Journal of Memory and Language 63 (2010) 274–294

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation ☆

Laura C. Dilley [a,b,*], Sven L. Mattys [c], Louis Vinke [d]

[a] Department of Communicative Sciences and Disorders, Michigan State University, United States
[b] Department of Psychology, Michigan State University, United States
[c] Department of Experimental Psychology, University of Bristol, UK
[d] Department of Psychology, Bowling Green State University, United States

ARTICLE INFO

Article history:
Received 17 February 2010
revision received 8 June 2010
Available online 15 July 2010

Keywords:
Prosody
Word segmentation
Lexical access
Word recognition
Rhythm
Perceptual organization

ABSTRACT

Recent work shows that word segmentation is influenced by distal prosodic characteristics of the input several syllables from the segmentation point (Dilley & McAuley, 2008). Here, participants heard eight-syllable sequences with a lexically ambiguous four-syllable ending (e.g., *crisis turnip* vs. *cry sister nip*). The prosodic characteristics of the initial five syllables were resynthesized in a manner predicted to favor parsing of the final syllables as either a monosyllabic or a disyllabic word; the acoustic characteristics of the final three syllables were held constant. Experiments 1a–c replicated earlier results showing that utterance-initial prosody influences segmentation utterance-finally, even when lexical content is removed through low-pass filtering, and even when an on-line cross-modal paradigm is used. Experiments 2 and 3 pitted distal prosody against, respectively, distal semantic context and prosodic attributes of the test words themselves. Although these factors jointly affected which words participants heard, distal prosody remained an extremely robust segmentation cue. These findings suggest that distal prosody is a powerful factor for consideration in models of word segmentation and lexical access.

Introduction

It is well-known that speech input lacks clear and reliable markers of word boundaries (Cole & Jakimik, 1980; Klatt, 1980). How human listeners locate boundaries between words has been shown to be driven by both characteristics of the speech signal and lexical-sentential knowledge (cf. Mattys, White, & Melhorn, 2005). Signal-driven characteristics include sub-lexical cues probabilistically associated with word boundaries, e.g., allophonic or phonotactic regularities (e.g., Christiansen, Allen, & Seidenberg, 1998; Mattys, 2004; Quene, 1992, 1993), as well as prosodic cues (e.g., pitch and timing), such as lexical stress (e.g., Cutler & Butterfield, 1992; Cutler & Norris, 1988) and phrasal boundaries (Cho, McQueen, & Cox, 2007; Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Gout, Christophe, & Morgan, 2004; Millotte, Rene, Wales, & Christophe, 2008). Knowledge-driven processes include the use of lexical-semantic knowledge and syntactic expectations (Mattys, Melhorn, & White, 2007). Uncovering the full set of cues that listeners use to identify word boundaries and determining how they interact with one another is an important step in understanding how humans communicate using spoken language.

The contribution of *context* to word segmentation is most often associated with knowledge-driven processes, such as lexical-semantic knowledge and syntax; these attributes have been shown to have large effects on segmentation behavior (Mattys, White, et al., 2005; Mattys et al., 2007) and to operate over long distances involving multiple syllables (Mattys et al., 2007). However,

☆ The authors thank three anonymous reviewers for helpful input and gratefully acknowledge the support of NSF grant BCS 0847653 to L. Dilley.

* Corresponding author at: Department of Communicative Sciences and Disorders, Michigan State University, East Lansing, MI 43403, United States. Fax: +1 517 353 3176.
E-mail address: ldilley@msu.edu (L.C. Dilley).

0749-596X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2010.06.003

contextual cues can also be acoustic in nature, e.g., the prosodic context in which a to-be-segmented word occurs. Prosodic context was omitted from the segmentation hierarchy of Mattys, White, et al. (2005), yet may be significant for segmentation.

The phonetics and phonology literature suggests that the structure of prosodic context often exhibits regularities in pitch, duration, and/or amplitude. That is, speech intonation and rhythm often show what listeners perceive to be patterning (Couper-Kuhlen, 1993; Dainora, 2001; Pierrehumbert, 2000). For example, listeners tend to hear stressed syllables as occurring at regular intervals, i.e., perceptual isochrony (e.g., Lehiste, 1977). Moreover, speakers tend to use similar intonation patterns on accented syllables within an intonational phrase (Couper-Kuhlen, 1993; Crystal, 1969; Dainora, 2001; Pierrehumbert, 2000). Repeated use of the same intonation pattern or phrasal boundary type is particularly common in lists of items (Beckman & Ayers Elam, 1997; Schubiger, 1958), but the same intonation pattern can also occur in coordinate syntactic constructions of various types (Wagner, 2005). Thus, the literature suggests that repeated intonation patterns persist over short stretches of speech, suggesting that such stretches, when they occur, may have communicative value in generating prosodic expectancies for listeners.

Sensitivity to pitch or rhythmic regularities could potentially facilitate speech segmentation if these regularities align with or help predict positions of linguistic structural significance, e.g., word or phrase boundaries. Work in non-speech auditory perception suggests that patterning in pitch and timing has predictable effects on percepts. In particular, when individuals hear simple tone sequences, the frequency, duration and amplitude patterning of the tones conveys a sense of sequence organization and structure (Boltz, 1993; Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley & Jones, 2003; Povel & Essens, 1985; Thomassen, 1982). Patterns of frequency and duration cause some sequence elements to sound like they belong together (i.e., to sound grouped), and within a group some elements to sound accented. For example, in an isochronous sequence of tones of equal amplitude and duration whose frequency alternates between high (H) and low (L), e.g., HLHLHL, listeners tend to hear a repeating strong–weak binary grouping with either the high or low tone as accented and beginning the group, i.e., (HL)(HL)(HL) or (LH)(LH)(LH) (Woodrow, 1909; Woodrow, 1911). Critically, these grouping effects have been shown to persist later in the sequences even when there are no explicit grouping cues in those later elements (Boltz, 1993; Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley & Jones, 2003; Povel & Essens, 1985; Thomassen, 1982).

In several experiments based on the above findings, Dilley and McAuley (2008) found that distal (i.e., distant or nonlocal) contextual prosodic regularities could persist to influence subsequent word segmentation. The present paper follows on their work and attempts to identify how these distal prosodic context cues can be reconciled with other types of signal- and knowledge-based cues within the theoretical framework advanced in Mattys, White, et al. (2005).

Distal prosody: the Dilley and McAuley (2008) experiments

Dilley and McAuley (2008) constructed auditory sequences beginning with two trochaic words (e.g., *channel dizzy*) and ending with four syllables that could form words in more than one way (*foot-note-book-worm* > *foot-note bookworm, foot notebook worm*, etc.). Participants provided a free report of the final word; the proportion of disyllabic responses was the dependent variable. Three types of distal prosody were created in which only the F0, only the duration, or both F0 and duration of the initial five syllables were manipulated (*channel dizzy foot*). Within each condition, the acoustic characteristics of the final three 'proximal' syllables (*note-book-worm*) were held constant. (Here, as in the original paper, the term 'proximal' refers to speech syllables which comprise or are adjacent to either of the possible final lexical items – e.g., *worm* vs. *bookworm* – while 'distal' refers to speech syllables preceding the proximal material.) The prosody of the initial five syllables was manipulated to create a prosodic context conducive to perception of either a disyllabic final word (Disyllabic context, e.g., *bookworm*) or a monosyllabic final word (Monosyllabic context, e.g., *worm*), as follows. In the F0 condition, the LHLHL pattern for the Disyllabic context was expected to yield a (LH)(LH)(L... organization and a H)(LH) grouping for the final three syllables (and hence, a disyllabic final word report), while the HLHLHL pattern for the Monosyllabic context was expected to yield a (HL)(HL)(HL) organization and thus a (HL)(H...) grouping for the final three syllables (and hence, a monosyllabic final word report). In the Duration condition, the entire sequence had monotone F0. Shortening the 5th syllable in the Disyllabic context made all syllables sound approximately isochronous perceptually, which was expected to yield a (SW)(SW)(S... organization and thus a W)(SW) grouping for the final syllables (and hence, a disyllabic final word report). In contrast, lengthening the 5th syllable in the Monosyllabic context was intended to induce the sense of a 'silent beat' on the second half of that syllable, which was expected to yield a percept of the fifth syllable as its own (SW) trochaic group, yielding a (SW)(SW)(SW) organization of context syllables and a (SW)(S...) grouping for the final syllables (and hence, a monosyllabic final word report).[1] Finally, in the F0 + Duration condition, the acoustic manipulations of the other two conditions were combined in a complementary manner, which was predicted to lead to strengthening of grouping percepts and a larger difference in rates of disyllabic word reports between the two prosody contexts relative to the other prosody conditions.

The following results were obtained. First, manipulating either distal F0 alone or distal duration alone affected word segmentation. Moreover, combining distal F0 and duration yielded the largest segmentation effects. In addition, the effects originated from the distal context as a whole and not simply from the 5th syllable, as shown by repeating the experiment but truncating the first four syllables of

[1] Here, S indicates a strong position in a metrical grid (grid height 2 or above), whereas W indicates a weak position in a metrical grid (grid height 1) (Hayes, 1995).

experimental items. Finally, confirmatory evidence of a distal prosody effect was found when a surprise recognition test was used instead of a free word report task. This latter finding suggests that the effect was not simply due to late-occurring meta-linguistic strategies.

Purpose of the present research

The purposes of the present experiments are to more specifically establish the mechanism behind the distal effect of prosody on segmentation as well as compare its strength relative to that of other cues. In particular, we propose that distal prosody may aid in segmentation by helping listeners to more reliably identify stressed syllables in the signal. Prior research has indicated that stressed syllables occur predominantly in word-onset positions in English (Cutler & Carter, 1987) and that listeners tend to perceive word boundaries before such syllables (Cutler & Butterfield, 1992), suggesting the utility for word segmentation of detecting stressed syllables. However, proximal prosodic cues to stress are quite variable and not always perceptually salient (e.g., Fear, Cutler, & Butterfield, 1995; Lehiste, 1970; Mattys, 2000). Distal prosodic cues could potentially aid through helping reliably identify stressed syllables even when proximal cues to stress are unclear (see also Pitt & Samuel 1990).

Lexical stress was assigned a low ranking in the segmentation hierarchy of Mattys, White, et al. (2005). However, Mattys et al. only examined *proximal* prosodic cues to lexical stress, and did not consider potential distal prosodic cues to segmentation. It could be that, compared to other cues, distal prosody does not aid much with segmentation, and thus would be low-ranked, just like proximal prosodic stress cues. On the other hand, if distal prosody reliably reinforces or disambiguates proximal prosody, it could be high-ranked and given significant 'weight' in determining segmentation, just like syntactic and semantic cues (Mattys et al., 2007).

In the present studies, we attempted to gain an initial understanding of how distal prosody might interact with other types of segmentation cues, in order to begin to integrate it with other cues in the theoretical framework developed by Mattys, White, et al. (2005). In the following experiments, we first explored the mechanism behind distal prosody by replicating and extending Dilley and McAuley's (2008) results. This was done by determining whether the distal prosodic effect generalizes to lexical sequences that are not based on compound words (Experiment 1a), assessing whether the effect is truly prosodic by using low-pass filtering (Experiment 1b), and testing the on-line nature of the effect with a cross-modal priming paradigm (Experiment 1c). Then, two experiments explored the strength of distal prosody relative to two other segmentation cues: semantic context (Experiments 2a and 2b) and proximal prosody (Experiments 3a and 3b), which were ranked high and low, respectively, in the segmentation hierarchy proposed by Mattys, White, et al. (2005). Distal prosody shares attributes with each type of cue and is therefore plausibly ranked equally well at the top or the bottom of the segmentation hierarchy. Indeed, like semantic context and unlike proximal prosody, distal prosody operates over relatively long domains, which would logically confine it to the upper tiers of the hierarchy. On the other hand, like proximal prosody, distal prosody is realized on the basis of suprasegmental characteristics, e.g., duration and F0. Proximal prosodic cues associated with word stress have relatively small effects on speech segmentation (Mattys, White, et al., 2005), so that distal prosodic effects, too, might be small. However, proximal prosodic cues associated with phrase boundaries have been noted to have relatively strong effects on segmentation (e.g., Christophe et al., 2004). In Experiment 2a, we established a semantic context manipulation and tested it in isolation and, in Experiment 2b, we pitted the semantic context manipulation against the distal prosody manipulation. Next, in Experiment 3a, we developed a relatively strong proximal prosody manipulation based on a combination of both pitch accent (cf. word stress) cues and phrasal boundary cues, following previous work. In Experiment 3b, we pitted these proximal prosodic cues against the distal prosodic cues.

Experiment 1a

This experiment was a replication of Dilley and McAuley's (2008) Experiment 1 for the condition in which distal prosody was strongest, i.e., where distal prosody was instantiated by both F0 and duration manipulations. The only difference was that the to-be-segmented portions of the utterances in the present experiment were comprised of non-compound words with simpler morphological structure. We made this change because the manner in which compound words are processed might not be reflective of general lexical processing. For example, it is possible that compound words are recognized via their constituent morphemes, which implies that they could undergo a form of pre-segmentation process independent of other cues (e.g., Taft, 1994; Wurm, 2000). Thus, Experiment 1 helped establish the generalizability of the findings of Dilley and McAuley to additional types of lexical materials with simpler morphological structure. As in the original study, the prosody of the beginning of the utterances (i.e., distal prosody) was manipulated to induce the perception of a segmentation point before the last two syllables of the utterances (e.g., *turnip*, disyllabic) or before the last syllable (e.g., *nip*, monosyllabic). Importantly, the last three syllables of the utterances were identical in both conditions (e.g., /sɪstɝnɪp/). Participants were asked to freely report the last word of the utterance and the proportion of disyllabic vs. monosyllabic responses was calculated.

The prosodic manipulation involved both F0 and duration cues—recall that Dilley and McAuley (2008) found that each cue independently was effective, but that the strongest effects were observed when both were combined. As in Dilley and McAuley, we focused on word lists, which the phonetics–phonology literature suggests is a particularly felicitous type of context for such variations (see e.g. Beckman & Ayers Elam 1997; Schubiger, 1958). A variety of real-world situations involve producing lists of lexical items, including reciting telephone numbers, spelling out

words, and reading aloud lists (grocery items for purchase, graduation honorees, etc.), suggesting the applicability of such constructions to many communicative contexts. Moreover, such contexts make it relatively straightforward to examine separately the contribution of lexical-semantic and proximal prosodic cues, which can be varied orthogonally, without the complicating factor of grammaticality.

Method

Participants

Twenty individuals from Bowling Green State University participated in the experiment in return for course credit or a nominal sum. All participants were at least 18 years of age, had self-reported normal hearing, and were native speakers of American English. Characteristics of participants were identical in all subsequent experiments.

Materials

Thirty eight-syllable experimental sequences were constructed (see Appendix A). Each experimental sequence consisted of two disyllabic words with initial primary stress, e.g., *magnet guilty*, followed by a four-syllable string that could be organized into words in more than one way, e.g., /krai sɪs tɝ nɪp/, which can be organized as *crisis turnip* or *cry sister nip* (see pronunciation rules below).[2] Each possible disyllabic word constructed from the final four syllables had initial primary stress. The majority of final disyllabic possible words were monomorphemic. For eight of 30, the two syllables of the disyllabic items were formed from different morphemes; in all of these cases, one of the morphemes was bound and the other was free, and in no case was a possible final disyllabic word also a compound word (cf. Dilley & McAuley, 2008). Moreover, 90 filler sequences were created, ranging in length from 6 to 10 syllables (3–6 words). Each filler sequence consisted of a mixture of monosyllabic and disyllabic words in varying positions within the string, all of which had unambiguous lexical structure. These sequences were intended to disguise the lexical ambiguity present in the experimental sequences. Disyllabic words in filler sequences always had initial primary stress; half of filler sequences ended in a monosyllabic word and half in a disyllabic word.

Experimental and filler sequences were read as connected speech by the first author, a native speaker of the General American English dialect from the Midwest US. Sequences were spoken with monotone F0; the final four syllables of experimental sequences were spoken as two disyllabic words. Multiple recordings were made for each sequence. Recordings were made in a sound-attenuated room directly to PC hard disk at a 16 kHz sampling rate with 16-bit quantization in Praat software (Boersma & Weenink, 2002) using a Shure SM58 microphone. One token of each sequence was then selected such that the final four syllables of the selected experimental sequences were judged to have segmental or allophonic pronunciations that were ambiguous between the two alternative lexical parses. This included ensuring that the vowels of the antepenultimate and final syllables (e.g., /sɪs/in *crisis* and /nɪp/in *turnip*), which are lexically unstressed, were produced with unreduced vowel quality. This was expected to facilitate these syllables' being reorganized into lexical items with a different pattern of lexical stress than the one spoken. This expectation is consistent with results of Fear et al. (1995) showing that syllables with unstressed unreduced vowel quality sound highly acceptable when put into lexically stressed syllable frames. Finally, syllables were spoken so as to sound approximately isochronous with respect to one another, with little phrase-final lengthening.

Next, two versions of each selected experimental sequence were created following the method outlined in Dilley and McAuley's (2008) Experiment 1 for the F0 + Duration condition, using a combination of waveform hand-editing and speech resynthesis with the pitch-synchronous overlap-and-add (PSOLA) algorithm (Moulines & Charpentier, 1990) implemented in Praat. First, the experimental sequence was spliced into three portions at zero crossings. The first portion consisted of the initial four syllables of the sequence plus the initial consonant of the fifth syllable (e.g., *magnet guilty* /k/-). The second portion consisted of the onset of the sonorant segment of the fifth syllable up to, but not including, the vowel onset of the sixth syllable (e.g., /rais/). The third portion consisted of the remaining material from the vowel onset of the sixth syllable through the end of the eighth syllable (e.g.,/ IstɝnIp/). The third portion was then resynthesized to have a HLH pitch (one tone per syllable); "H" and "L" targets corresponded to short stretches of monotone F0 of 165–175 and 235–245 Hz, respectively, aligned with the end of each corresponding syllable,[3,4] where each successive pair of tones was always separated by a linear rising or falling F0 interpolation. The result was then set aside for later concatenation with one of two types of distal prosodic context, as follows.

The choice of prosodic patterns for monosyllabic and disyllabic distal contexts was motivated by manipulations from work in non-speech auditory perception, as well as descriptions of prosodic patterns in spoken language (see

[2] Possible disyllabic final words in the present experimental materials always had primary stress on their initial syllables (e.g., turnip). This was done intentionally; it was predicted that such a lexical stress pattern would continue the "strong-weak" alternation set up by distal prosodic manipulations, thereby carrying over to proximal material in a way that would strengthen the sense of rhythm.

[3] Consistent with the choice of a fairly narrow range of F0 values for H and L tones, production experiments have revealed that speakers frequently demonstrate substantial regularity in F0 level at particular points in utterances. For instance, the low (L) endpoint of a declarative F0 fall varies little for any given speaker (Liberman & Pierrehumbert, 1984; Maeda, 1976), even in spontaneous speech (Anderson & Cooper, 1986; Menn & Boyce, 1982). Similarly, the F0 levels of successive high (H) tones can exhibit remarkable consistency in absolute F0 (Ashby, 1978; Liberman & Pierrehumbert, 1984). Crystal (1971) commented that (p. 26): "most speakers... produce most onset syllables within a narrow band of frequencies, which can be considered an absolute physical norm." For discussion of these issues, see also Crystal (1969: 235–252) and Ladd (2008: 62–72).

[4] For each non-final target (H or L), if the initial consonant segment of the following syllable was voiced and non-sonorant, the right edge of the short monotone region corresponding to that target was made to extend through that consonant.

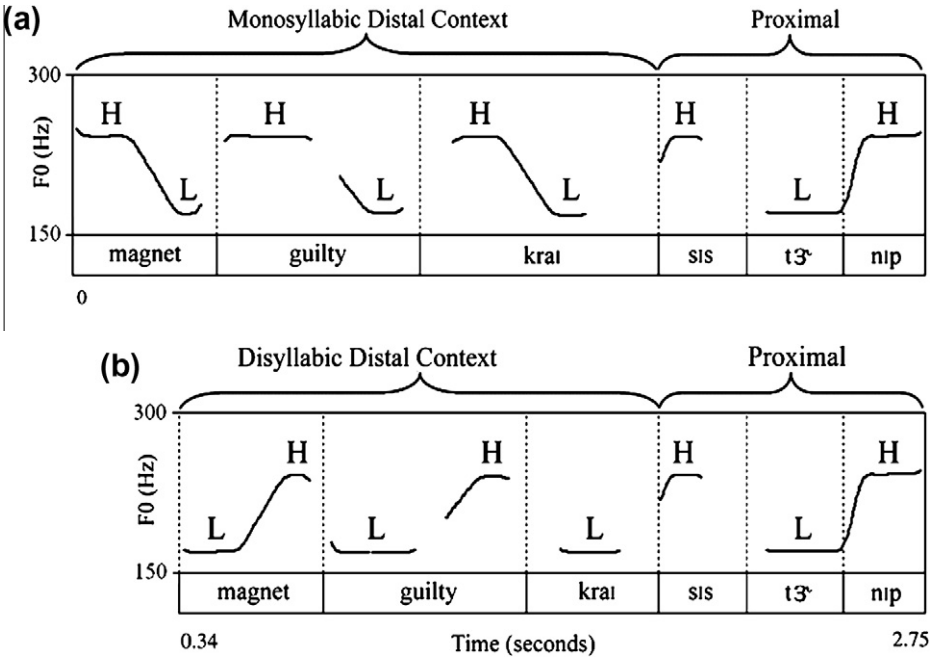

Fig. 1. Experiment 1a. (a) Example of an experimental sequence with Distal Prosody in the monosyllabic condition. (b) Example of an experimental sequence with Distal Prosody in the disyllabic condition.

Dilley & McAuley, 2008, for details). To create the monosyllabic distal context condition, the first five syllables (e.g., *magnet guilty* /kraɪs/) received a H_1–L_2–H_3–L_4–HL_5-pattern with one F0 target for each of the first four syllables, and a fall from H to L on the fifth syllable (subscripts indicate syllable numbers, while hyphens indicate syllable boundaries). For example, *mag-*, *-net*, *guil-*, and *-ty* were paired with H, L, H, and L, respectively, while *cry* was paired with a HL fall. Next, the second portion (vowel onset of the fifth syllable to vowel onset of the sixth syllable) was lengthened by a factor of 1.8 using PSOLA resynthesis in Praat and resynthesized to have a falling (HL) F0 pattern.[5] Any irregular pitch periods resulting from this manipulation were spliced off. Finally, the first, second, and third portions were concatenated in order to create the final monosyllabic experimental sequence. An example of a monosyllabic sequence is shown in Fig. 1a. The F0 manipulation of the first five syllables was predicted to yield a (H_1–L_2-)(H_3–L_4-)(HL_5-)(H_6–L_7-)(H_8…) grouping of the eight-syllable experimental sequence, creating perception of a larger prosodic phrase boundary before syllable 8 than 7, and hence, the perception of a juncture before the monosyllabic final word (e.g., *nip*). Moreover, based on findings from Dilley and McAuley (2008) that a combination of distal F0 and duration cues yielded the strongest effects on word segmentation, lengthening the second portion (cf. the fifth syllable) in the Monosyllabic context was expected to strengthen perception of the grouping structure induced by the F0 manipulation. In particular, the lengthening was expected to induce the sense of a 'silent beat' on the second half of the lengthened syllable, causing listeners to hear the fifth syllable as its own trochaic group (Sw_5-). As a result of the lengthening manipulation, the inter-beat-interval (IBI) between syllables 5 and 6 was approximately twice the IBI of all other pairs of successive syllables. This was expected to induce a grouping of experimental sequences as (S_1–W_2-)(S_3–W_4-)(Sw_5-)(S_6–W_7-)(S_8…). The shifted perceptual grouping of syllables introduced by the missing 'beat' was expected to move the location of the stronger prosodic boundary to before S_8 and thus cause listeners to report a monosyllabic final word (e.g., *nip*).

To create the disyllabic distal context condition, the first five (distal) syllables of each experimental sequence received a L_1–H_2–L_3–H_4–L_5-pattern, with one F0 target, H or L, per syllable. For example, *mag-*, *-net*, *guil-*, and *-ty* were paired with L, H, L and H tones, respectively. Next, the second portion was then slightly shortened, using a time compression factor of 0.9 through PSOLA resynthesis in Praat. This portion was then resynthesized to have a low (L) F0 pattern; any irregular pitch periods resulting from this manipulation were spliced off. Finally, the first, second, and third portions were concatenated in order to create each disyllabic experimental sequence. An example of a disyllabic sequence is shown in Fig. 1b. The F0

[5] The use of linear time-expansion via Praat resynthesis in our stimuli is unlikely to have led to issues with naturalness or intelligibility. The resultant speech sounded very natural, consistent with findings that time-domain PSOLA algorithm used in Praat affords very high-quality, intelligible speech relative to competing algorithms (Dutoit, 1994). Moreover, previous research has shown that linear time-expansion of speech by up to a factor of two yields speech which is comparably intelligible to unexpanded speech (Korabic, Freeman, & Church, 1978). Similarly, Gordon-Salant, Fitzgibbons, and Friedman (2007) showed listeners demonstrated excellent performance in perceiving time-expanded speech, regardless of which of several methods of nonlinear expansion were employed. Likewise, Janse, Nooteboom, and Quene (2003) found that speech which was time-compressed using nonlinear scaling modeled after natural production patterns was less intelligible than speech time-compressed using linear scaling.

manipulations were predicted to yield a (L_1–H_2-)(L_3–H_4-)(L_5–H_6-)(L_7–H_8) grouping of syllables, with a larger prosodic boundary before syllable 7 than 8, thereby inducing a perceptual juncture before the disyllabic final word (e.g., *turnip*). Based on results from Dilley and McAuley (2008), this grouping was expected to be strengthened by the duration manipulation on the second portion. As a result of the duration manipulation, the IBI between syllables 5 and 6 was approximately equal to the IBI of all other pairs of successive syllables. Assuming a continuation of the alternating pattern of stress created by the initial two S–W words, the disyllabic context was predicted to yield a (S_1–W_2-)(S_3–W_4-)(S_5–W_6-)(S_7–W_8) perceptual grouping with a stronger prosodic boundary before S_7 than before S_8. As a result, it was predicted that participants would tend to report a disyllabic final word (e.g., *turnip*).

Moreover, approximately half of filler sequences were resynthesized to have a rising (LH) pattern on each word. The remaining filler sequences were resynthesized to have a falling (HL) pattern on each word; for approximately half of these, the final word ended in a falling pattern while, for the remainder, the final word ended in a sustained H pitch. F0 values for H and L were in the range 220–260 Hz and 150–190 Hz, respectively. Finally, the amplitude of the experimental and filler sequences was normalized to 70 dB SPL and then upsampled to 22.05 kHz for compatibility with Eprime 1.1 experimental software (Psychology Software Tools, Inc., Pittsburgh, PA).

Design and procedure

There was a single within-subjects independent variable, Distal Prosodic Context (disyllabic, monosyllabic). Participants were asked to give a free report of the last word that they heard after listening to each sequence. Stimuli were presented over headphones using Eprime 1.1 running on a Dell Optiplex GX620 desktop computer. Responses were made by typing the word using a computer keyboard. Before starting the experiment, participants completed 16 practice trials which did not include any experimental sequences. Each participant then heard 30 experimental sequences and 90 filler sequences, pseudo-randomly ordered with the constraint that at least one filler sequence separated each pair of successive experimental sequences. Half of the 30 experimental sequences were presented in a disyllabic prosodic context, and the other half were presented in a monosyllabic prosodic context. The pairing of sequences with levels of Distal Prosodic Context was counterbalanced across sequences to create two lists. Two additional lists were then created by reversing the order of presentation of the sequences, for a total of four unique lists. Equal numbers of participants were randomly assigned to each list.

Results and discussion

All typed responses to experimental sequences were coded with respect to the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (fewer than 0.5% of trials). Next, a mixed-effect generalized model (Baayen, Davidson, & Bates, 2008) was applied to the ratio of disyllabic responses relative to the total of valid disyllabic and monosyllabic responses; in this analysis, both participants and items were considered simultaneously as random factors and Distal Prosodic Context (monosyllabic vs. disyllabic) was a fixed factor. Consistent with our expectation, disyllabic responses were far more frequent when the distal prosodic context induced segmentation of the disyllabic word (e.g., *turnip*) than segmentation of the monosyllabic word (e.g., *nip*), $F(1, 596) = 143.23$, $p < .001$ (Fig. 2), showing generalizability of the effect across participants and items. This effect is a clear replication of Dilley and McAuley's (2008) main finding, confirming the distal effect of prosody on speech segmentation. The data also show that this effect is robust enough to induce segmentation of non-compound words, therefore ruling out the possibility that the original effect was facilitated by morphologically-defined word boundaries.

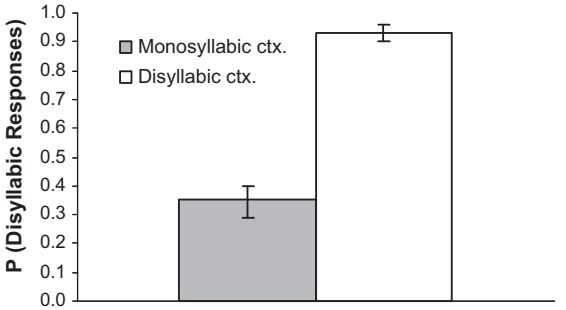

Fig. 2. Experiment 1a. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by the distal prosodic context (monosyllabic vs. disyllabic).

Experiment 1b

In an attempt to further confirm the prosodic origin of the segmentation effect in Experiment 1a, the sequences of this experiment were low-pass filtered to minimize access to subsegmental and segmental information. In doing so, any lexical or semantic bias that could have been present in the stimuli in Experiment 1a was eliminated or, at least, drastically reduced. Reduced intelligibility meant that the reporting task needed to be adjusted. In Experiment 1b, participants were asked to guess what the last word was. Of primary interest was the length of the guessed word (monosyllabic vs. disyllabic) as a function of the Distal Prosodic Context (monosyllabic-inducing vs. disyllabic-inducing).

Method

Participants, design, and materials

There were twenty participants in the experiment. The sequences for this experiment were low-pass filtered versions of the experimental and filler sequences from Experiment 1a. The Hann band filter in Praat was used for filtering; the pass band ranged from 0 to 800 Hz, and the width of the region between the pass band and stop band was 100 Hz. The design of the experiment was identical to that of Experiment 1a.

Procedure

Participants were asked to guess the last word that they heard after listening to each sequence. They were informed that the sequences consisted of only one- or two-syllable words. They produced their responses using a setup and procedure identical to that of Experiment 1a.

Results and discussion

All responses to experimental sequences were coded for the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (about 1% of trials). As expected from the low-pass manipulation, a vast majority of responses (88%) were segmentally incorrect, confirming that the manipulation achieved the desired reduction of segmental and lexical information (Fig. 3). The remaining 12% of responses matched the actual signal (e.g., reporting "nip" or "turnip" when hearing low-pass filtered/. . .kraisɪstɜ·nɪp/). A mixed-effect generalized model was applied to the ratio of disyllabic responses relative to the total of disyllabic and monosyllabic responses, with participants and items as random factors and Distal Prosodic Context (monosyllabic vs. disyllabic) and Segmental Match (match vs. mismatch) as fixed factors. There was no main effect of Segmental Match, $F(1, 583) = 2.59$, $p = .11$. However, as predicted, disyllabic responses were more frequent when the distal prosodic context induced segmentation of the disyllabic than monosyllabic word, $F(1, 583) = 51.60$, $p < .001$. This effect did not interact with whether the responses segmentally matched or mismatched the target, $F(1, 583) = 1.88$, $p = .15$. Separate analyses for the matched and mismatched responses revealed a significant Distal Prosodic Context effect in both cases: Match, $F(1, 62) = 11.73$, $p = .001$; Mismatch: $F(1, 583) = 48.26$, $p < .001$. Thus, the effect of distal prosody was clearly present even when the sequence was segmentally unintelligible. The results thus provide strong confirmatory evidence that Dilley and McAuley's (2008) effect of distal prosody on segmentation is indeed prosodic in nature.

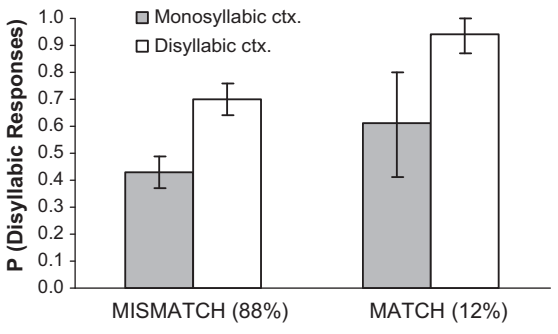

Fig. 3. Experiment 1b. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by the distal prosodic context (monosyllabic vs. disyllabic) and whether the participants' responses matched or mismatched the segmental content of the signal. The utterances were low-pass filtered versions of those of Experiment 1a.

Experiment 1c

In this experiment, we asked whether the effect of distal prosody can be observed on-line—as listeners hear the utterances—or whether it can only be accounted for at a strategic stage, when listeners have had time to activate meta-linguistic knowledge. Given that participants in the previous experiments were under no time pressure, a strategic locus for the effect is possible. However, the fact that Dilley and McAuley (2008) replicated their main pattern in an incidental recognition memory task suggests that distal prosody had an effect on how words were segmented and encoded in the absence of any incentive to listen to the utterances strategically (participants simply performed a phoneme-monitoring task during the presentation of the utterances). The earliness of the phenomenon still needs to be addressed, however.

To do so, we used a cross-modal identity priming task. Listeners heard one of the test utterances and then performed lexical decision on a visually presented monosyllabic or disyllabic letter string. On critical trials, the end of the utterance overlapped phonologically with the visual target. We predicted that lexical-decision latencies would be faster when the distal prosody of the utterance induced the perception of a word boundary that aligned with the beginning of the target word than when it did not.

Method

Participants, design, and materials

There were 48 participants in the experiment. The stimuli were those in Experiment 1a, consisting of 30 experimental sequences and 90 filler items. There were two within-subjects independent variables, the Distal Prosodic Context (disyllabic, monosyllabic) of the auditory prime and the number of syllables in the visual target (disyllabic, monosyllabic). Four lists were constructed from a single pseudorandom ordering of the 30 experimental sequences and 90 filler sequences; this ordering had the constraint that at least one filler sequence separated each pair of successive experimental sequences. One list was first constructed by pairing half of the 30 experimental sequences with the disyllabic distal prosodic context and the other half with the monosyllabic prosodic context; of these, half of the experimental sequences were paired with monosyllabic visual word "identity" targets (e.g., *nip*) and half with disyllabic visual word "identity" targets (e.g., *turnip*). The pairing of experimental sequences with levels of Distal Prosodic Context (monosyllabic, disyllabic) and of visual "identity" target length (monosyllabic, disyllabic) was then counterbalanced across items to create four lists. For filler items, 30 of the visually-presented items were real words in English (5 monosyllabic, 25 disyllabic) while the remaining 60 items were nonwords which followed English phonotactics (35 monosyllabic, 25 disyllabic); thus across the experiment, exactly 50% of trials involved real word targets. To ensure that form overlap was not a reliable cue to whether the target was a word or not, 20 of the nonwords had phonological overlap with the final word in the auditory filler primes (e.g., *hazard–hazap, taper–tapel, prior–priner*).

Equal numbers of participants were randomly assigned to each list. Participants were instructed to listen to each word list and then to judge whether the string of letters displayed on the computer screen formed a real word in English. Responding both quickly and accurately was emphasized. Stimuli were presented over studio-quality headphones using Eprime 1.1 running on a Dell Optiplex GX620 desktop computer. Responses were made by pressing a button on a response box; "Yes" responses were always made with the participant's dominant hand. Before starting the experiment, participants completed 16 practice trials which did not include any experimental sequences.

Results and discussion

Lexical-decision latencies were measured from the onset of visual target presentation. Incorrect responses and correct responses two standard deviations from the mean (computed separately for each participant) were discarded. Mean lexical-decision latencies and accuracy for the test conditions are plotted in Fig. 4.

The results show that, consistent with our main hypothesis, latencies were shorter when the visual target aligned with the lexical boundary induced by the distal prosody of the prime utterance than when it did not. A mixed-effect model, with participants and items as random factors and Distal Prosodic Context (monosyllabic vs. disyllabic) and Visual Target (monosyllabic vs. disyllabic) as fixed factors, showed a main effect of Distal Prosodic Context, $F(1, 1235) = 4.13$, $p < .05$, a main effect of Visual Target, $F(1, 1235) = 17.93$, $p < .05$, and, critically, a cross-over interaction between the two factors, $F(1, 1235) = 18.13$, $p < .001$. As expected from our previous experiments, latencies to the monosyllabic and disyllabic visual targets were affected in opposite directions by the preceding distal prosodic context: Monosyllabic visual targets, $F(1, 647) = 4.13$, $p < .05$; Disyllabic visual targets, $F(1, 588) = 32.37$, $p < .001$. A similar cross-over pattern was found in the accuracy data, $F(1, 1436) = 9.16$, $p < .005$, with monosyllabic visual targets responded to more accurately after a monosyllabic distal prosodic context, $F(1, 718) = 7.63$, $p < .01$, and disyllabic targets responded to more accurately after a disyllabic prosodic context, $F(1, 718) = 3.58$, $p = 06$. Of lesser importance for the present purpose was a main Visual Target effect $F(1, 1436) = 20.03$, $p < .001$, with lower performance for disyllabic than monosyllabic words, probably reflecting to the somewhat lower frequency of the former, $t(58) = 2.67$, $p < .01$ (Kucera & Francis, 1967).

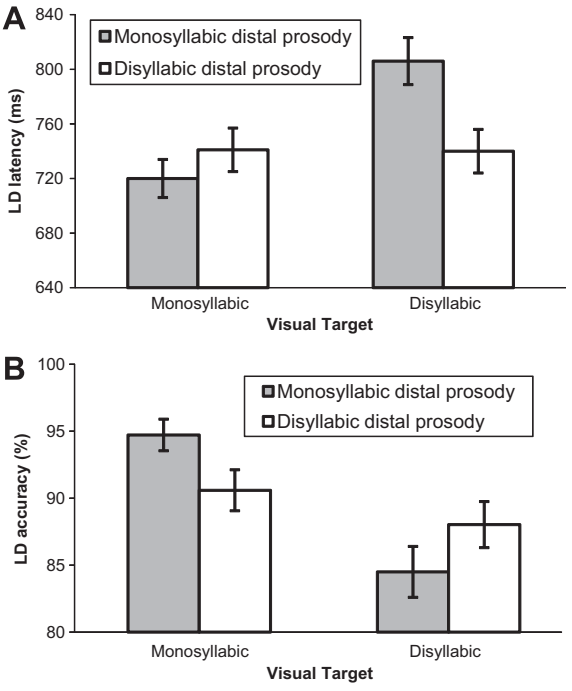

Fig. 4. Experiment 1c. Mean latencies (A) and accuracy (B) to the monosyllabic and disyllabic visual targets (and standard error bars) as a function of the type of segmentation induced by the distal prosodic context (monosyllabic vs. disyllabic) in the prime utterance.

This pattern confirms Dilley and McAuley's (2008) conclusion that the effect of distal prosody on word segmentation is not solely the consequence of late, strategic decisions. The fact that the critical interactive pattern emerged strongly within a second or so after the auditory onset of the primes indicates that distal prosody biases lexical activation at a fairly early stage of processing.

Experiment 2a

Experiments 1a–c showed that the segmentation of an utterance can be greatly affected by the prosody of nonadjacent portions of the signal. However, in those experiments, as well as Dilley and McAuley's (2008), the effect of distal prosody was tested in isolation, with potential contributions of other cues held constant. In particular, the words making up the distal context were the same in both distal prosodic conditions (e.g., *magnet guilty*), and they were chosen to be as semantically neutral as possible relative to the segmentation alternatives. However, Mattys, White, et al. (2005) have shown that the effect of semantic information and sentential context on speech segmentation can be substantial. For example, they found that the segmentation of a word in connected speech was faster if it was preceded by a semantically related than unrelated word, even when phonotactic and coarticulatory cues favored the latter (e.g., "gap" detected faster in "deepening gap" than in "pseudonym#gap," with # denoting a decoarticulation point and the underline denoting a phonotactically favorable diphone).

To tease out the relative weight of distal prosody and semantic context on the segmentation of our test sequences, we manipulated both variables orthogonally (Experiment 2b). First, however, we measured the effect of semantic context alone by neutralizing distal prosody (Experiment 2a). To do so, we replaced the first two words of each sequence from Experiment 1a (e.g., *magnet guilty*) with words that were semantically associated with either the final disyllabic word (e.g., *garden veggie crisis turnip*) or the final monosyllabic word (e.g., *puppy biting cry sister nip*). Context words were selected based on a pilot study – see Method section. Recall that in real-world situations involving produced lists of lexical items – telephone

numbers, spelling words verbally, reading aloud grocery items for purchase, etc. – the items are often or usually semantically related; thus, a semantic context manipulation based on meaningfully-related context words can clearly be tied to many naturalistic communicative situations. In this experiment and the following ones, we used our original free-report paradigm because of its unconstrained nature compared to the lexical-decision task of Experiment 1c, as well as its relatively high degree of external validity.

Method

Participants

There were eighteen participants in Experiment 2a.

Materials

A pilot study was first conducted in order to identify words that are close semantic associates of each monosyllabic or disyllabic final word in the experimental sequences. For the pilot, 20 participants were presented with a list of 60 written words corresponding to the final monosyllabic and disyllabic words. They were instructed to write down between two and four two-syllable words that were semantically related to each word on the list. An association strength was then calculated for each semantic associate, defined as the proportion of the 20 participants who wrote down that word. For each word, we identified the two semantic associates that: (a) had the highest association strength, (b) had a strong–weak stress pattern, and (c) were semantically related with each other.

Of the 30 experimental sequences used in Experiments 1a–c, a subset was selected in order to create sequences for Experiments 2a and 2b. Specifically, three sequences were excluded because the meanings of the disyllabic and monosyllabic final words (*cheese/munchies*, *fume/perfume*, and *plus/surplus*) were judged to be too closely related. One additional sequence ending in *pose/depots* was excluded because the pilot study suggested that the semantic associates for the monosyllabic possible final word (*picture* and *model* for *pose*) had association strengths that were almost four times those of semantic associates of the disyllabic possible final word (*railroad* and *station* for *depots*). For the remaining 26 sequences, average association strengths of the two associates favoring disyllabic final words ("disyllabic semantic contexts") and monosyllabic final words ("monosyllabic semantic contexts") were 0.25 and 0.27, respectively; this difference was not significant, $t(25) = 0.54$, $p = 0.59$. For these 26 sequences, the two initial words from Experiment 1a were replaced with the two newly-identified semantic associates, giving rise to a new set of eight-syllable strings serving as the basis for experimental sequences in Experiments 2a and 2b (Appendix B).

The experimental sequences were recorded by the same speaker as in Experiments 1a-c, using an identical recording setup. Sequences were produced with consistent, moderate speaking rate and monotone F0. For each sequence, syllables 1–4 plus the consonantal onset of the 5th syllable (the "semantic context", e.g., *garden veggie* /k/ related to disyllabic *turnip* or *puppy biting* /k/ related to monosyllabic *nip*) were spliced off and set aside for use in Experiment 2a, while the remaining portion of the recording was discarded. These semantic contexts were then combined with portions of existing materials from Experiment 1a. First, the portion of the speech from the vowel onset of the 5th syllable to the vowel onset of the 6th syllable (e.g., /rais/) was spliced out of the experimental sequences in Experiment 1a (the "original 5th syllable"). The duration of the original 5th syllable was then manipulated using PSOLA resynthesis in Praat. Specifically, the base duration of this portion was multiplied by a lengthening factor that resulted in a duration that equaled the average of the durations of the corresponding intervals in the monosyllabic and disyllabic conditions of Distal Prosodic Context in Experiment 1a; the average lengthening factor across sequences was 1.22. Any irregular pitch periods resulting from the duration manipulation were spliced off using waveform editing; the resulting portion is referred to as the "time-altered 5th syllable".

Because Experiment 2a aims to test the effect of semantic context on segmentation independent of distal prosody, PSOLA resynthesis was used to flatten the pitch of each semantic context and each time-altered 5th syllable to a monotone F0 of 202 Hz, which is approximately halfway between 170 Hz and 240 Hz (roughly the F0 values of L and H, respectively) on a logarithmic scale. The amplitude of each portion was subsequently normalized to 70 dB SPL. Next, the final part of each Experiment 1a sequence ranging from the onset of the vowel nucleus of the 6th syllable through the end of the stimulus (the "final portion", e.g., /Istɜ˞nIp/) was spliced out; recall that this portion had a HLH pitch pattern. All splices were taken at zero crossings. Finally, the flattened semantic context, the flattened time-altered 5th syllable, and the corresponding final portion were concatenated in sequence to create the experimental sequences for use in Experiment 2a.

Ninety new filler sequences were constructed consisting of three to six semantically related words generated from a thesaurus. Each filler sequence was 6–10 syllables in length. Filler sequences were recorded by the same speaker with the same setup as before. Half of filler sequences were resynthesized using Praat to have monotone F0 (202 Hz) across the initial 2 (for shorter sequences) to 7 syllables (for longer ones), followed by repeated rising patterns across subsequent words. Of these, approximately half ended in a rising pattern on the final word, and the other half ended in a sustained low pitch on the final word. The other half of filler sequences were resynthesized to have monotone F0 (202 Hz) across the initial 2 (for shorter sequences) to 8 syllables (for longer ones), followed by repeated falling patterns on each word. For approximately half of these, the final word ended in a falling pattern, and for the remainder, the final word ended in a sustained high pitch. F0 values for H and L were in the range 220–260 Hz and 150–190 Hz, respectively. Finally, the average amplitude of filler and experimental sequences was normalized to 70 dB. All sequences were then upsampled to 22.05 kHz for compatibility with Eprime 1.1 software.

Design and procedure

The manipulated variable was Semantic Context (monosyllabic, disyllabic). Two lists were constructed consisting

of 26 experimental sequences and 90 filler sequences in a pseudorandom order with the constraint that at least one filler sequence separated each pair of successive experimental sequences. For each list, half of the experimental sequences were paired with a disyllabic semantic context, and the other half were presented in a monosyllabic semantic context, with the pairing between sequences and context types counterbalanced across lists. An equal number of participants was randomly assigned to each list. Participants completed 6 practice trials before beginning the experiment. The experimental setup and procedure were otherwise identical to those of Experiment 1a.

Results and discussion

Typed responses to experimental sequences were coded for the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (approximately 1% of trials). A mixed-effect generalized model, with participants and items as random factors and Semantic Context (monosyllabic vs. disyllabic) as a fixed factor, revealed a clear effect of Semantic Context, $F(1, 461) = 34.01$, $p < .001$. Disyllabic responses were more numerous when the semantic context was consistent with the disyllabic word than when it was consistent with the monosyllabic word (Fig. 5), which confirms the contribution of semantic information to speech segmentation.

An analysis performed across Experiments 1a and 2a, with Type of Context (distal prosodic context [Experiment 1a] vs. semantic context [Experiment 2a]) and Implied Boundary (monosyllabic vs. disyllabic), showed an interaction between the two factors, $F(1, 1057) = 62.17$, $p < .001$, suggesting that the effect of Implied Boundary was larger in Experiment 1a than in Experiment 2a. There was no main effect of Type of Context, $F(1, 1057) = 1.86$, $p = .17$. Thus, the semantic manipulation was less effective than the prosodic manipulation. While this difference suggests that distal prosody is a particularly strong segmentation cue, it could also be due to the particular words we chose for the semantic contexts in Experiment 2a and/or the particular strength of prosodic cues we used in Experiment 1a. On the other hand, it is important to note that the larger effect size for distal prosody than semantic context was found even though the semantic manipulation was more likely to lend itself to strategic responding. Indeed, given the high proportion of semantically related trials, participants' responses could have been due not only to segmentation preferences but also to a propensity to respond in a way that made intuitive sense in the context of the experiment. This issue is more directly addressed in Experiment 2b.

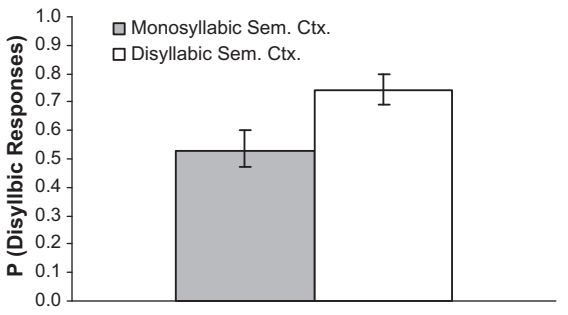

Fig. 5. Experiment 2a. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by the semantic context (monosyllabic vs. disyllabic).

Experiment 2b

In this experiment, distal prosody and semantic context were manipulated orthogonally. Of particular interest was whether these two variables would exert their effect in an additive fashion, with the original contribution of each segmentation cue unaffected by the presence of the other cue, or whether one cue would attenuate the effect of the other cue, as would be expected by an account that gives cues different weights when found in combination. An interaction between the cues would indicate a more complex relationship between the two cues, e.g., dominance of one cue over the other when the cues are in conflict and/or a synergetic effect when the cues converge (i.e., the combined effect of the two cues is greater than the sum of their individual effects).

Method

Participants

Twenty individuals participated in the experiment.

Materials

The monosyllabic and disyllabic levels of Semantic Context were, respectively, the initial four syllables of the monosyllabic and disyllabic sequences from Experiment 2a (e.g., monosyllabic *garden veggie*, disyllabic *puppy biting*). As will be described later, these semantic contexts were created by splicing out portions of speech from the onset of the first syllable to the offset of the fourth syllable from the corresponding Experiment 2a sequences.

To create a distal prosodic contrast on experimental sequences, the resynthesis manipulations used in Experiment 1a were applied to the initial five syllables of each experimental sequence. In particular, for the monosyllabic level of Distal Prosodic Context, we altered the F0 of the initial four syllables of each sequence, giving each a HLHL (i.e., falling) pattern, with one tone per syllable, following the method for the monosyllabic condition described in Experiment 1a. To the end of each initial four-syllable sequence, we concatenated a fragment from the monosyllabic condition of Experiment 1a materials consisting of the portion from the consonantal onset of the 5th syllable through the end of the initial consonant of the 6th syllable, e.g., /kraɪs/ of *cry s*(*ister#nip*); recall that this fragment had a HL F0 pattern and relatively long duration. Finally, each of these two fragments (e.g., *garden veggie* /kraɪs/ and *puppy biting* /kraɪs/, each with a HLHLHL F0 pattern) was concatenated with the final three syllables (e.g., /ɪstɚnɪp/ of *(s)ister#nip* from the vowel onset of the 6th syllable to the end of the 8th syllable of the corresponding sequence) of Experiment 1a materials; recall that these were acoustically identical in monosyllabic and disyllabic conditions in Experiment 1a and had a HLH pitch pattern.

To create the sequences for the disyllabic level of Distal Prosodic Context, we altered the F0 of the initial four syllables of each sequence, giving each a LHLH (i.e., falling) pattern, with one tone per syllable, following the method for the disyllabic condition described in Experiment 1a. To the end of each initial four-syllable sequence, we concatenated a fragment from the disyllabic condition of Experiment 1a materials consisting of the portion from the consonantal onset of the 5th syllable through the end of the initial consonant of the 6th syllable, e.g., /kraɪs/ of *cry s*(*ister#nip*); recall that this fragment had a L F0 pattern and relatively short duration. Finally, each of these two fragments (e.g., *garden veggie* /kraɪs/ and *puppy biting* /kraɪs/, each with a LHLHL F0 pattern) was concatenated with the final three syllables (e.g., /ɪstɔ·nɪp/ of *(s)ister#nip* from the vowel onset of the 6th syllable to the end of the 8th syllable of the corresponding sequence) of Experiment 1a materials; this was the same speech material as was appended to the end of each sequence in the monosyllabic level of Distal Prosodic Context in the present experiment.

As a result of these manipulations, the final three syllables of each experimental sequence (e.g., /ɪstɔ·nɪp/ of *(s)ister#nip*) were preceded by one of four types of contexts: (1) a monosyllabic Semantic Context (e.g., *puppy biting* /kraɪs/) paired with either (a) HLHLHL prosody and a lengthened 5th syllable (monosyllabic Distal Prosodic Context) or (b) LHLHL prosody and a shortened 5th syllable (disyllabic Distal Prosodic Context); or (2) a disyllabic Semantic Context (e.g., *garden veggie* /kraɪs/) paired with either (a) HLHLHL prosody and a lengthened 5th syllable (monosyllabic Distal Prosodic Context) or (b) LHLHL prosody and a shortened 5th syllable (disyllabic Distal Prosodic Context).

Design and procedure

We used a 2×2 within-subjects factorial design, in which two levels of Semantic Context (monosyllabic, disyllabic) were crossed with two levels of Distal Prosodic Context (monosyllabic, disyllabic). Four experimental lists were created. One list was constructed by pseudo-randomly ordering experimental and filler sequences, with the constraint that successive pairs of experimental sequences were separated by at least one filler item; approximately one-fourth of the experimental sequences were paired with each of the four experimental conditions in this list. The remaining three lists corresponded to the same ordering of experimental and filler sequences, but the pairing of experimental sequences with the four conditions was cycled across lists, so that each sequence occurred exactly once in each of the four conditions across the four lists. An equal number of participants was randomly assigned to each list. Participants completed six practice trials before beginning the experiment; the setup and procedure was otherwise identical to that of Experiment 1a.

Results and discussion

Responses to experimental sequences were coded for the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (less than 1% of trials). A mixed-effect model, with participants and items as random factors, and Distal Prosodic Context (monosyllabic vs. disyllabic) and Semantic Context (monosyllabic vs. disyllabic) as fixed factors, showed a Distal Prosodic Context effect, $F(1, 515) = 179.44$, $p < .001$, a Semantic Context effect, $F(1, 515) = 40.51$, $p < .001$, and no interaction, $F(1, 515) < 1$ (Fig. 6).

An analysis comparing the data of this experiment with those of Experiment 1a (distal prosodic context alone) indicated that the presence of a semantic context did not reduce the size of the Distal Prosodic Context effect, $F(1, 1113) < 1$. In fact, the effect of Distal Prosodic Context was numerically larger in this experiment (.69 compared to .58 on the 0-to-1 scale). Thus, even though the semantic manipulation was more likely to engage strategic responding, distal prosody came out as a highly robust and reliable resource for segmentation. Likewise, an analysis comparing the data of this experiment with those of Experiment 2a (Semantic Context alone) indicated that the presence of a distal prosodic context did not significantly reduce the size of the Semantic Context effect, $F(1, 978) = 1.18$, $p = .28$ (.17 vs. .21).

This experiment shows a clear additive effect between distal prosodic context and semantic context; that is, there was a main effect of both of these factors, with no evidence that the effect of distal prosody was attenuated by the presence of a semantic context. Thus, unlike word stress (cf. pitch accent placement), which is a type of "proximal" prosody that Mattys, White, et al. (2005) have shown to be outweighed by sentential information, distal prosody is robust and continues to fully operate in the presence of high-level information. This conclusion must be made with caution, however. The present manipulation of semantic context was restricted to lexical semantics (as opposed to sentential semantics) and was intrinsically constrained by the words selected from the pilot study. The effect of semantic context alone (Experiment 2a) was indeed smaller than that of distal prosody alone (Experiment 1a). Thus, it could be argued that the semantic manipulation was not strong enough to mitigate the effect of distal prosody. However, the fact that the magnitude of the semantic effect was itself not affected by conflicting distal prosody highlights the effectiveness and robustness of the semantic

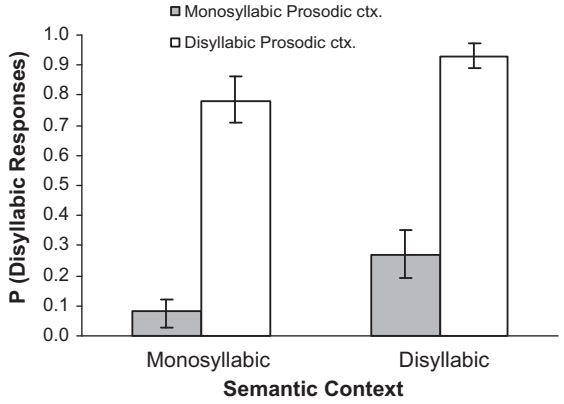

Fig. 6. Experiment 2b. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by the distal prosodic context (monosyllabic vs. disyllabic) and the semantic context (monosyllabic vs. disyllabic).

manipulation. Moreover, a very similar type of semantic manipulation in Mattys, White, et al. (2005) was shown to clearly outweigh other sub-lexical cues (phonotactics and coarticulation).

Experiment 3a

The above results raise the question of how distal prosody compares with proximal prosody when both cues are available in the input. According to prosodic theories, differences in F0, duration, and other suprasegmental cues arise from two kinds of prosodic phonological constructs: phrasal boundaries and pitch accents (Beckman & Pierrehumbert, 1986; Bolinger, 1958; Ladd, 2008; Nespor & Vogel, 1986; Pierrehumbert, 1980). Previous research suggests that proximal prosodic phrase boundaries influence word segmentation (Cho et al., 2007; Christophe et al., 2004; Millotte et al., 2008). Moreover, previous research also suggests that proximal placement of pitch accents may influence word segmentation. A focus of earlier work in speech segmentation has been the level of stress of a syllable, e.g., strong vs. reduced (Cutler & Norris, 1988) or primary vs. secondary (Mattys & Samuel, 2000), with primary stressed syllables tending to be perceived as word initial. It is known that differences in lexical stress also influence the placement of pitch accents (i.e., pitch excursions) on syllables, which would be expected to reinforce acoustical cues to lexical stress differences (Ladd, 2008; Shattuck-Hufnagel, 1995). Thus, studies suggesting differences in word segmentation arising from lexical-stress distinctions can be re-cast in terms of differences in locations of pitch accents. To create a strong proximal prosodic manipulation, we built on this prior work and thus manipulated distributions of both proximal phrasal boundary and pitch accent cues.

In Experiment 3a, we neutralized distal prosody and manipulated proximal prosody only. One proximal prosodic context was intended to induce segmentation of the final monosyllabic word (e.g., *nip*) and the other was intended to induce segmentation of the final disyllabic word (e.g., *turnip*). To create stimuli for Experiments 3a and 3b, we manipulated prosodic boundaries and pitch accents by altering the proximal F0 and duration on the final three syllables. We created two proximal prosodic environments: one with phrasal boundary and pitch accent placement that favored a final disyllabic word, and another with phrasal boundary and pitch accent placement that favored a final monosyllabic word. Manipulating both phrasal boundaries and pitch accents was expected to give proximal prosody the best chance of having an effect, compared with manipulating phrasal boundaries or pitch accents alone. Following the logic of Experiments 2a and 2b, we first measured the effect of proximal prosody in isolation (Experiment 3a), and then the orthogonal combination of proximal prosody and distal prosody (Experiment 3b).

Method

Participants

Twenty individuals participated in the experiment.

Materials

The experimental sequences in Experiment 3a were modified versions of the sequences in Experiment 1a (Appendix A). To create the two proximal prosodic contexts, the final three syllables (from the vowel onset of the 6th syllable through the end of the 8th syllable) were first spliced from the Experiment 1a sequences. The choice of acoustic manipulations to these syllables was motivated by assumptions within prosodic theory about the relationship between acoustic–phonetic attributes and major phrasal boundaries and pitch accents. Prosodic phrase boundaries come in a variety of sizes (Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986; see Shattuck-Hufnagel & Turk, 1996, for a review). For English and other languages, two of the largest prosodic constituents are the full intonational phrase (FIP) and the intermediate intonational phrase (IIP).[6] Boundaries of these constituent types are associated with distinct patterns of duration and F0 (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980). With respect to duration, both FIPs and IIPs are assumed to show local lengthening of segments (e.g., Fougeron & Keating, 1997). With respect to F0, both IIPs and FIPs are assumed to be paired with a tonal marker (i.e., a phrase accent or boundary tone, respectively), which is typically cued by an F0 change (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980). Pitch accents are assumed to be cued primarily by F0 attributes on the accented syllable (Ladd, 2008; Pierrehumbert, 1980). A number of pitch accent types have been proposed for English (e.g., L^*, H^*, $L+H^*$), corresponding to particular distinctive F0 attributes on the accented syllable and/or adjacent syllables (Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980). These widespread assumptions helped to motivate selection of acoustic parameters for proximal prosody manipulations.

To create the monosyllabic level of Proximal Prosodic Context, the F0 contour of the final three syllables was altered to be consistent with (1) pitch accents ('*') on the 6th and 8th syllables and (2) major prosodic phrase boundaries (']') at the edges of the 7th and 8th syllables. This was expected to yield a distribution of prosodic cues consistent with a monosyllabic final word, e.g. /(kraɪ) sɪs* tɚ] nɪp*]/. The prosodic manipulations were as follows: First, the 6th syllable was paired with a H^* pitch accent, corresponding to a high, level F0 of 220–260 Hz across the syllable's rhyme. Next, the end of the rhyme of the 7th syllable was paired with a L–L% phrase-accent/boundary tone sequence; this corresponded to a linear decrease in F0 ending in a low, level value of 120–130 Hz. The 8th syllable was then paired with a H^* pitch accent and following H–H% phrase accent/boundary tone sequence; this corresponded to a brief (~100 ms), level F0 at 140–150 Hz, followed by a steep linear interpolation to a final F0 of 370–380 Hz timed to occur with the end of voicing on the syllable. Finally, the

[6] Note that each large prosodic constituent is assumed to contain nested, embedded constituents of each successively smaller type (Selkirk, 1984). Thus, the boundary of each large prosodic constituent also corresponds to the boundary of a smaller constituent. However, the reverse is not true: Every small prosodic boundary does not correspond to the edge of a large prosodic constituent. For clarity, we only refer to the largest prosodic boundary occurring at each potential juncture point.

7th syllable was spliced out; the initial splice point was just before the consonantal onset if that onset was sonorant, and at the vowel onset if it was not. This portion was then time-expanded by a factor of 1.3 to simulate phrase-final lengthening. Any pitch period irregularities or transients arising from the time expansion were spliced off, and the resulting modal fragment was spliced back between the 6th and 8th syllables. The specific tonal pattern for the monosyllabic condition (H* L–L% H* H–H%) was selected to bear phonetic similarity to H and L tones of Distal prosodic context conditions to be used in Experiment 3b. An example of an experimental sequence in the monosyllabic condition is shown in Fig. 7a.

To create the disyllabic level of Proximal Prosodic Context, the F0 contour of the final three syllables was altered to be consistent with (1) a pitch accent on the 7th syllable and (2) major prosodic phrase boundaries at the edges of the 6th and 8th syllables. This was expected to yield a distribution of prosodic cues consistent with a monosyllabic final word, e.g. /(krai) sɪs] tɜ-* nɪp]/. The prosodic manipulations were as follows. First, the end of the 6th syllable was paired with a H- phrase accent, corresponding to a linear rise to a F0 value of 230–240 Hz. Next, the 7th syllable was paired with a L* pitch accent, corresponding to a low, level F0 of 120–130 Hz across the whole syllable. The 8th syllable was then paired with a H-H% phrase accent/boundary tone sequence; this corresponded to a F0 contour which rose steeply and linearly to 340–350 Hz across the first third to half of the syllable, followed by a shallower linear interpolation to a final F0 of 370–380 Hz across the last half to two-thirds of the syllable. The specific tonal pattern for the disyllabic condition (H– L* H–H%) was selected to bear phonetic similarity to H and L tones of Distal prosodic context conditions to be used in Experiment 3b. Finally, the 6th syllable was spliced out; the initial splice point was just before the consonantal onset if that onset was sonorant and at the vowel onset if it was not. This portion was then time-expanded by a factor of 1.3 to simulate phrase-final lengthening. Any pitch period irregularities or transients arising from the time expansion were spliced off, and the resulting modal fragment was spliced back before the 7th and 8th syllables. All splices were made at zero crossings. Stimuli resulting from these manipulations were checked closely for naturalness. An example of an experimental sequence in the monosyllabic condition is shown in Fig. 7b.

To neutralize distal prosodic contexts across both Proximal Prosodic Context conditions, a portion of speech from the onset of the first syllable through the consonantal onset of the 5th syllable was spliced out of the sequences of Experiment 1a. PSOLA resynthesis was used to flatten the pitch of this portion to 202 Hz (the midpoint on a log scale between the F0 values of L and H tones). This portion was then concatenated with each "flattened 5th syllable" (also 202 Hz) that had been created for Experiment 2a materials. The resultant portion is termed the "initial five syllables". The amplitude of the initial five syllables and corresponding final three syllables of each sequence was then normalized to 70 dB SPL and upsampled to 22.05 kHz for compatibility with Eprime 1.1. These were then concatenated to form the final experimental sequences. Finally, the filler sequences from Experiment 1a were modified for use in the present experiment by flattening the first three (for short sequences) to six (for the longest sequences) syllables of each sequence to 202 Hz, while between two and four syllables at the end of each filler sequence retained pitch variation.

Design and procedure

The within-subjects factor was Proximal Prosodic Context: monosyllabic (i.e., monosyllabic-inducing) or disyllabic (i.e., disyllabic-inducing). Four experimental lists

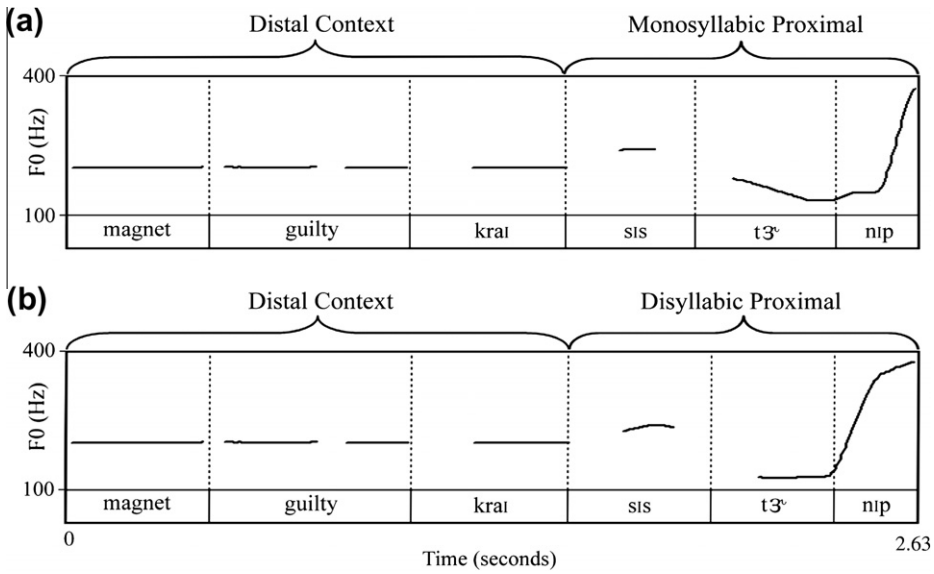

Fig. 7. Experiment 3a. (a) Example of an experimental sequence with Proximal Prosody in the monosyllabic condition. (b) Example of an experimental sequence with Proximal Prosody in the disyllabic condition.

were created from the 30 experimental sequences and 90 fillers. In one list, half of the experimental sequences received the monosyllabic Proximal Prosodic Context manipulation, and the other half the disyllabic Proximal Prosodic Context manipulation. A second list received the opposite assignment. The two remaining lists were created by reversing the order of the first two lists (first-to-last > last-to-first), for a total of four unique lists. Experimental and filler sequences within each list were pseudo-randomly ordered, with the constraint that there were no more than two consecutive experimental sequences. An equal number of participants was randomly assigned to each list. Participants completed six practice trials before beginning the experiment. The experimental setup and procedure were otherwise identical to those of Experiment 1a.

Results and discussion

All responses to experimental sequences were coded with respect to the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (1% of trials). A mixed-effect generalized model, with participants and items as random factors and Proximal Prosodic Context (monosyllabic vs. disyllabic) as a fixed factor, showed an effect of Proximal Prosodic Context, $F(1, 592) = 85.00$, $p < .001$. Thus, disyllabic responses were more numerous when proximal prosody favored segmentation of disyllabic than monosyllabic words (Fig. 8). This strong effect of proximal prosody on segmentation is an interesting departure from the literature showing that lexical stress is effective in segmenting speech in noise, but much less so with intact speech (Mattys, 2004; Mattys, White, et al., 2005). The kind of proximal prosody we used here is therefore clearly more impactful than lexical stress, even though they both rest on similar suprasegmental characteristics. The implications for models of speech segmentation will be discussed in the General Discussion.

In order to estimate the magnitude of the effect of proximal prosody relative to that of distal prosody, the data from this experiment and from Experiment 1a were entered into a single analysis, with Type of Prosody (Proximal Prosodic Context [Experiment 3a] vs. Distal Prosodic Context [Experiment 1a]) and Implied Boundary (monosyllabic vs. disyllabic) as fixed factors. A significant interaction between Type of Prosody and Implied Boundary, $F(1, 1188) = 41.97$, $p < .001$, revealed that the effect of proximal prosody was smaller than that of distal prosody. However, a similar analysis comparing the effects of Proximal Prosodic Context (Experiment 3a) and Semantic Context (Experiment 2a) showed that the effect of proximal prosody was larger than that of semantic context, $F(1, 1053) = 9.24$, $p < .005$. Again, these differences might reflect in part the specific manipulations used in each experiment. For example, using different proximal prosodic calibrations might have led to more or less effective proximal segmentation. However, the differences suggest that, in our best attempt to maximize the contrasts under study, distal prosody came out as a stronger cue than proximal prosody, and proximal prosody as a stronger cue that semantic context.

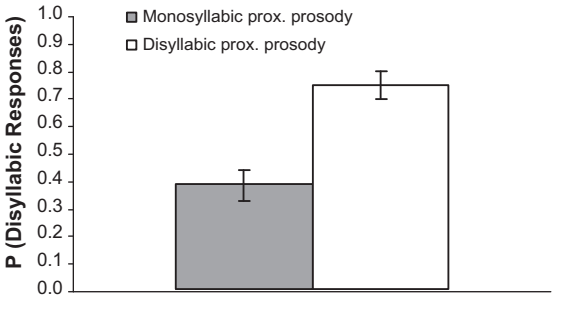

Fig. 8. Experiment 3a. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by proximal prosody (monosyllabic vs. disyllabic).

Experiment 3b

In this experiment, we manipulated proximal and distal prosody orthogonally. As before, of interest is whether the two types of prosody have a simple additive effect or whether patterns of dominance can be found.

Method

Participants

Twenty individuals participated in the experiment.

Materials

The two levels of Distal Prosodic Context were created by taking the experimental sequences from Experiment 1a and keeping only the portion stretching from the onset of the first syllable through the end of the consonantal onset of the 6th syllable. The two levels of Proximal Prosodic Context were created by taking the experimental sequences of Experiment 3a and keeping only the portion stretching from the end of the consonantal onset of the 6th syllable through the end of the 8th syllable. The amplitude of each portion was then normalized to 70 dB SPL. For each sentence of experimental sequences, the two levels of Distal Prosodic Context were then orthogonally concatenated with the two levels of Proximal Prosodic Context, giving rise to four versions of each experimental sequence. The filler sequences were those of Experiment 1a.

Design and procedure

The study used a 2×2 within-subjects factorial design, with Proximal Prosodic Context (monosyllabic, disyllabic) and Distal Prosodic Context (monosyllabic, disyllabic) as independent variables. Four experimental lists were created from the 30 experimental sequences and 90 fillers. A single list was first constructed by pseudo-randomly ordering experimental and filler sequences, with the constraint that there were no more than two experimental sequences in a row. Each of the experimental sequences was paired with one of the four experimental conditions in this list, with approximately equal proportions of experimental items in each condition. The remaining three lists corresponded to the same ordering of experimental and filler

sequences, but the pairing of experimental sequences with the four conditions was cycled across lists, so that each sequence occurred exactly once in each of the four conditions across the four lists. An equal number of participants was randomly assigned to each list. Before starting the experiment, participants completed six practice trials which did not include any experimental items. The instructions to participants and the procedure were otherwise identical to Experiment 1a.

Results and discussion

All responses to experimental sequences were coded with respect to the number of syllables they contained. Nonword responses and word responses with three or more syllables were discarded (about 1% of trials). A mixed-effect generalized model, with participants and items as random factors, and Proximal Prosodic Context (monosyllabic vs. disyllabic) and Distal Prosodic Context (monosyllabic vs. disyllabic) as fixed factors, showed a Proximal Prosodic Context effect, $F(1, 590) = 40.89$, $p < .001$, a Distal Prosodic Context effect, $F(1, 590) = 129.33$, $p < .001$, and an interaction, $F(1, 590) = 11.67$, $p < .001$. All four pairwise comparisons reached $p < .001$. This interaction clearly indicates that proximal and distal prosodic cues, when available in concert, impact each other's capacity to influence segmentation (Fig. 9).

Specifically, the effect of distal prosody was reduced when proximal prosody encouraged the segmentation of the disyllabic words. An analysis comparing the effect of Distal Prosodic Context in isolation (Experiment 1a) and in the presence of Proximal Prosodic Context (Experiment 3b) showed that the effect of distal prosody in the context of monosyllabic proximal prosody was similar to that of distal prosody in isolation, $F(1, 894) = 2.16$, $p = .10$, but that the effect of distal prosody in the context of disyllabic proximal prosody was smaller than that of distal prosody in isolation, $F(1, 888) = 21.81$, $p < .001$. Thus, proximal prosody attenuates the effect of distal prosody when proximal prosody signals an early word boundary (disyllabic Proximal Prosodic Context condition) but not when it signals a later word boundary (monosyllabic Proximal Prosodic Context condition). This interaction could be due to a number of factors, e.g., unexpected percepts resulting from the particular prosodic manipulations used in this experiment, or differential processing timecourses of distal and proximal prosody. A more parsimonious explanation at this stage, however, is that the strength of the disyllabic proximal prosody context was such that any additional contribution of distal prosody was masked by a ceiling effect. Although testing these possibilities will require followup work, a tentative conclusion from this experiment is that both distal and proximal prosodies have a substantial effect on speech segmentation and that neither seems to strongly dominate the other.

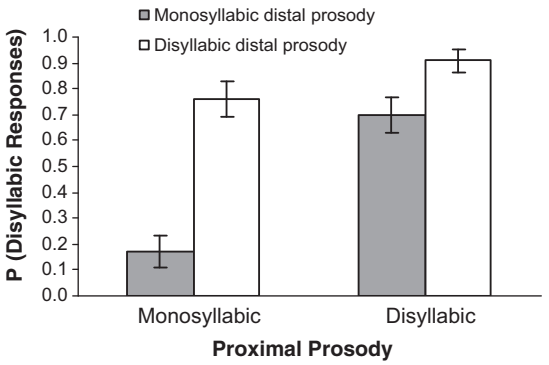

Fig. 9. Experiment 3b. Mean proportion of disyllabic responses with 95% confidence interval as a function of the type of segmentation induced by the distal prosodic context (monosyllabic vs. disyllabic) and proximal prosody (monosyllabic vs. disyllabic).

General discussion

The present paper investigated a factor recently identified by Dilley and McAuley (2008) as affecting word segmentation, namely distal prosody. Experiment 1a replicated the findings of Dilley and McAuley (2008) showing an effect of distal prosodic characteristics on word segmentation using items with more subtle end-embedding and more typical morphological structure than used in the earlier study. Experiment 1b confirmed that the distal prosody effect was, indeed, truly prosodic by using speech which had been low-pass filtered, thereby removing segmental, and hence, lexical-semantic information. Experiment 3c showed that the distal prosody effect was not solely the consequence of late, strategic, and/or meta-linguistic decisions, but that it biases lexical activation at a fairly early stage of processing.

Next, the effects of semantic context on segmentation were examined alone (Experiment 2a) and in combination with distal prosody (Experiment 2b). The results of these two experiments together confirmed the robustness of distal prosody even in the face of incongruent semantics. Moreover, the effects of proximal prosody were examined alone (Experiment 3a) and in combination with distal prosody (Experiment 3b). These experiments showed an interactive pattern between proximal and distal prosody, but with a clear unique contribution of distal prosody as well. Perhaps most striking was the magnitude of the impact of distal prosody on segmentation, as the report of disyllabic words in proximally un-manipulated stimuli varied by an average of 60% simply in response to differences in distal prosody.

These results constitute a significant departure from the mainstream literature on signal-based cues to speech segmentation, which has so far focused almost exclusively on proximal cues, especially with respect to prosody. For example, proximal prosodic cues affect the lexical activation of embedded words, e.g., *ham* in *hamster* (Cho et al., 2007; Christophe et al., 2004; Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003; Salverda et al., 2007; Shatzman & McQueen, 2006a) as well as perception of locations of word boundaries in lexically ambiguous segmental strings (Banel & Bacri, 1994; Nakatani & Schaffer, 1978; Shatzman & McQueen, 2006b). Such effects have been interpreted as resulting from prosodic

phrasal boundaries of different sizes as proposed by the theory of the prosodic hierarchy (e.g., Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986); see Shattuck-Hufnagel and Turk (1996) and Cutler, Dahan, and van Donselaar (1997) for reviews. Quantitative differences in proximal prosodic stress – e.g., whether a syllable has primary or secondary stress – affect word segmentation as well (Mattys, Jusczyk, Luce, & Morgan, 1999; Mattys & Samuel, 2000; Morgan, 1996; Vroomen & de Gelder, 1997; Vroomen, Tuomainen, & de Gelder, 1998). Considered more broadly, however, distal or nonlocal influences have previously been investigated in the areas of segmental perception (Holt, 2005; Kidd, 1989; Summerfield, 1981), segmental production (Hawkins & Nguyen, 2004) and implicit prominence judgments (Niebuhr, 2009).

Our results have obvious implications for models of spoken-word recognition and segmentation. Indeed, recall that one of our initial goals, besides replicating the basic distal prosodic effect using materials with more representative (i.e., simpler and non-compound) morphological structure, was to evaluate the strength of distal prosody relative to other segmentation cues, namely, semantic information and proximal prosody. In particular, based on Mattys, White, et al. (2005) hierarchical organization, we asked whether distal prosody would show the robustness of high-level cues, such as lexical-semantic information, or rank relatively low, as is the case for lexical stress. This question was motivated by the fact that, like high-level sources of information, distal prosody builds up over time, acting as a long-distance anticipatory cue for utterance structure but, like lexical stress, distal prosody is realized on the basis of suprasegmental characteristics, e.g., duration and F0. Based on the results of Experiment 2b, in which distal prosody and semantics were pitted against each other, we conclude that distal prosody, while is it is signal-derived, could not be filed as a low-weight cue, unlike lexical stress. Indeed, the presence of an incongruent semantic context had no mitigating effect on distal prosody. A question for future research is whether the robustness of distal prosody in the face of conflicting semantics will extend to sentence-level semantics. The latter differs from our semantic manipulation by its progressive build-up and its close relationship to syntactic structure. In contrast, our semantic manipulation, though distal, did not involve semantic-syntactic integration; in theory, our semantic effect could have happened entirely at the lexical level.

Our findings are important because they show that, contrary to Mattys, White, et al. (2005) proposal, segmentation cues need not be of a lexical-semantic nature to rank high. Instead, the evidence so far suggests that the size of the domain within which the cues operate might be the determining factor, with cues operating over large domains being particularly strong. This possibility is consistent with Christophe et al.'s (2004) claim that the effect of prosody on word segmentation occurs within the domain of phonological phrases. In their experiments, they found lexical activation of an overlapping candidate when the overlap was located inside a phonological phrase (e.g., activation of "chagrin" in "...chat grincheux][...", with the brackets denoting phrase boundaries), but not when the overlap straddled a phonological phrase boundary (e.g., "...chat][grimpait..."). A difference, however, is that, while the *origin* of the prosodic effect in Christophe et al.'s study was distal (the phrase), its *locus* was proximal (mostly restricted to the phonetic details at the critical juncture). According to the authors, the finding was consistent with one of two scenarios. In one scenario, segmental and prosodic details are considered simultaneously at the activation stage and they compete online for an optimal segmentation solution. In another scenario, lexical representations themselves contain juncture-specific prosodic details, and hence, no separate contribution from prosody is needed. The latter possibility must be rejected because our experiments (and Dilley & McAuley's 2008) showed clear effects of prosody even in the absence of any acoustic differences in to-be-segmented phrases. Thus, our results are consistent with independent contributions of segmental and prosodic details.

How the distal contribution of prosody can be implemented in models of spoken-word recognition is unclear, mainly because most models have so far failed to take into account factors outside the word domain. It now seems evident that a model capable of accounting for our results must have provision for distal context effects on lexical activation. One way of beginning to incorporate these effects into spoken-word recognition models is to view them within a traditional information-processing framework as the outcome of short-term memory and temporal selective attention processes. With respect to short-term memory, at issue is the extent to which the information held in auditory short-term memory can affect lexical activation, in particular: (1) the format of the memory store (how is prosody encoded?), (2) the span of the processing window (how distal can prosodic effects be?), and (3) the way in which the content of the memory store interacts with the activation of the lexical representations held in long-term memory. These questions, although already present in the literature on the time-course of lexical-semantic integration (Mattys, Pleydell-Pearce, Melhorn, & Whitecross, 2005; van Petten, Coulson, Rubin, Plante, & Parks, 1999) have not yet found a satisfactory answer. The issue of temporal attention has a similar status. Attentional effects on speech segmentation and lexical access have been reported before (e.g., Mattys, Brooks, & Cooke, 2009), with an emphasis on the role of salient acoustic cues (Astheimer & Sanders, 2009) and rhythm (Pitt & Samuel, 1990) in constraining lexical access. For instance, not unlike our own results, Pitt and Samuel (1990) showed that a repeating rhythmic pattern leads to a build-up of attention to stressed syllables later down a speech stream (the *attentional bounce hypothesis*). However, how attentional allocation actually modulates the activation levels of lexical representations has seldom been addressed. The few studies that have attempted to do so (e.g., Mirman, McClelland, Holt, & Magnuson, 2008; Norris, McQueen, & Cutler, 2000) have unfortunately limited their computational domain to the word level. Our results clearly show that models of spoken lexical access must extend their algorithms to the utterance level.

Another approach which might be pursued in incorporating distal context effects into models of spoken-word

recognition is to view these effects within the framework of dynamical systems as the outcome of entrainment via one or more endogenous oscillators or "clocks" (Barbosa, 2007; Cummins & Port, 1998; Jones, 1976; Large & Jones, 1999; McAuley, 1995; McAuley & Jones, 2003; Port, 2003). According to such approaches, endogenous internal oscillators are responsible for judgments about the timing of events in the environment and for coordinating motor actions in response to them. Oscillators, which attune to temporal periodicities and quasi-periodicities in auditory stimuli, have been proposed to play a role in both perception and production of speech (Byrd & Saltzman, 2003; Cummins & Port, 1998; Nam, Goldstein, & Saltzman, 2006; Port, 2003). Importantly, no explicit short-term memory component is needed to account for distal context effects, since periodic or quasi-periodic stimuli are assumed to affect entrainment dynamics of the oscillators directly, thereby influencing their subsequent behavior. Moreover, modulation of attention by periodic or quasi-periodic aspects of distal context has been accounted for by proposing that attention is focused on a window around "expected" moments in time coinciding with peaks in oscillator amplitude (Jones, 1976; Large & Jones, 1999). Evidence for endogenous oscillators comes from perception and production studies involving both speech- and non-speech related tasks (e.g., Cummins & Port, 1998; McAuley, 1995; McAuley & Jones, 2003). A challenging but fruitful avenue for future work will be to integrate distal context effects as modeled by oscillator accounts of speech perception and production into models of spoken-word recognition.

In this regard, on-line measures of lexical activation—beyond the "semi-on-line" method used in Experiment 1c—will surely prove useful in elucidating the time-course of distal context effects on spoken-word recognition and beginning to tease apart the predictions of information theoretic vs. entrainment perspectives about the temporal dynamics of such effects. For example, eye-tracking paradigms or on-line lexical-decision tasks will be useful in determining how early distal context effects become available to the perceptual system and how they are used over time in recognizing spoken words. Previous work using on-line paradigms suggests that information about spoken words is used as soon as it is available (e.g., Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008), suggesting that on-line spoken-word recognition might be influenced by distal context several syllables prior to the acoustic onset of a given spoken word.

Even if distal context is shown to influence not only word segmentation, as demonstrated here, but also on-line lexical activation, a legitimate question is the extent to which cues of the sort used in our experiments are available in everyday spoken language. Evidence for patterning in each of two kinds of acoustic–phonetic dimensions can be considered. With respect to pitch, evidence from linguistic descriptions suggests that repeating pitch patterns may commonly occur in a number of languages (Beckman & Pierrehumbert, 1986; Hayes & Lahiri, 1991; Ladd, 1986; Pierrehumbert, 2000). With respect to duration and timing, it is well-known that speech tends to sound perceptually isochronous, although it does not show measurable acoustic isochrony (Dilley, 1997; Lehiste, 1977; McAuley & Dilley, 2004); moreover, pitch pattern regularity and perceptual isochrony have been reported to commonly co-occur (Couper-Kuhlen, 1993). These observations suggest that speech sometimes contains perceptual cues to cyclic patterns of pitch and/or timing which listeners are sensitive to, as demonstrated by our work and that of others (e.g., Pitt & Samuel, 1990). However, the pervasiveness with which speech shows perceptual isochrony and repeating patterns of pitch is not known. Our findings demonstrating that such cues impact word segmentation, rather than just attention to speech (Cutler, 1976; Pitt & Samuel, 1990), provide motivation for further investigations about both the relative frequency with which repeating pitch patterns and/or perceptual isochrony occur in speech, as well as the communicative conditions under which such patterns arise.

To the extent that distal prosodic regularities are present in spoken language, why and how would the perceptual system use such information? We argue that the main advantage of distal prosody is that it allows listeners to anticipate the occurrence of stressed syllables. Increased attention to stressed syllable would be beneficial for two reasons: the more informative segmental structure of stressed syllables aids in lexical identification, and the distributional properties of these syllables (specifically, their tendency to be word-initial) makes them useful in word segmentation. First, compared to reduced syllables, stressed syllables *reduce more uncertainty* about the segmental compositions of spoken words and thus *provide more and better information* about lexical identity, as suggested by a number of observations and findings. Indeed, it is well-known that stressed syllables provide clearer acoustical cues to segmental identity than reduced syllables (e.g., Beckman, 1986; Lehiste, 1970), thus providing more reliable acoustical information about segmental content. This holds true not only for vowels, but also for consonants: stressed syllables tend to have consonants with phonetic realizations which more closely match these segments' canonical forms (de Jong, 1998; Shockey, 2003) and which are more resistant to deletion than reduced syllables (Bell et al., 2003; Johnson, 2004), thus eliminating more uncertainty about segmental content than reduced syllables. Additionally, stressed syllables also provide more information about lexical identity than reduced syllables due to the relative (un)predictability of segmental content in stressed syllables (Altmann & Carter, 1989; Huttenlocher, 1984; Piantadosi, Tily, & Gibson, 2009). For example, vowels in reduced syllables are more predictable, and hence less informative, about the segmental composition of words than vowels in stressed syllables (Altmann & Carter, 1989; Piantadosi et al., 2009). This appears to be true of consonants as well, with differential degrees of reduction in uncertainty for consonants in stressed syllables for some languages more than others (Piantadosi et al., 2009). Given limited attention and memory resources, it is advantageous for the perceptual system to use distal prosodic cues to predict the location of stressed than

reduced syllables, since the former allow for greater reductions in uncertainty about segmental composition, and hence the identities of spoken words.

Second, we propose that using distal prosodic regularities to attend to rhythmically stressed syllables is advantageous because it helps the perceptual system to more reliably identify *which* syllables are lexically stressed. Previous research has established the importance of lexical stress for early word segmentation (Jusczyk, Houston, & Newsome, 1999; Nazzi, Dilley, Jusczyk, Shattuck-Hufnagel, & Jusczyk, 2005) in addition to statistical cues (Thiessen & Saffran, 2003). However, acoustic cues to lexical stress are variable and often unreliable (Beckman, 1986; Fry, 1955; Sluijter & van Heuven, 1996), making the apparent reliance on lexical stress *per se* by early language learners rather puzzling. Moreover, so-called unstressed, unreduced syllables (such as the last syllable of *veto* and the first and last syllables of *piano*) are potentially confusable perceptually with lexically stressed syllables. In particular, unstressed unreduced syllables have a number of the properties of lexically stressed syllables, such as full vowel quality (Bolinger, 1981), and perception research has established that unstressed unreduced vowels sound natural and acceptable to listeners when they are cross-spliced into vocalic positions of lexically stressed syllables (Fear et al., 1995). Indeed, our six experiments exploited the ambiguity of unstressed unreduced syllables to sound acceptable as primary stressed syllables: we predicted correctly that distal prosody could affect the perceived location of lexical stress. We therefore propose a role for distal prosody in disambiguating which syllables are truly lexically stressed; that is, distal prosodic cues could potentially be more reliable indicators of lexical stress than proximal acoustic cues alone. Sensitivity to distal prosodic regularities, such as rhythmic stress and quasi-periodic pitch patterns, could be of particular use in early stages of language acquisition by more reliably guiding language learners to locations of stressed syllables in continuous speech than is possible from proximal acoustic cues to stress alone, enabling young learners to form "chunks" of speech material for segmentation and develop accurate knowledge of lexical stress. As higher-level knowledge of their language grows, reliance on distal prosody to identify which syllables are lexically stressed is expected to diminish, so that for adult speakers of a language, the benefit of sensitivity to distal prosody might be due to enhanced attention to information-laden (stressed) syllables, as discussed above. An interesting implication of our hypothesis that distal prosody helps to more reliably identify stressed syllables is that distal prosodic cues may be more useful in identifying stressed syllables in "syllable-timed" languages such as French and Spanish, in which unstressed syllables seldom show vowel reduction, than in English, in which unstressed syllables usually exhibit vowel reduction (Dauer, 1983; Roach, 1982).

Finally, as a note for future research, it will be interesting to find out the extent to which reliance on distal prosody for segmentation is subject to individual differences. Given that the perception of prosody has been shown to be related to musical abilities (e.g., Patel, Peretz, Tramo, & Labreque, 1998; Patel, Wong, Foxton, Lochy, & Peretz, 2008), one could anticipate that individuals with musical training would be more sensitive to our prosodic manipulation than individuals with less or no musical training at all. In all our experiments, participants were asked to indicate their level of musical training (number of years). To test if a link existed between distal prosody and musical training, we calculated the effect of distal prosody for the participants involved in experiments that used the same task and which included distal prosody as a factor (Experiments, 1a, 1b, 2b, 3b; $N = 80$ in total), and included their number of years of musical training as a covariate. The effect of distal prosody, $F(1, 2294) = 876.87$, $p < .001$, was significantly modulated by musical training, $F(1, 2294) = 5.31$, $p = .02$. To explore this link further, we split the 80 participants into three sub-groups of approximately equal size: No musical training ($N = 28$); One to five years of musical training (3.3 years on average, $N = 27$); More than five years of musical training (8.7 years on average, $N = 25$). The effect of distal prosody, measured as the difference between the ratio of disyllabic responses in the monosyllabic and disyllabic distal conditions, was .41 (i.e., .75 − .34), .55 (i.e., .87 − .32), and .52 (i.e., .90 − .38), respectively. Again, distal prosody and musical training were found to interact, $F(2, 2292) = 4.71$, $p < .01$, with the most noticeable difference between the no-training group and the more-than-five-years group, $F(1, 1524) = 9.71$, $p = .002$. Thus, individuals with musical training exhibited greater sensitivity to distal prosody than individuals with no musical training. Interestingly, the greater sensitivity to distal prosody in the trained musicians manifested itself as an increased proportion of disyllabic responses in the disyllable-inducing condition, $F(2, 1147) = 8.32$, $p < .001$, but not as a change in the monosyllable-inducing condition, $F(2, 1145) < 1$. Such findings would tend to support the idea that musical training can influence linguistic perception, stimulating the ongoing debate about the modularity of linguistic vs. musical processing (Fedorenko, Patel, Casasanto, Winawer, & Gibson, 2009; Patel et al., 2008; Peretz & Coltheart, 2003; Slevc, Rosenberg, & Patel, 2009; Zatorre & Gandour, 2007).

In sum, distal prosody was shown to be an extremely robust segmentation cue, indicating a new, powerful factor for consideration by models of word segmentation and lexical access. While speech is expected to vary in the extent to which rhythmically regular distal prosodic cues are present in the signal, the results obtained here suggest that when they are, they have robust effects on word segmentation. Such sensitivity to distal prosodic cues could serve both to reduce uncertainty about segmental composition of spoken words, as well as to more reliably guide listeners, particularly neonates, to the locations of lexically stressed syllables. Future work will be aimed at testing these hypotheses.

Appendix A. Stimuli for Experiments 1a, 1b, 1c, 2a, and 2b. Shown in parentheses are the parses of the final four syllables that are consistent with either a monosyllabic or a disyllabic final word for each experimental string.

1. banker helpful (tie murder *bee*/timer derby)
2. kettle heaven (Tim burrow bow/timber oboe)
3. pebble dollar (bar lever chew/barley virtue)
4. gossip oyster (pan treaty coy/pantry decoy)

5. plenty fluid (tray dirty crease/traitor decrease)
6. angry index (lay birdie fence/labor defense)
7. feather onion (bay beaker few/baby curfew)
8. chapter elbow (rue beaver gin/ruby virgin)
9. magic notice (gang sterling go/gangster lingo)
10. kitchen dealer (may beanie grow/maybe negro)
11. hero vacuum (sell early gull/cellar legal)
12. bullet junior (come feeding key/comfy dinky)
13. liquid perish (broad leasing king/broadly sinking)
14. lumpy danger (chair eager knee/cherry gurney)
15. lender dentist (hare umber lap/harem burlap)
16. plasma honey (pigs typo low/pigsty polo)
17. forest pepper (pee canter might/pecan termite)
18. blanket mounted (ham mercy nick/hammer scenic)
19. magnet guilty (cry sister nip/crisis turnip)
20. tourist robin (draw musty plea/drama steeply)
21. sandwich rosy (far gopher meant/Fargo ferment)
22. trouble wealthy (limb burner sing/limber nursing)
23. nicely equal (gray veto stir/gravy toaster)
24. nature lazy (faux meaty tour/foamy detour)
25. lady jacket (bran diesel tree/brandy sultry)
26. fever pencil (lie bully word/libel leeward)
27. husband lemon (fan seaman cheese/fancy munchies)
28. fortune decade (win deeper fume/windy perfume)
29. center northern (two cancer plus/toucan surplus)
30. mixture pleasure (class seedy pose/classy depots).

Appendix B. Semantic contexts used in Experiments 3a and 3b

Monosyllabic Semantic context (*monosyllabic final word*)/Disyllabic Semantic context (*disyllabic final word*)

1. honey stinger (*bee*)/horses racing (*derby*)
2. arrow ribbon (*bow*)/woodwind music (*oboe*)
3. eating dinner (*chew*)/moral values (*virtue*)
4. shyly sneaky (*coy*)/hunting faking (*decoy*)
5. paper folding (*crease*)/lower minus (*decrease*)
6. picket gated (*fence*)/football offense (*defense*)
7. little tiny (*few*)/midnight parents (*curfew*)
8. drinking liquor (*gin*)/Mary bible (*virgin*)
9. moving forward (*go*)/language talking (*lingo*)
10. bigger older (*grow*)/color Spanish (*negro*)
11. flying water (*gull*)/jury lawyer (*legal*)
12. locksmith doorknob (*key*)/tiny little (*dinky*)
13. ruler royal (*king*)/swimming drowning (*sinking*)
14. body bending (*knee*)/doctor injured (*gurney*)
15. sitting baby (*lap*)/fabric knapsack (*burlap*)
16. under little (*low*)/water playing (*polo*)
17. maybe hopeful (might)/chewing insect (*termite*)
18. shaving razor (*nick*)/pretty driving (*scenic*)
19. puppy biting (*nip*)/garden veggie (*turnip*)
20. guilty begging (*plea*)/mountain incline (*steeply*)
21. meaning purpose (*meant*)/liquor smelly (*ferment*)
22. music lyrics (*sing*)/baby doctor (*nursing*)
23. mixing cooking (*stir*)/oven bagel (*toaster*)
24. travel looking (*tour*)/driving shortcut (*detour*)
25. maple branches (*tree*)/sexy dancing (*sultry*)
26. writing language (*word*)/leaning forward (*leeward*).

References

Altmann, G. T. M., & Carter, D. (1989). Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? *Computer and Speech Language, 3*, 265–275.

Anderson, S., & Cooper, W. E. (1986). Fundamental frequency patterns during spontaneous picture description. *Journal of the Acoustical Society of America, 79*, 1172–1174.

Ashby, M. (1978). A study of two English nuclear tones. *Language and Speech, 21*, 326–336.

Astheimer, L. B., & Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology, 80*, 23–34.

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, 12*, 390–412.

Banel, M. H., & Bacri, N. (1994). On metrical patterns and lexical parsing in French. *Speech Communication, 15*, 115–126.

Barbosa, P. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication, 49*, 725–742.

Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

Beckman, M., & Ayers Elam, G. (1997). Guidelines for ToBI labeling, version 3. Ohio State University.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

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*(2), 1001–1024.

Boersma, P., & Weenink, D. (2002). Praat, a system for doing phonetics by computer (Version 4.0.26): Software and manual available online at <http://www.praat.org>.

Bolinger, D. (1958). A theory of pitch accent in English. *Word, 14*, 109–149.

Bolinger, D. (1981). *Two kinds of vowels, two kinds of rhythm*. Indiana University Linguistics Club.

Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics, 53*, 585–600.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*, 149–180.

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

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

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language, 51*, 523–547.

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

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction*. John Benjamins.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Crystal, D. (1971). Relative and absolute in intonation analysis. *Journal of the International Phonetic Association, 1*, 17–28.

Cummins, F., & Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics, 20*, 55–60.

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). The 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. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113–121.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507–534.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*, 292–314.

Dainora, A. (2001). *An empirically based probabilistic model of intonation in American English*. Ph.D. dissertation, University of Chicago.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51–62.

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.

de Jong, K. (1998). Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics, 26*, 283–310.

Dilley, L. C. (1997). Some factors influencing duration between syllables judged perceptually isochronous. *Journal of Acoustical Society of America, 102*, 3205–3206.

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

Dutoit, T. (1994). High quality text-to-speech synthesis: A comparison of four candidate algorithms. In *Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP'94)* (pp. 565–568). Adelaide, Australia.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America, 97*(3), 1893–1904.

Fedorenko, E., Patel, A. D., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory and Cognition, 37*, 1–9.

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

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America, 27*, 765–768.

Gordon-Salant, S., Fitzgibbons, P. J., & Friedman, S. A. (2007). Recognition of time-compressed and natural speech with selective temporal enhancements by young and elderly listeners. *Journal of Speech, Language and Hearing Research, 50*, 1181–1193.

Gout, A., Christophe, A., & Morgan, J. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language, 51*, 547–567.

Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics, 32*, 199–231.

Hayes, B. (1995). *Metrical stress theory*. Chicago: University of Chicago Press.

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory, 9*, 47–96.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*, 305–312.

Huttenlocher, D. (1984) *Acoustic–phonetic and lexical constraints in word recognition: Lexical access using partial information*. MA thesis, MIT.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58*, 541–573.

Janse, E., Nooteboom, S. G., & Quene, H. (2003). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication, 41*, 287–301.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29–54). Tokyo, Japan: The National International Institute for Japanese Language.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83*, 323–355.

Jones, M. R., & Boltz, M. G. (1989). Dynamic attending and responses to time. *Psychological Review, 96*, 459–491.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in the English-learning infants. *Cognitive Psychology, 39*, 159–207.

Kidd, G. R. (1989). Articulatory rate-context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 736–748.

Klatt, D. H. (1980). Speech perception: A model of acoustic–phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 243–288). Hillsdale, NJ: Erlbaum.

Korabic, E. W., Freeman, B., & Church, G. T. (1978). Intelligibility of time-expanded speech with normally-hearing and elderly subjects. *International Journal of Audiology, 17*, 159–164.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Ladd, D. R. (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook, 3*, 311–340.

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 119–159.

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

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5*, 253–263.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), *Language Sound Structure* (pp. 157–233). Cambridge, MA: MIT Press.

Maeda, S. (1976). *A characterization of American English intonation*. Ph.D. dissertation, MIT.

Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception and Psychophysics, 62*, 253–265.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 397–408.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59*, 203–243.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465–494.

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., Pleydell-Pearce, C. W., Melhorn, J. F., & Whitecross, S. E. (2005). Detecting silent pauses in speech – A new tool for measuring on-line lexical and semantic processing. *Psychological Science, 16*, 958–964.

Mattys, S. L., & Samuel, A. G. (2000). Implications of stress-pattern differences in spoken-word recognition. *Journal of Memory and Language, 42*, 571–596.

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.

McAuley, J. D. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing*. Unpublished Ph.D. dissertation, Indiana University.

McAuley, J. D., & Dilley, L. C. (2004). Acoustic correlates of perceived rhythm in spoken English. *Journal of the Acoustical Society of America, 115*, 2397.

McAuley, J. D., & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1102–1125.

Menn, L., & Boyce, S. (1982). Fundamental frequency and discourse structure. *Language and Speech, 25*, 341–383.

Millotte, S., Rene, A., Wales, R., & Christophe, A. (2008). Phonological phrase boundaries constrain the on-line syntactic analysis of spoken sentences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 874–885.

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science, 32*, 398–417.

Morgan, J. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language, 35*, 666–688.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453–467.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63*, 234–245.

Nam, H., Goldstein, L., & Saltzman, E. (2006). Dynamical modeling of supragestural timing. In *Proceedings of the 10th laboratory phonology conference*. Paris, France.

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., & Jusczyk, P. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech, 48*, 279–298.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris Publications.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica, 66*, 95–112.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299–370.

Patel, A. D., Peretz, I., Tramo, M., & Labreque, R. (1998). Processing prosodic and musical patterns: A neuropsychological investigation. *Brain & Language, 61*, 123–144.

Patel, A. D., Wong, M., Foxton, J., Lochy, A., & Peretz, I. (2008). Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception, 25*, 357–368.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6*, 688–691.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). *The communicative lexicon hypothesis*. Paper presented at the proceedings of the annual meeting of the Cognitive Science Society, Amsterdam.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

Pierrehumbert, J. (2000). Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and experiment* (pp. 11–36). Dordrecht: Kluwer Academic Publishers.

Pitt, M., & Samuel, A. G. (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 564–573.

Port, R. F. (2003). Meter and speech. *Journal of Phonetics, 31*, 599–611.

Povel, D. J., & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2*, 411–440.

Quene, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics, 20*, 331–350.

Quene, H. (1993). Segment durations and accent as cues to word segmentation in Dutch. *Journal of the Acoustical Society of America, 94*, 2027–2035.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. *Linguistic Controversies*, 73–79.

Salverda, A. P., Dahan, D., & McQueen, J. (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.

Schubiger, M. (1958). *English intonation. Its form and function*. Verlag/ Tübingen: Max Niemeyer.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shattuck-Hufnagel, S. (1995). *Pitch accent patterns in adjacent-stress vs. alternating-stress words in American English.* Paper presented at the International Congress of Phonetic Sciences, Stockholm.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*, 193–247.

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

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

Shockey, L. (2003). *Sound patterns of spoken English*. Cambridge: Blackwell.

Slevc, L. R., Rosenberg, J. C., & Patel, A. D. (2009). Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic Bulletin and Review, 16*, 374–381.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America, 100*, 2417–2485.

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

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9*, 271–294.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- and 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Thomassen, J. M. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America, 71*(6), 1596–1605.

van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 394–417.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 710–720.

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language, 38*, 133–149.

Wagner, M. (2005). *Prosody and recursion*. Ph.D. dissertation, Massachusetts Institute of Technology.

Woodrow, H. (1909). A quantitative study of rhythm. *Archives of Psychology, 14*, 1–66.

Woodrow, H. (1911). The role of pitch in rhythm. *The Psychological Review, 18*, 54–77.

Wurm, L. H. (2000). Auditory processing of polymorphemic pseudowords. *Journal of Memory and Language, 42*, 255–271.

Zatorre, R. J., & Gandour, J. (2007). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B, 363*, 1087–1104.