Duration. That is, increasing Word Duration tended to attenuate the effect of Distal Speech Rate cue on critical word reports.

Discussion

In Experiment 3, the distal speech rate effect persisted even when the proximal cue also involved a manipulation of temporal information. The proximal cue of word duration also significantly affected whether participants reported a critical word. The two cues were not independent, with less discrimination between levels of Distal Speech Rate for larger values of Word Duration than smaller values when the interaction was analyzed across sentence fragments. That the interaction between these cues was reliable even when analyzed separately for both sentence fragments speaks to the strength of this interaction. This is consistent with the idea that listeners used a comparison of the relative speed of the distal and proximal regions to determine where to place word boundaries, rather than using some sort of absolute tempo metric. The strength of each cue depends on the strength of the other, suggesting that, regardless of which cue is stronger in word segmentation, the stronger cue is influenced by the weaker.

Though the lengthening in Experiment 3 may bear some similarity to that used in, for example, contrastive emphasis (which persists across speech rates, Cummins, 1999), contrastive emphasis is unlikely to explain the results of Experiment 3, as it is unlikely to co-occur with the reduced speech used here. Regardless of whether listeners interpreted the increased lengthening on "*er*" as stress or emphasis (Cummins, 1999) or as boundary-related lengthening (Turk & Shattuck-Hufnagel, 2000), it remains true that the distal speech rate manipulation caused listeners to differentially report hearing an extra word within the target region for a given proximal word length.

EXPERIMENT 4

In Experiment 4, an unambiguously segmental cue, that of frication noise, was compared to distal speech rate, in order to see whether proximal acoustic cues can rank higher than distal prosodic cues under any circumstances. The combination of noise and intensity drop reliably corresponds to a fricative consonant in English (Stevens, 1998). Thus, across the four experiments, it was predicted that the proximal acoustic manipulation in Experiment 4 (/h/-like frication noise) would be most effective in segmentation due to its conclusively segmental nature, compared to the indeterminate cues of Experiments 1–3. It was therefore expected that the frication noise cue could clearly and distinctly be put on Tier II of the Mattys et al. (2005) hierarchy, in contrast to the cues explored in Experiments 1–3. Frication was selected in order to evoke recovery of the original /h/ in each sentence fragment through combining a naturalistic /h/ sound with versions of each stimulus, causing changes in

intensity and high-frequency noise for the sentence fragments. Furthermore, the manipulation used in Experiment 4 combined two acoustic dimensions of variation (frication noise and a local intensity dip). These dimensions were expected to reinforce each other, resulting in a stronger effect on word segmentation and lexical perception than the single dimension of variation in the other three experiments. Still, given its robustness so far, it was predicted that distal speech rate effects would persist, even when listeners heard definitive segmental cues to word boundary placement.

Methods

Participants

Twenty-six participants (22 female, 4 male) were recruited for research credit at Michigan State University. All were native speakers of English who self-reported normal hearing and were at least 18 years of age ($M = 21.9$ years, Range = 20–40 years).

Stimuli and design

Once more, this experiment had a 2 (Sentence Fragment) $\times$ 4 Noise Strength) $\times$ 13 (Distal Speech Rate) design. The base sentence fragments and distal duration multipliers were identical to the previous experiments.

Natural voiceless /h/ tokens were selected from the talkers who had produced the original speech fragments for materials in Dilley and Pitt (2010); voiceless tokens were chosen to avoid perceptual effects of F_0 discontinuity which could have arisen from voiced /h/ tokens. For the "rich" speech fragment, a token of /h/ spoken by the same talker in another sentence context in materials recorded for the original Dilley and Pitt (2010) experiments was selected as the basis of proximal manipulations. For the "rude" speech fragment, because of the infrequent use of voiceless /h/ tokens in that talker's productions, a token of /h/ from the same sentence spoken by a different talker was used. Both tokens had their intensities adjusted so as to create a 10 ms intensity transition into and out of the /h/ token, to make the transition into and out of the token less abrupt.

Natural voiceless /h/ tokens show a change in source spectrum to a glottal frication source as well as a dip in intensity relative to source properties of context vowels. Therefore, a multi-step approach was taken. First, the duration of the target region of each stimulus was augmented in length by the duration of the selected /h/ token through multiplying the target region's duration by the relevant multiplicative factor in Praat. The manipulation of target duration was performed simultaneously to the imposition of the 13 possible distal duration multipliers. To mimic changes in intensity that occur in naturally-produced /h/ segments, a 5 dB intensity dip was created at the midpoint of each target region, similar to the intensity dips created for Experiment 1. The trough length of this intensity dip was set to equal the duration of the natural /h/ token being spliced in.

Finally, a continuum of stimuli ranging in strength from no frication noise to high frication noise was created, with frication noise being centred at the temporal midpoint of the target region. The stimuli with no /h/ token spliced into them, and therefore with the lowest level of frication noise (i.e., none), will be referred to as the adjusted stimuli, to differentiate them from stimuli from previous experiments without proximal word duration or intensity adjustment. The second-lowest level of frication

noise was the result of inserting the /h/ token directly around the midpoint of the target region of the adjusted stimuli. The relatively lower intensity of the frication noise in comparison to the target led to a weak /h/ percept. An approach based on the principle of signal-to-noise ratio (SNR) was used to create the other steps on the frication noise continuum.

For this experiment, the "signal" was the portion of the target region within the modified stimulus around the midpoint of the region, where the frication noise was to be inserted; the "noise" was the /h/ token itself. The SNR, then, reflected the ratio of the power of the part of the adjusted target where the frication noise was to be inserted (the "signal") to the power of the /h/ token (the "noise"). Praat was used to multiply the amplitude of the adjusted target "signal" by a scalar designed to lower the power of the "signal" region to match a given SNR, thereby decreasing the strength of the "signal" region and producing a stronger /h/ percept. One of the SNRs selected was 0 (i.e., the power of the adjusted target and the power of the /h/ token were the same). Another SNR chosen was equal to half the SNR between the adjusted target and frication noise, representing the midpoint in the SNR between the other two steps. The SNR values used varied between the sentence fragments used, as initial SNRs differed as a result of the different powers of the adjusted target regions and /h/ tokens selected. They also varied slightly from distal duration multiplier to distal duration multiplier because of inconsistencies in the handling of power by Praat. Table 1 gives the means and standard deviations for the SNRs used (Figure 8).

Apparatus

The apparatus used were identical to Experiments 2 and 3.

Procedure

The procedure was identical to that of Experiments 2 and 3.

Results

Figure 9 shows the effects of Noise Strength and Distal Speech Rate on the percentage of critical word reports for the "rich" and "rude" Sentence Fragments. Different lines represent different strengths of Noise Strength. The effects of the fixed factors Noise Strength, Distal Speech Rate, and Sentence Fragment were determined using logit mixed-effect models, with subjects as a random factor. The saturated model was the

TABLE 1
Means and standard deviations for the SNRs of Experiment 4 stimuli, arranged by sentence fragment and Noise Strength

		Noise Strength				
		Weak		*Moderate*		
Phrase	None	*Mean*	*SD*	*Mean*	*SD*	Strong
"rich"	n/a	+17.51	0.02	+8.76	0.01	0
"rude"	n/a	+11.82	0.05	+5.91	0.03	0

Notes: No standard deviation is given for stimuli with None or Strong Noise Strength because both classes of stimuli had no variation in SNR.

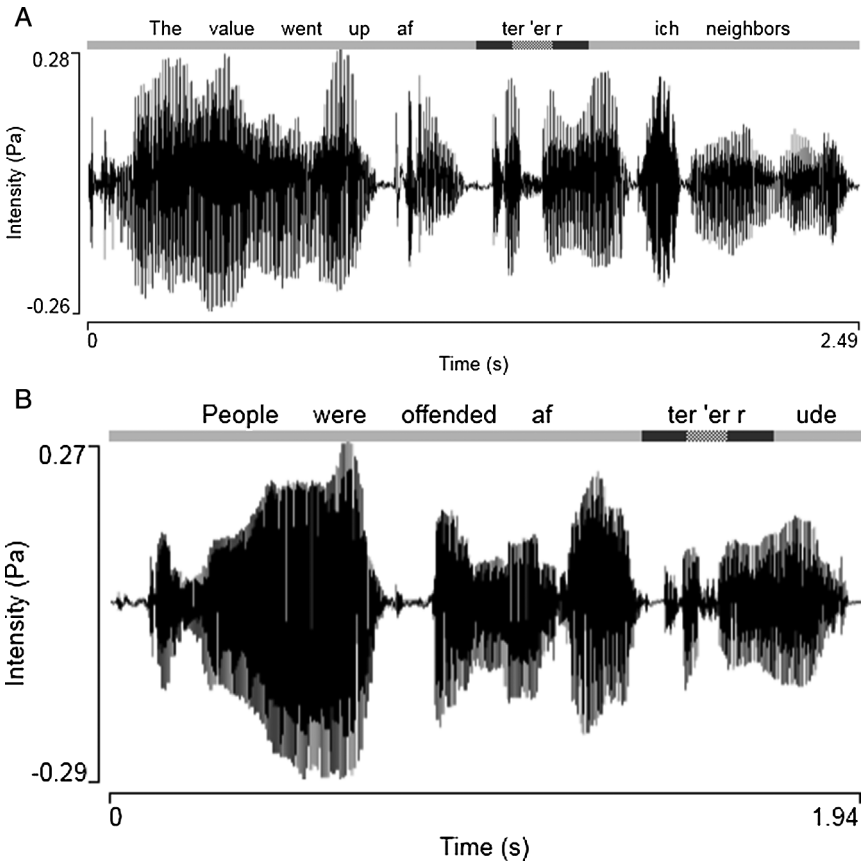

Figure 8. Waveform representations of examples from the "rich" (a) and "rude" (b) sentence fragments with a distal word rate factor of 1.55 and SNR of 0. The lines above the spectrogram give the approximate durations of the context (light gray) and target (dark gray) regions, with the text above that denoting the orthographic transcriptions of each region. The dotted area within the target region of the line reflects the locus of /h/ manipulation, including the surrounding intensity dip and addition of frication noise.

best-fitting model. As in the preceding experiments, slowing Distal Speech Rate decreased critical word reports ($b = –0.11$, $p < .01$). For the Noise Strength manipulation, more noise resulted in more critical word reports ($b = 0.99$, $p < .01$). Though the "rude" sentence fragment ($M = 70\%$, $SD = 13\%$) had higher critical word report rates than the "rich" sentence fragment ($M = 65\%$, $SD = 11\%$), the difference was not reliable ($b = –0.069$, $p = .44$). Distal Speech Rate and Noise Strength did not interact ($b = –0.002$, $p = .70$). However, there was a reliable three-way interaction between Sentence Fragment, Distal Speech Rate and Noise Strength ($b = 0.037$, $p < .01$). To unpack this interaction, data from each sentence fragment were analyzed separately. This analysis revealed a reliable interaction between Distal Speech Rate and Noise Strength for the "rich" fragment ($b = 0.04$, $p < .01$), but not for the "rude" fragment ($b = –0.001$, $p = .88$).

Finally, the Medium and High levels of Noise Strength were also analyzed separately, in order to determine whether the effects of Distal Speech Rate, and any interactions between the factors, were apparent even when critical word report rate was near ceiling. The best-fitting model for the Medium and High levels of Noise Strength was one in which Distal Speech Rate, Noise Strength, and Sentence

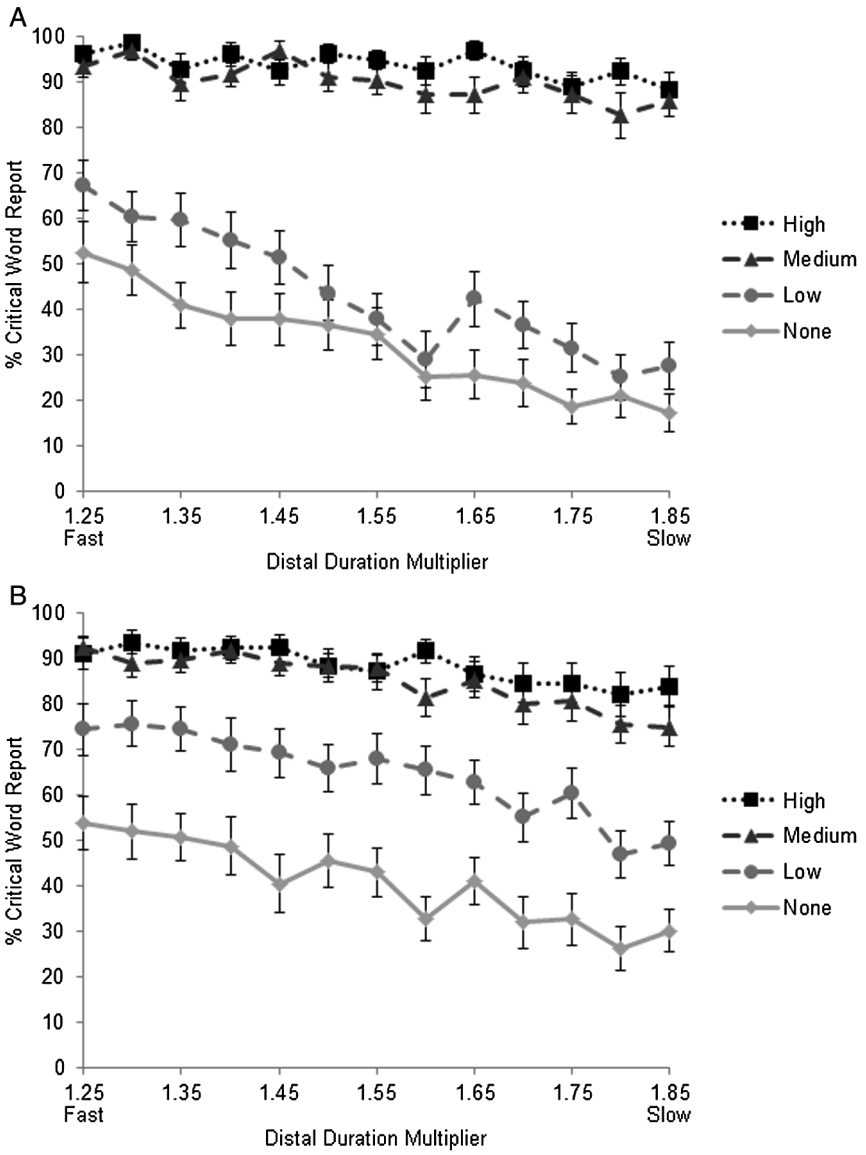

Figure 9. Mean percentage of participant critical word reports as a function of Distal Speech Rate and Noise Strength for the "rich" (a) and "rude" (b) sentence fragments in Experiment 4.

Fragment, and the interaction between Distal Speech Rate and Noise Strength were used in the model, but Sentence Fragment was not crossed with the other two factors. As before, increased Noise Strength led to more critical word reports ($b = 0.41$, $p < .01$). Despite the fact that critical word report rate was close to ceiling, slowing Distal Speech Rate still reliably decreased critical word report rate ($b = -0.10$, $p < .01$). The "rude" sentence fragment had lower critical word report rates than the "rich" sentence fragment ($b = 0.63$, $p < .01$). However, there was no significant interaction between Distal Speech Rate and Noise Strength ($b = -0.02$, $p = .30$).

Discussion

In Experiment 4, distal speech rate still motivated placement of word boundaries, and adding two related proximal cues to the speech stream—frication noise and intensity dip—also led to an increase in the frequency of word boundary placement. Though the two cues did not directly trade off, there was a three-way interaction between the distal cue, the proximal cue, and the sentence fragment, such that there was an interaction between the distal and proximal cues for the "rich" fragment but not for the "rude" fragment. For the "rich" fragment, Distal Speech Rate had a robust effect on word boundary perception for the two conditions with smallest Noise Strength. For the largest two Noise Strengths, varying Distal Speech Rate maintained an effect on critical word report rates even though critical word reports were near ceiling for all distal speech rates. However, for the higher levels of Noise Strength, there was no interaction between Distal Speech Rate and Noise Strength, and interactions between Sentence Fragment and the other two cues were unreliable. The lack of a two-way interaction between the proximal and distal cues when both sentence fragments were considered together may reflect differences between the two sentence fragments in cue strength; further studies will need to assess the potential interaction between the cues in a wider variety of contexts. Compared to Experiments 1–3, here the predictions of Mattys et al. (2005) are relatively supported; the Tier II cue of segmental noise is a larger determinant of word boundary placement than the distal speech rate cue, which, under the hierarchy, would likely fall in Tier III.

GENERAL DISCUSSION

The present studies involved a simultaneous manipulation of distal speech rate, building on the work of Dilley and Pitt (2010), as well as one of several proximal acoustic cues, in order to investigate the independent and interactive effects of these cues in vowel-initial word boundary placement. In the present studies, listeners selected one of two lexical interpretations for each test sentence fragment that was spoken with a critical vowel-initial word (here, a word with an elided initial consonantal onset), in order to indicate whether they had heard the word. Results across four experiments extend findings of Dilley and Pitt (2010) by showing that distal speech rate can be a robust determinant of whether a spectrally reduced, vowel-initial word is heard, even in the face of conflicting acoustic cues. Moreover, the current work extends recent studies of distal prosodic cues (Dilley & McAuley, 2008; Dilley et al., 2010; Dilley & Pitt, 2010) by showing that distal speech rate effects persisted under a wide range of proximal acoustic realisations of the critical word onset: intensity (Experiments 1 and 4), F_0 (Experiment 2), proximal word duration (Experiment 3), and frication noise (Experiment 4). By using direct manipulation of several proximal acoustic variables, the present results demonstrate causally that these proximal acoustic variables have effects on lexical perception and word segmentation. This contrasts with previous, largely observational studies, which successfully demonstrated that each individual cue studied here is deployed in sentence production (Byrd & Saltzman, 2003; Dilley et al., 1996; Fougeron & Keating, 1997; Hanson, 2009; Turk & Shattuck-Hufnagel, 2000) but which have left many questions unanswered about their use by listeners (see Hillenbrand & Houde, 1996, Pardo & Fowler, 1997; Shatzman & McQueen, 2006).

The pattern of interactions across experiments suggests that the effectiveness of distal speech rate as a segmentation cue depends on the type and strength of a given proximal acoustic cue. The present experiments therefore clearly demonstrate that different kinds of cues "trade off" dynamically to determine word segmentation. This is the first series of studies to specifically examine the robustness of the distal speech rate effect on word segmentation in the face of conflicting proximal acoustic cues, as well as the relative strength of those cues, allowing for a more-nuanced examination of the strength of the distal speech rate effect. The results clearly show that distal speech rate is a robust cue to word boundary placement which can have large effects on the number of words (and word boundaries) perceived by a listener, even in the face of conflicting proximal acoustic information.

The results revealed that distal and proximal cues sometimes interact, but not always. Significant interactions were found between the distal and proximal variables in Experiments 3 and 4, and a marginally significant interaction was observed in Experiments 1 and 2 (for the "rich" fragment). The variability of the results may be a function of the large number of distal speech rates employed for this experiment, the small number of sentence fragments, and/or the reduced power of analyses to detect significant effects when analyzing responses separately by fragment.

Most of the effects observed in this study were found to hold for both sentence fragments examined, with a few exceptions. Though the sentence fragments yielded different magnitudes for the main effects and interactions observed, particularly for proximal acoustic cues, the direction of those main effects and interactions were largely consistent across both test sentence fragments. To the extent that there were differences in effects between the sentence fragments used, the results showed that the effectiveness of distal speech rate and proximal acoustic cues can depend on the specific lexical speech context and pre-existing acoustic information in each fragment. Further exploration of the interplay between distal speech rate and proximal acoustic cues across a wider variety of proximal and distal contexts would help elaborate on the strength and applicability of each of these types of cues in normal speech perception.

As discussed in the introduction, a number of frameworks have begun to be developed which attempt to account for effects of multiple sources of information on phonetic perception and/or word segmentation (Clayards et al., 2008; Feldman et al., 2009; Goldwater et al., 2009; Mattys et al., 2005; Norris & McQueen, 2008; Toscano & McMurray, 2010). Of these, the most fully elaborated model so far proposed for word segmentation is the promising framework of Mattys et al. (2005), which combines the strengths of lexical (e.g., TRACE, McClelland & Elman, 1986; Shortlist, Norris, 1994; Shortlist B, Norris & McQueen, 2008) and pre-lexical (e.g., Christiansen et al., 1998) theories of word segmentation. That proposal aimed to synthesise multiple segmentation cues into a fairly strict hierarchy of cue strengths, and has three principal strengths. First, completeness; by combining insights from both lexical and pre-lexical theories, the Mattys et al. (2005) hierarchy served as a near-comprehensive list of the cues which had been shown to affect word segmentation. Second, stratification; by classifying cues into one of three separate layers (Tier I, segmental; Tier II, segmental and acoustic-phonetic; Tier III, metrical prosodic), Mattys et al. (2005) allowed for a relatively simple, uncluttered division of cues, with each type of cue being classified as wholly a member of exactly one tier. Third, ranking; by ranking cues with respect to each other, and postulating that higher-level cues almost always overpowered lower-ranked ones, Mattys et al. (2005) allowed for robust predictions about the relative power of each type of cue.

The experiments presented here, however, indicate that revisions need to be made to the Mattys et al. (2005) hierarchy. First, distal prosodic cues represent a significant point of incompleteness for the hierarchy. Their absence is understandable, considering how recently the effects of the distal cues were demonstrated. As the results from Dilley et al. (2010) and the present paper show, distal prosodic cues are potent cues to word boundaries. Despite their prosodic nature, it seems clear that they pattern differently from so-called "metrical" prosodic cues discussed in Mattys et al. (2005). As such, the framework could be improved through the addition of distal prosodic cues to the hierarchy, perhaps as a separate tier.

Moreover, the experiments here call into question the strong hypothesis assumed in Mattys et al. (2005) that each cue can be stratified into a single level of a segmentation hierarchy. The distal speech rate effect could not clearly be sorted into one of the three tiers discussed; it is not knowledge-based (Tier I), not clearly segmental (Tier II), and though prosodic, is not linked to the metrical prosody described in the original hierarchy, that is, word stress (Tier III). However, more subtly, the proximal acoustic cues manipulated for this experiment also cannot be reconciled with any one tier. Consider the discussion of the segmental status of the proximal acoustic manipulations. It is not clear whether, for example, F_0 is segmental (Tier II), subsegmental (perhaps also Tier II), or suprasegmental (Tier III) in nature. Placing it on the hierarchy, then, is quite challenging, especially considering it was outranked in Experiment 2 by distal speech rate, a cue that would most likely fall under Tier III under the current formulation of the hierarchy.

Finally, the results here speak to the difficulty of assigning cues with gradient rankings to a strictly-ranked hierarchy, something that Mattys et al. (2005) were clearly aware of. Indeed, subsequent results by Mattys and Melhorn (2007) suggested that varying cue strength might result in different patterns of relative ranking of segmentation cues; their data suggested that strong segmental acoustic information favouring a word boundary could apparently outweigh knowledge-based information, in contrast to the earlier working hypotheses of Mattys et al. (2005). Here, across experiments, both distal speech rate and proximal acoustic cues were consistent in that they affected critical word report rates, but their relative rankings differed in a graded fashion as a function of the strength of each cue. The gradient rankings of each cue, as opposed to the absolute rankings exemplified by a strict hierarchy, can be seen easily by examining the interactions between distal and proximal cues across experiments. The use of distal and proximal cues traded off in a way that was unexpected given a narrow reading of the Mattys et al. (2005) hierarchy, but reminiscent of trade-offs in segmental phonetics (Miller, 1994; Repp, 1982) and subsequent work by Mattys and colleagues (e.g., Mattys & Melhorn, 2007). The strength of each cue appears to quantitatively affect the ranking of each cue; there is no qualitative point beyond which any particular cue is simply "used" or "not used", as implied by a strict hierarchy.

In addition to updating the Mattys et al. (2005) hierarchy, it might be fruitful to explore developing a flexible word segmentation model that allows for the use of multiple cues to word segmentation with changing cue strengths that unfold over time. One promising framework is the adaptive resonance theory (ART) model ARTWORD (Grossberg & Myers, 2000), which explicitly takes into account the effects of distal phonetic information on phonetic judgments. Similarly, endogenous oscillator approaches to prosodic boundary placement (e.g., Byrd & Saltzman, 2003) provide a window to understanding the effects of distal temporal context on word segmentation. In addition, Bayesian models of phonetic perception (Clayards et al.,

2008; Feldman et al., 2009; Toscano & McMurray, 2010) and mixture-of-Gaussians models (e.g., Toscano & McMurray, 2010) may be particularly suitable frameworks for modelling the kinds of results demonstrated here. In particular, further development of Bayesian approaches to word segmentation (Goldwater et al., 2009; Shortlist B: Norris & McQueen, 2008) would seem particularly promising given our results.

In summary, the present results indicate that different types of prosodic cues (e.g., proximal intensity vs. distal speech rate) can have very different segmentation strengths, depending on the strengths of each cue. These findings provide evidence that different types of cues interact or "trade off" in the word segmentation process. Our results thus indicate that a wider variety of specific acoustic-phonetic and prosodic cues will need to be investigated in order to generate a comprehensive picture of how word boundary cues are ranked and integrated with one another during processing. In our studies, though each cue could have independent effects on word segmentation and lexical perception, the congruent combination of these cues yielded the strongest lexical percepts. These findings suggest that listeners use all acoustic information available to them regarding the structure of the speech signal in order to make sense of the lexical content of speech.

Manuscript received 8 March 2011
Revised manuscript received 21 February 2012
First published online 3 August 2012

REFERENCES

Bates D., Maechler, M., Bolker, B., & Vasishth, S. (2011). *lme4: Linear mixed-effects models using S4 classes*. Retrieved June 6, 2011, from http://CRAN.R-project.org/package = lme4

Boersma, P. P., & Weenink, D. (2009). Praat, a system for doing phonetics by computer (Version 5.1) [Computer program]. Retrieved August 2, 2009, from http://www.praat.org/

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180. doi:10.1016/S0095-4470(02)00085-2

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2/3), 221–268. doi:10.1080/016909698386528

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. doi:10.1016/j.cognition.2008.04.004

Cole, R. A., Jakimik, J., & Cooper, W. E. (1980). Segmenting speech into words. *Journal of the Acoustical Society of America*, *67*(4), 1323–1332. doi:10.1121/1.384185

Cummins, F. (1999). Some lengthening factors in English speech combine additively at most rates. *Journal of the Acoustical Society of America*, *105*(1), 476–480. doi:10.1121/1.424576

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121. doi:10.1037/0096-1523.14.1.113

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*(1), 218–244. doi:10.1037/0096-1523.28.1.218

Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*(3), 274–294. doi:10.1016/j.jml.2010.06.003

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*(3), 294–311. doi:10.1016/j.jml.2008.06.006

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. doi:10.1177/0956797610384743

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*(4), 423–444. doi:10.1006/jpho.1996.0023

Feldman, N. H., Griffiths, T., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. doi:10.1037/a0017196

Fernandes, T., Ventura, P., & Kolinsky, R. (2007). Statistical information and coarticulation as cues to word boundaries: A matter of signal quality. *Perception and Psychophysics*, *69*, 856–864. doi:10.3758/BF03193922

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*(6), 3728–3740. doi:10.1121/1.418332

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. doi:10.1016/j.cognition.2009.03.008

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, *107*(4), 735–767. doi:10.1037/0033-295X.107.4.735

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *Journal of the Acoustical Society of America*, *125*(1), 425–441. doi:10.1121/1.3021306

Hillenbrand, J. M., & Houde, R. A. (1996). Role of F_0 and amplitude in the perception of intervocalic glottal stops. *Journal of Speech and Hearing Research*, *39*(6), 1182–1190.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Kenstowicz, M. J. (1994). *Phonology in generative grammar*. Cambridge, England: Blackwell.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica*, *5*(Suppl.), 5–54. doi:10.1159/000258062

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Mattys, S. L., & Jusczyk, P. W. (*2001*). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 644–655. doi:10.1037/0096-1523.27.3.644

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*(1), 554–567. doi:10.1121/1.2735105

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500. doi:10.1037/0096-3445.134.4.477

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi:10.1016/0010-0285(86)90015-0

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46. doi:10.1006/jmla.1998.2568

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory and Cognition*, *39*(6), 1085–1093. doi:10.3758/s13421-011-0074-3

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Erlbaum.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, *50*(1–3), 271–285. doi:10.1016/0010-0277(94)90031-0

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*(6), 505–512. doi:10.3758/BF02308147

Newman, R. S.,Sawusch, J. R., & Wunnenberg, T. (2011). Cues and cue interactions in segmenting words in fluent speech. *Journal of Memory and Language*, *64*(4), 460–476. doi:10.1016/j.jml.2010.11.004

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. doi:10.1016/0010-0277(94)90043-4

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. doi:10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243. doi:10.1006/cogp.1997.0671

Pardo, J. S., & Fowler, C. A. (1997). Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perception and Psychophysics*, *59*(7), 1141–1152. doi:10.3758/BF03205527

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978–996. doi:10.1037/a0021923

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110. doi:10.1037/0033-2909.92.1.81

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 621–637. doi:10.1037/0096-1523.4.4.621

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. doi:10.1006/jmla.1996.0032

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*(1), 51–89. doi:10.1016/S0010-0277(03)00139-2

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, *68*(1), 1–16. doi:10.3758/BF03193651

Shimizu, K., & Dantsuji, M. (1980). A study on perception of internal juncture in Japanese. *Studia Phonologica*, *14*(1), 1–15.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5): 1074–1095. doi:10.1037/0096-1523.7.5.1074

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. doi:10.1111/j.1551-6709.2009.01077.x

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440. doi:10.1006/jpho.2000.0123

Viemiester, N. F., & Bacon, S. P. (1988). Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones. *Journal of the Acoustical Society of America*, *84*(1), 172–178. doi:10.1121/1.396961

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, *95*(5), 2694–2701. doi:10.1121/1.409838