Attention, Perception, & Psychophysics (2019) 81:571–589
https://doi.org/10.3758/s13414-018-1626-4

Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables

Melissa M. Baese-Berk[1] · Laura C. Dilley[2] · Molly J. Henry[3] · Louis Vinke[4] · Elina Banzina[5]

Published online: 28 November 2018

Abstract
Listeners resolve ambiguities in speech perception using multiple sources, including non-local or distal speech rate (i.e., the speech rate of material surrounding a particular region). The ability to resolve ambiguities is particularly important for the perception of casual, everyday productions, which are often produced using phonetically reduced forms. Here, we examine whether the distal speech rate effect is specific to a lexical class of words and/or to particular lexical or phonological contexts. In Experiment 1, we examined whether distal speech rate influenced perception of phonologically similar content words differing in number of syllables (e.g., *form*/*forum*). In Experiment 2, we used both transcription and word-monitoring tasks to examine whether distal speech rate influenced perception of a reduced vowel, causing lexical reorganization (e.g., *cease*, *see us*). Distal speech rate influenced perception of lexical content in both experiments. This demonstrates that distal rate influences perception of a lexical class other than function words and affects perception in a variety of phonological and lexical contexts. These results support a view that distal speech rate is a pervasive source of information with far-reaching consequences for perception of lexical content and word segmentation.

Keywords Speech perception · Spoken word recognition · Word perception

Introduction

Even under ideal conditions, speech perception is a challenging task. When listening to everyday speech, a listener is exposed to a signal that contains multiple ambiguities that could be parsed in a number of different ways. Because speech does not generally contain clear acoustic cues to word boundaries, the issue of how a listener segments the stream into individual words has been a long-standing topic of interest. Listeners have been shown to use a multitude of different types of cues to cope with this challenge, including fine phonetic detail (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; McMurray, Tanenhaus, & Aslin, 2002), phonotactic information (McQueen, 1998; Vitevitch & Luce 1999), suprasegmental cues, e.g., timing (Beach, 1991; Cutler & Norris, 1988), and syntactic and semantic predictability (Mattys, Melhorn, & White, 2007; Morrill, Baese-Berk, Heffner, & Dilley, 2015; Staub & Clifton, 2006). These cues, ranging from very fine-grained and acoustic to broader and more knowledge-based, are integrated to interpret the signal in a relatively seamless way. How do listeners determine the most likely meaningful units – words, syllables, and phonemes – that comprise the speaker's message?

Recent years have seen an upsurge of interest in the role of *timing* as a cue that may be used by listeners to recover a speaker's intended meaning (Bell et al., 2009; de Ruijter, Mitterer & Enfield, 2006; Levinson, 2016; Dilley & Pitt, 2010; Peelle & Davis, 2012; Gahl, Yao, & Johnson, 2012; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013;

✉ Laura C. Dilley
ldilley@msu.edu

Melissa M. Baese-Berk
mbaesebe@uoregon.edu

[1] Department of Linguistics, 1290 University of Oregon, Eugene, OR 97403, USA

[2] Department of Communicative Sciences and Disorders, Michigan State University, Oyer Center, 1026 Red Cedar Road, East Lansing, MI 48824, USA

[3] Department of Psychology, Brain and Mind Institute, University of Western Ontario, Social Science Centre Rm 7418, London, Ontario N6A 5C2, Canada

[4] Center for Systems Neuroscience, Boston University, One Silber Way, Boston, MA 02215, USA

[5] Department of Linguistics, Stockholm School of Economics in Riga, Strelnieku iela 4a, Riga LV-1010, Latvia

Ravignani, Honing, & Kotz, 2017; Reinisch, 2016; Seyfarth, 2014; van de Ven & Ernestus, 2017; Ding, Patel, Chen, Butler, Luo, & Poeppel 2017). Of particular interest for our work is how timing information in the speech *context* may influence – or even determine – perceived lexical interpretations in cases of ambiguity that relate specifically to the *number of syllables* in the utterance. For example, the disyllabic word *align* is potentially confusable with the monosyllabic word *line* in the phrase *I align the paper*; not only are *align* and *line* different lexical items, but they also differ in numbers of syllables. Phonetically, timing cues may distinguish productions of similar-sounding yet distinct lexical parses differing in syllable count (Browman & Goldstein, 1990; Dilley, Arjmandi, & Ireland, 2017; Fougeron & Steriade, 1997; Manuel et al., 1992). Importantly, such cases of ambiguities in numbers of syllables associated with different lexical items (e.g., *align* vs. *line*) are distinguished from cases of optional syllable reduction for the same lexical item. For example, *camera* has a variant pronunciation *cam'ra* (Bürki, Fougeron, Gendrot, & Frauenfelder, 2011; Davidson, 2006; LoCasto & Connine, 2002; Patterson, LoCasto, & Connine, 2003). While *camera* and *cam'ra* are merely variants of one another that differ in numbers of syllables, phonetically similar *align* versus *line* correspond to different lexical entries with distinctive meanings.

It is critically important to understand the kinds of information that listeners might use to narrow ambiguous lexical interpretations, given a potentially large set of similar-sounding lexical candidates. An often-overlooked problem is that, in deciding how many syllables are present in the lexical items that form the final parse, listeners also must somehow arrive at a decision of how many word boundaries are present. Word boundaries are not consistently marked acoustically, posing a challenge to understanding the process by which listeners recover a speaker's intended words (e.g., Mattys, Davis, Bradlow & Scott, 2012; Mattys, White & Melhorn, 2005; Norris & McQueen, 2008). Thus, phonetically similar but lexically distinctive parses involving different numbers of syllables may form a single lexical item flanked by a word boundary on either side (e.g., *align* vs. *line* in contexts like *I align (the paper)* vs. *I line (the paper*), or they may correspond to different numbers of syllables, as well as different numbers of word boundaries (e.g., *guide* vs. reduced *guy had* [ga # d]). Such ambiguities in the speech signal have been documented in multiple corpus and acoustic-phonetic studies (Ernestus & Warner, 2011; Johnson, 2004; Kohler, 1998, 2006; Niebuhr & Kohler, 2011; Schuppler, Ernestus, Scharenborg, & Boves, 2011). The existence of such varied phonetic ambiguities greatly increases the set of meanings and structures that might be entertained in lexical search (e.g., Brouwer, Mitterer & Huettig, 2012; Poellmann, Bosker, McQueen, & Mitterer, 2014). Further, semantic and/or syntactic context are not guaranteed to sufficiently narrow down the set of candidate interpretations (cf. Snedeker & Trueswell, 2003).

The significance of the problem of how listeners narrow the set of lexical interpretations that might be considered, given a large set of phonetically overlapping candidates, is further highlighted by the fact that many lexical items begin with a vowel. Vowel-initial words often show few proximal acoustic discontinuities at their onsets that might provide overt local parsing cues for initiating segmentation (Dilley et al., 2017; Dilley, Shattuck-Hufnagel, & Ostendorf, 1996). While glottal irregularities may sometimes be observed at the onsets of vowel-initial words (Dilley et al., 1996), more frequently, heavy co-articulation of the vowel-initial word is observed. In such cases, the word's onset will often lack any detectable amplitude dip, voicing irregularity, noise, or other discontinuity that could serve as a proximal segmentation cue (Dilley et al., 2017). This contrasts with cases in which a word begins with a consonant, where the oral constriction gesture for the consonant will often result in local discontinuities (e.g., in amplitude, voicing, or fundamental frequency; Stevens, 2000); such consonant-like acoustic discontinuities provide segmentation cues to listeners about numbers of syllables and associated lexical interpretations (Heffner, Dilley, McAuley, & Pitt, 2013; Hillenbrand & Houde, 1996). The regular alternation of vocalic syllable nuclei and consonant onsets/codas leads to characteristic amplitude envelope modulations in speech, which have further been shown to be important for intelligibility and to drive entrainment of neural oscillations (Bosker & Ghitza, 2018; Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Doelling, Arnal, Ghitza, & Poeppel, 2014). These entrainment responses to amplitude envelope variation reflect categorical knowledge of speech-specific phonemes rather than simply a passive response to the envelope (Di Liberto, O'Sullivan, & Lalor, 2015).

The challenge of understanding how listeners tell apart lexically distinct but phonetically similar parses differing in numbers of syllables is compounded by the fact that many words whose careful or canonical pronunciations are consonant-initial often have reduced or variant pronunciations that are vowel-initial. For example, canonically /h/-initial words may be spoken with reduced or elided cues to /h/ (e.g., *had* can be carefully produced as [hæd] or casually produced as vowel-initial [d], or *his* [h z] as [z]; Dilley et al., 2017; Pierrehumbert & Talkin, 1991). Similar to words that are lexically vowel-initial (e.g., *align*), variant vowel-initial word forms (e.g., [d] for *had*) may show reduced or totally absent proximal acoustic discontinuities (e.g., glottalization or frication noise for /h/), which could otherwise serve as overt segmentation cues to the start of the word or presence of an extra syllable. The problem of how listeners recover representations under the highly variable pronunciations that occur in casual, everyday speech is a critically important and

challenging problem in understanding the neurocognitive mechanisms of speech perception (Alexandrou, Saarinen, Kujala, & Salmelin, 2018; Drijvers, Mulder & Ernestus, 2016; Ernestus & Warner, 2011; Tucker & Ernestus, 2016).

It is well known that casual speech can result in extreme reductions in pronunciation (e.g., Ernestus & Warner, 2011). However, surprisingly little research has focused on how listeners recover lexical representations when the speech results in ambiguities in numbers of syllables arising from potential co-articulation of vowel-initial words (or word variants), e.g., one-syllable *line* versus two-syllable *align*, one-syllable *guide* versus two-syllable *guy had* [ga # d]. By contrast, the vast majority of speech perception studies have focused on ambiguities where the number of syllables does not change across lexical parses, including studies of ambiguities in consonant-initial words, e.g., *run picks* versus *rum picks* (Gaskell & Marslen-Wilson, 2001; Gow, 2001), or in ambiguities in numbers of word boundaries, e.g., *two lips* versus *tulips* (Gow & Gordon, 1995), or *topic* versus *top pick* (Pickett & Decker, 1960).

The present studies tested the hypothesis that a specific type of timing cue – namely, *distal speech rate* (Dilley & Pitt, 2010) – would be an influential factor in listeners' resolution of lexical ambiguities arising from different candidate numbers of syllables in possible parses. Distal speech rate is defined here as the speech rate of syllables surrounding a region with potential co-articulation across adjacent syllables (e.g., syllables that begin and/or end with sonorants, such as vowels, liquids, or glides). In the present studies, we demonstrate that distal speech rate influences the number of syllable listeners perceive – and that this subsequently influences the lexical parses that they hear. In so doing, we hoped to determine whether effects of distal speech rate can be generalized from affecting vowel-initial monosyllabic function words – previously shown by Dilley and colleagues to be affected by distal speech rate – to influence perception of a wide variety of reduced, vowel-initial syllables. If so, this would show that the effects of distal speech rate are not just limited to function words but instead affect a wide variety of reduced vowel-initial syllables and words.

Timing cues in perception and production

A sizeable body of research has shown that timing cues are critical for understanding speech. For example, even when spectral cues are severely degraded, listeners are capable of robust speech understanding using primarily timing cues (Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Listeners also use duration of words and syllables to disambiguate potentially ambiguous stretches of speech (e.g., whether /kæp/ is the first syllable of *captain*; Davis, Marslen-Wilson, & Gaskell, 2002). The duration of single segments influences perception of ambiguous speech in both offline (Kemps, Ernestus, Schreuder, & Baayen, 2004; Quené, 1992) and online tasks (Gow & Gordon, 1995; Shatzman & McQueen, 2006). Timing is also used in *production* to differentiate potentially ambiguous stretches of speech. For example, speakers lengthen portions of ambiguous sentences to differentiate between possible interpretations (Turk & Shattuck-Hufnagel, 2007). Timing cues are also important for phoneme perception. For example, Repp, Liberman, Eccardt, and Pesetsky (1978) demonstrated that relative timing information of silence and noise influenced perception of ambiguous sounds constituting either a stop-fricative sequence or an affricate (e.g., *great ship* vs. *grey chip*; see also, e.g., Dorman et al. 1976; Liberman et al. 1956; Lisker & Abramson, 1967; Miller & Liberman, 1979). Speakers have also been shown to normalize for speech rate in a manner that can influence which phoneme they perceive (Miller, Aibel, & Green, 1984).

In addition to highly local cues such as the duration of individual words, syllables, or segments, non-local cues such as context, or distal, speech rate has also been shown to influence perception. Phoneme identification is subject to distal rate effects, particularly when the phonemes have similar spectral qualities and are typically differentiated by durational information (e.g., voice onset time distinctions in English; Liberman, Delattre, Gerstman, & Cooper, 1956; Miller & Liberman, 1979). When the surrounding speech rate is relatively slow, listeners perceive a segment as being shorter than when that same segment is presented in a situation where the distal speech rate is relatively fast. Listeners are more likely to perceive a consonant with an ambiguous duration as being a singleton when the distal speech rate is relatively slow and as a geminate when the distal speech rate is relatively fast (Pickett & Decker, 1960). That is, in a sentence such as *He was the topic of the year*, which could be perceived as *He was the top pick of the year*, listeners were more likely to interpret the ambiguous region as a geminate (i.e., *top pick*) if the speech rate surrounding the ambiguous region was relatively fast. Previous studies suggest that listeners normalize for distal speech rate in making judgments about phoneme identity (Bosker, 2017; Miller & Liberman, 1979; Pisoni, Carrell, & Gans, 1983; Sawusch & Newman, 2000).

Distal context speech rate and spoken word recognition

The studies reviewed above show that timing can influence phonetic perception. A newer and growing body of work has shown *distal* context speech rate influences can influence appearance or disappearance of entire lexical items. A seminal study in this line by Dilley and Pitt (2010) involved examination of highly co-articulated, ambiguous regions of

speech as in the sentence *Don must see the harbor or boats …* in which the word (e.g., *or*) was pronounced with a great degree of co-articulation and minimal amplitude envelope cues to the presence of the function word (i.e., *Don must see the harbor boats*). Dilley and Pitt manipulated the distal speech rate of the area surrounding the target. When the surrounding speech rate was relatively fast, either due to compressing the speech rate of the distal speech or expanding the speech rate of the target region, listeners were more likely to report having heard the function word than when the distal speech rate was relatively slow. They interpreted this "disappearing word effect" through a similar lens of rate-normalization, that listeners create expectations for upcoming spoken material that influence whether a function word is present in an otherwise ambiguous region. Such an account is compatible with a growing theoretical framework involving data-explanation/predictive coding frameworks. Under such a data-explanation view, a syllable is heard or not heard based on generative processes within perceptual systems that give rise to probabilistic expectations about incoming sensory input based on high-level representations and contextual information (Brown et al., 2014; Brown & Kuperberg, 2015; Kurumada, Brown, & Tanenhaus, 2017; Norris, McQueen, & Cutler, 2016; Olasagasti, Bouton, & Giraud, 2015; Park, Thut, & Gross, 2018; Pickering & Garrod, 2013; Pitt, Szostak, & Dilley, 2016; Tavano & Scharinger, 2015).

Subsequent studies of this "disappearing word effect" demonstrated that distal speech rate interacts with a number of factors during spoken word recognition, including intensity, frequency, and word duration (Heffner et al., 2013), speech rhythm (Morrill, Dilley, McAuley, & Pitt, 2014), linguistic knowledge (Morrill, Baese-Berk, Heffner, & Dilley, 2015), global speech rate (Baese-Berk et al., 2014), and speech signal intelligibility (Pitt et al., 2016). Of particular interest, Heffner et al. (2013) manipulated acoustic cues that affected the clarity of amplitude envelope cues to the presence of a word. They demonstrated that the distal speech rate effect was highly robust and could cause a word to disappear, even when proximal amplitude envelope cues to the word were *clear and salient*. However, distal speech rate had the largest effects when proximal amplitude envelope cues to the word onset were acoustically weak. While it is clear that distal speech rate can influence perception of ambiguous speech, the types of lexical items and phonological contexts over which this mechanism operates are unclear.

Motivation for present study

Critically, studies of distal context speech rate have so far focused on whether whole words disappear or appear, specifically using stimuli that involved ambiguity in function word presence or absence. Further, many of the prior studies have only examined cases where the ambiguous region is flanked by a word boundary on both sides, with less work generalizing this phenomenon to other kinds of phonetic and structural ambiguities. In the present study, we ask about the scope of this distal speech rate effect, examining a range of word classes and phonetic environments, to better understand whether this effect is limited to function words in a restricted set of lexical/phonological contexts. The present work thus aimed to test the hypothesis that distal speech rate is a factor that constrains lexical dynamics, allowing listeners to efficiently recover a speaker's message, given a large array of phonetically similar-sounding parses potentially differing in numbers of words, syllables, and/or phonemes (*align* vs. *line*, *guy had* vs. *guide*).

Specifically, we hypothesized that distal speech rate is applicable to perception of (metrically and/or acoustically) weak syllables more generally, providing important cues to presence of a syllable boundary (and potentially a word boundary) when amplitude envelopes are not clear and other cues (e.g., F0) do not appreciably vary.[1] Consonants typically have a relatively clear acoustic boundary, but vocalic segments, such as those examined in the present study, often do not. In the present study, we investigate the hypothesis that distal speech rate provides general and pervasive effects in lexical perception and word segmentation across contexts.

Some initial evidence in support of this hypothesis came from a study by Dilley, Morrill and Banzina (2013). They demonstrated that the disappearing word effect emerges in Russian for both cases with clear word boundaries surrounding the ambiguous region and those without (see also O'Dell & Niemenen, 2018 for a similar finding in Finnish). However, their Russian stimuli involved a variety of heterogeneous items, and it is possible that their effects were driven by item heterogeneity. That is, many of their stimuli were of the "disappearing word" type, and it is possible that these items drove their effect. In Experiment 1, we directly test cases where the two possible interpretations are different lexical items, specifically content words (e.g., *form* vs. *forum*), rather than function words.

Related to the question of whether the distal rate effect operates most powerfully over word boundaries is the possibility that distal speech rate affects only function words, since function words behave differently from content words in a great many ways. The meaning of a sentence may in many cases be clear to the listener with or without a function word. Morrill et al. (2015) examined function

[1] It should be noted that there are many cues to word segmentation other than the amplitude envelope (e.g., F0, intensity, and word duration, among others). These cues clearly interact with distal speech rate (see Heffner et al., 2013); however, the focus of the present study is unclear amplitude envelope cues and the use of distal speech rate to resolve ambiguities resulting from these unclear cues.

word perception in conjunction with the distal speech rate, demonstrating that listeners show the disappearing word effect and interpret function words as missing even when the resulting sentence is not grammatically well formed in English. Thus, function words may carry a different informational load to content words. Moreover, function words are typically reduced acoustically (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Some theories of syntactic reduction predict that the presence or absence of a function word modulates the information density of an utterance (Jaeger, 2010). In some forms of professional jargon, such as Aviation English, function words are entirely omitted, and yet communicative clarity is not lost (Farris & Barshi, 2013). As a result, function words as a lexical class may be particularly vulnerable to the disappearing word effect. If listeners are aware that function words are especially likely to be reduced, they may be more willing to perceptually repair utterances that are lacking function words if evidence in the distal speech rate is consistent with the presence of a function word. It is possible that this type of repair is primarily limited to function words and does not operate as robustly over content words.

Recent studies have begun to more systematically investigate how distal context speech rate can affect the interpretation of upcoming speech material, systematically focusing on the role of distal speech rate in segmentation (e.g., Canadian notes vs. Canadian oats; Heffner, Newman & Idsardi, 2017). However, the study by Heffner et al. (2017) did not focus on how distal rate may affect numbers of syllabic units. Our hypothesis is that distal context speech rate affects the dynamics of neural entrainment for distinct lexical parses (Brown, Dilley, & Tanenhaus, 2014; Tavano & Scharinger, 2015), which would be expected to result in different numbers of *syllabic units*.

In the present study, we examine perception of content words (e.g., *tear* vs. *terror*) to address whether distal speech rate effects are limited to the lexical class of function words. We hypothesize that distal speech rate influences perception of phonological material not just for the lexical class of function words, but rather is a more general mechanism impacting perception of syllables, especially those that are weak and that lack clear amplitude onsets. If so, then distal speech rate should influence the perception of content words, in addition to function words. If the distal speech rate effect, which influences perception of lexical content and word boundaries, extends to content words as well as function words and to other phonological contexts, these results would suggest that distal speech rate may have a far greater influence on word segmentation and spoken word recognition as they are traditionally defined. Further, if these effects emerge in both untimed and timed processing, we can conclude that distal speech rate is a powerful factor in spoken word recognition.

Overview of current studies

Here, we present a series of studies examining the role of distal speech rate in speech perception. In Experiment 1, we ask whether distal speech rate can influence which of two content words the listener perceives when those words are rendered ambiguous in casual speech (e.g., *form/forum*). In Experiment 2a, we examine how distal speech rate influences perception of ambiguous words and phrases with and without schwa (e.g., *press/oppress* and *cease/see us*). Unlike previous studies, we examine ambiguity word/phrase initially (*press/oppress*), as well as word/phrase finally (*cease/see us*). In Experiment 2b, we examine how distal speech rate influences perception in a complementary on-line task, word detection. In all three experiments we manipulate distal speech rate and ask whether relative rate information influences lexical access and spoken word recognition.

Experiment 1

In Experiment 1, we test whether the distal speech rate effect is limited to function words by investigating whether this rate effect influences lexical interpretations where alternative interpretations of phonological material are content words. For example, in the word *forum*, the second syllable begins with a sound potentially lacking a salient amplitude envelope dip at onset and possibly being ambiguous with the phonologically related word *form*.

Methods

Participants

Participants (n = 40, age range = 18–42 years, mean = 20.1 years, SD = 3.9 years, 21 female) were adult native speakers of General American English with no reported speech or hearing difficulties. Informed consent was obtained for all participants in this experiment and all subsequent experiments. No individual participated in more than one experiment.

Materials

Sentences were based on target two-syllable trochaic words (e.g., *forum*) that contained a weak syllable preceded by a rhotic sound; the amplitude envelope cues for this weak syllable therefore lacked a clear spectro-temporal discontinuity. Each trochaic word had a corresponding phonologically-related, one-syllable lexical counterpart omitting the weak syllable (e.g., *form*). For example, in the sentence *Jim has to tick the forum today,* the word *forum* could be reduced and produced in a way that would make it ambiguous with *form*. A total of 18 pairs of words were created.

Stimuli were recorded using a memory task following Pitt, Dilley, and Tat (2011). In the task, participants were asked to memorize a sentence on a screen and were then presented with an additional word to integrate into the sentence. They were instructed to speak this new sentence, including the integrated extra word into the microphone. Previous work has suggested that this paradigm results in sufficiently informal speech allowing for the co-articulation of interest in the present study. The 11 talkers (age range 18–23 years, mean = 19.5 years, SD = 1.5 years, six female) for the experiment were students at Bowling Green State University who spoke General American English. The original (i.e., "unaltered") speech rates varied, and therefore the slowed and sped distal rates also varied, following other studies of distal speech rates.

Following Dilley and Pitt (2010), we defined a target region as consisting of the syllable before the ambiguous portion of the sentence and the phoneme immediately following the ambiguous region. Therefore, for the sentence *Jim has to tick the forum today*, the target region is *forum t-*. The context region was the rest of the sentence. We used the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2015) to resynthesize the stimuli for all distal rate conditions.

We created three versions of each stimulus, one for each distal rate condition. Each distal rate condition manipulated the relationship between the speech rate of the target region and the speech rate of the context region. Figure 1 shows a schematic for these manipulations. For the Unaltered Rate Condition, the speech rate matched the original speech rate of the stimulus and the target and context regions were resynthesized at the same rate. The other two distal rate conditions involved a change of 40% relative to the original stimulus duration, following the conventions in Dilley and Pitt (2010). In the Context-Expanded Condition, the duration of all segments in the context portion of the stimuli were slowed by a factor of 1.4. This resulted in a duration of this portion that was 140% of the original stimulus (an increase of 40% relative to the original stimulus). For this distal rate condition, the speech rate of the target region was resynthesized at the originally spoken rate (i.e., a duration factor of 1.0). This resulted in a distal rate condition where the context was relatively slow compared to the target region. In the Target-Compressed Condition, the target region was made faster by a factor of .6, resulting in a duration that was 60% of the original stimulus (a decrease by 40% relative to the original stimulus). The speech rate of the context region was resynthesized at the original spoken rate (i.e., a duration factor of 1.0). As in the Context-Expanded Condition, the context was relatively slow compared to the target region. Because one portion of the utterance was held constant in each of the two distal rate conditions, comparing each distal rate condition to the Unaltered Rate Condition allowed us to examine the effect of differing relationships in speech rate between the context and target regions.

It is important to note that our previous work demonstrated that the distal speech rate effect is *not* simply an effect of manipulating the speech rate of any part of the sentence. Previous work has demonstrated that when both the target and context regions are manipulated, no such distal rate effect emerges. That is, when both the target and context are compressed, native listeners report a similar number of function words as cases where the speech rate is presented at the originally spoken rate (Baese-Berk, Morrill, & Dilley, 2016; Dilley & Pitt, 2010). Therefore, any rate effects seen here

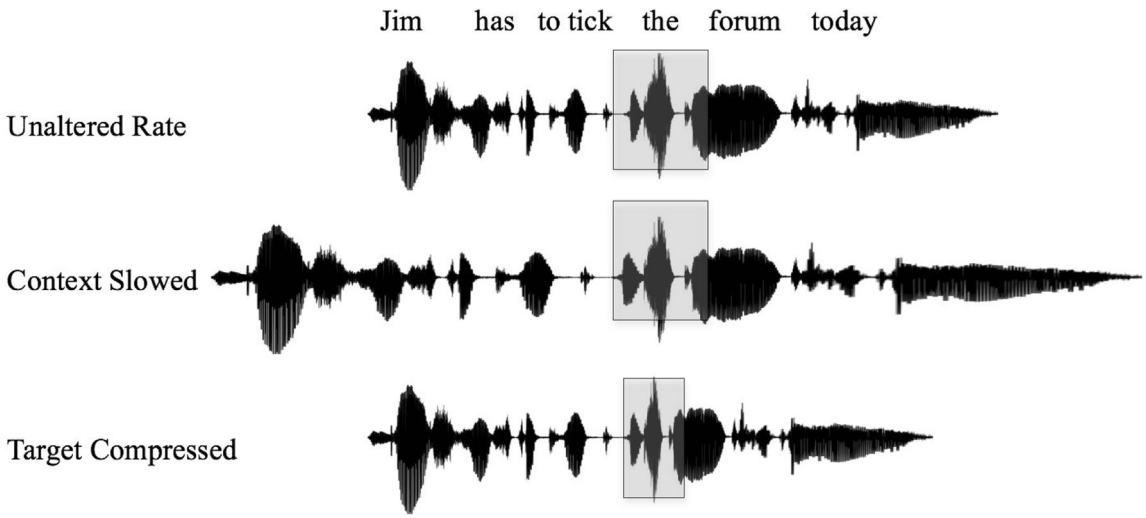

Fig. 1 Schematic of the three rate manipulations for each distal rate condition in the experiments. The target region is highlighted in grey, and the context region is not highlighted. In this example, the Unaltered Rate stimulus is 2.6 s long, and the Context Slowed stimulus is 3.64 s long. (Note that the target region is identical in duration in the Unaltered Rate and Context Slowed conditions.) The Target Compressed version is approximately 2.4 s, as the target region is compressed by a factor of .6. (Note, however, that the context region is identical in the Target Compressed and Unaltered Rate conditions)

are unlikely to emerge from artifacts in the recordings or manipulations of speech rate generally. Rather, rate effects can be attributed to the specific rate manipulations made here.

Procedure

Participants listened to utterances and saw a portion of the utterance on a computer screen. For example, in the case of *Jim has to tick the forum/form today*, participants saw "Jim has to tick …" on the screen. They were asked to type the continuation of the sentence on the keyboard provided to them. Each participant heard the target in only one of the three experimental distal rate conditions. Assignment of utterances to distal rate conditions was counterbalanced across participants, such that each participant heard each sentence, but the distal rate condition in which the sentence was presented differed across participants.

Analysis

Following Dilley and Pitt (2010), we excluded any trials (<2%) that indicated inattention to the task (e.g., cases where participants did not report hearing something phonologically similar to the target). We analyzed the transcriptions to determine if participants reported hearing the "long" version of the target or not, where a long response is defined as a disyllabic word (e.g., *forum*). These responses were analyzed in R (R Development Core Team, 2014) using mixed effects logistic regressions in the lme4 package (Bates, Maechler, Bolker, & Walker, 2014), with the proportion of "long" reports as the dependent variable (0 being the short response and 1 being the long response). We used model comparisons to determine what fixed and random effects were included in the final model, and to determine significance of the fixed effects. Random effects included in the model were the maximal effects allowing for the model to converge and included random intercepts for item and participant, as well as random slopes for the two distal rate comparisons by subject, including correlations between these random effects. The fixed effect included in the model was distal rate condition (unaltered, context-expanded, target-compressed), which was contrast-coded. Two comparisons of this contrast are reported: Unaltered versus Rate Altered (Unaltered Rate: -1, Context Expanded: 0.5, Target Compressed: 0.5) and Context Expanded versus Target Compressed (Unaltered Rate: 0, Context Expanded: -1, Target Compressed: 1). Model comparisons were used to determine significance of each factor.

Results

Figure 2 shows the proportion of "long" responses for our participants. Similar to previous studies, even in the Unaltered Rate Condition participants do not report hearing the long response in all cases, even though it was produced in the recordings. Importantly, the figure also demonstrates that when stimuli were presented in the Context-Expanded and Target Compressed Conditions, participants reported hearing "longer" versions less often than in the Unaltered Rate Condition (Mean proportion longer response (standard deviation): Unaltered Rate = .79 (.14), Context-Expanded = .49 (.17), Target Compressed = .65 (16)).

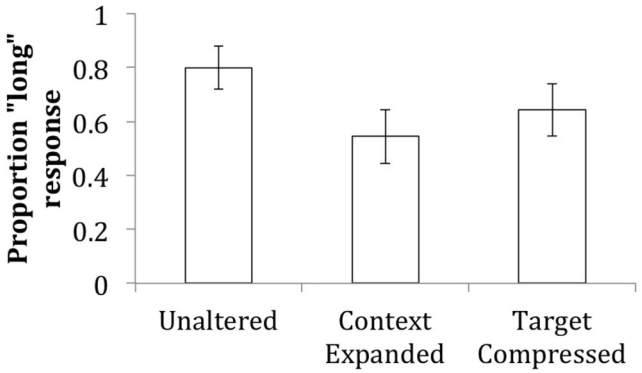

Fig. 2 Proportion of "long" responses (e.g., *forum*) in each of the three distal rate conditions. Error bars represent two standard deviations of the mean

The mixed-effect model supports these observations. The inclusion of the comparison between Unaltered Rate and Rate Altered conditions significantly improved model fit (β = -2.6645 *s.e.* = 0.5559, z = -4.793, χ^2 = 22.314, p < .001). The inclusion of the comparison between Context-Expanded and Target-Compressed conditions did not significantly improve model fit (β = -0.6910, *s.e.* = 0.3684, z = -1.876, χ^2 = 3.5, p = .06).

These results show that distal speech rate affected perception of content words, as has previously been shown for function words. In particular, distal speech rate caused weak syllables of trochaic words to be perceived when the target region was relatively long compared to the context. Conversely, distal speech rate caused weak syllables to disappear from perception when the target region was relatively short compared to the context (which was brought about apparently equally well by expanding/slowing the context or compressing the target). Importantly, distal rate cues caused perceptual reorganization of sonorant material into different numbers of syllables (and hence, different numbers of phonemes), where either interpretation resulted in a percept of a content word. Further, the type of speech rate manipulation (i.e., whether the context was slowed or the target was compressed) does not seem to matter as much as the relative speech rates of the two portions. These results therefore supported the hypothesis that the distal speech rate effect is not one that is limited to function words, but rather one that applies to weak syllables lacking clear amplitude envelope cues to syllable onset, regardless of lexical form class.

Experiment 2a

The results of Experiment 1 showed that a disyllabic content word containing a weak syllable could be perceived as a monosyllabic content word lacking that syllable. Thus, distal rate influenced the manner in which phonetic content present in the signal was perceived to be organized phonologically. Specifically, a content word like *forum* could be heard as another content word, *form*, where the latter lacked the weak syllable and was essentially missing a schwa (/ /).

However, it is possible that distal speech rate is a cue that potentially disambiguates phonological structures in an even wider set of contexts and circumstances. In Experiment 2, we hypothesized that distal speech rate could cause a reorganization of phonetic content into phonological structures in ambiguous sequences typically lacking clear amplitude envelope cues to the start of a new unit. In particular, we designed materials to test whether locations of word boundaries could be assigned variably to different positions with respect to the phonetic speech signal within the sonorant stretch, as a function of distal speech rate. If so, then we predicted that adjoining phonetic content likewise would be assigned variably to very different phonological positions within lexical structures. In Experiment 2, the stimuli differed with regard to which of two content words might be perceived in the ambiguous region or, in some cases, whether one or two words might be perceived. For example, a sonorant stretch containing a lengthened /i/-like segment, such as /si:s/, might either potentially correspond to a single lexical item with /s/-segments in both word-onset and word-final position (i.e., *cease*), or it might contain a word boundary that lacked clear amplitude envelope cues to its onset (as in the phrase *see us* spoken in a reduced fashion). That is, we predicted that distal speech rate would influence listeners to hear different phonological structures involving different numbers of phonetic segments and syllables and reassignment of phonetic material into phonological "slots." For example, we predicted that distal speech rate would cause a structure like /si:s/ to, on the one hand, sound like cease – which would correspond to a phonemic parsing as /#sis#/ (consisting of one monosyllabic word with a closed, CVC syllable structure), versus, on the other hand, *see us* – which would correspond to a phonemic parsing as /#si# s#/ (consisting of two monosyllabic words, with an open CV syllable followed by an onsetless closed syllable VC). If distal speech rate were shown to operate in such a fashion that it could influence parsing of phonetic material into quite different phonological and lexical structures, it would support the hypothesis that distal speech rate is a powerful factor influencing lexical perception and word segmentation, shaping perception in a wide range of phonological and phonetic contexts.

Methods

Participants

Participants (n = 40; age range = 18–38 years, mean = 20.1 years, 22 female) were adult native speakers of General American English with no reported speech or hearing difficulties. No individual participated in any of the other experiments presented here.

Materials

Eighteen pairs of words or phrases were created for this experiment. Recording conditions were identical to those described above in Experiment 1. Talkers were nine students at Bowling Green State University, who did not record stimuli for Experiment 1. All pairs contained an ambiguous target region, which contained a schwa (or schwar). For example, in the sentence *John thought he might see us talking in the theatre*, "*see us*" could be produced with a substantial amount of co-articulation rendering it ambiguous with *cease*. Seven pairs of words contained ambiguity word or phrase finally, as in the case of *cease/see us* above. Eleven pairs of words contained ambiguity word or phrase initially (e.g., *Being a nosy person I peer/appear around the corner*). Note that the "long" version was always the intended target. In both the Unaltered Rate and Context-Expanded conditions, the signal actually contained the extra syllable (i.e., *see us* was always produced). The stimuli were altered in the same manner as those in Experiment 1, creating three experimental distal rate conditions: Unaltered Rate, Context-Expanded, and Target Compressed. Additionally, 36 fillers were included in the experiment.

Procedure

The procedure was identical to that in Experiment 1.

Analysis

Analysis procedures were similar to those in Experiment 1. Fixed effects included in the model were location of ambiguity (initial vs. final) and distal rate condition, which was contrast-coded. Two comparisons of this contrast are reported: Unaltered Rate versus Rate Altered and Context Expanded versus Target Compressed. The interactions between each type of context rate comparison and location of ambiguity were also included in the models. Random effects included in the model included a random intercept for item and random slopes for distal rate condition (unaltered vs. rate altered) and location of

ambiguity by participant.[2] Model comparisons were used to determine significance of each factor.

Results

Figure 3 shows the proportion of "long" responses (i.e., responses where the ambiguous region is parsed as two syllables, rather than one syllable) in each of the three experimental context rate conditions. As in Experiment 1, participants reported fewer "long" responses in the Context-Expanded and Target-Compressed conditions (mean proportion longer response (SD): Unaltered Rate = .89 (.13), Context-Expanded = .56 (.15), Target Compressed = .62 (14)). Results of our mixed model analysis support these observations. Unaltered Rate versus Rate Altered conditions were a significant predictor of model fit (β=-4.051, *s.e.*=1.685, *z*=-2.90, χ^2=7.5, *p*<.01). Comparing the two rate-altered conditions, we see no significant difference in proportion of long responses (χ^2<1, *p*>.3). Interestingly, whether the ambiguity was in the initial or final position did not significantly predict whether participants reported a "long" response (χ^2<1.5, *p*>.2), suggesting that both types of ambiguity are susceptible to effects of speech rate. This observation is supported additionally by the lack of significant interaction between location of ambiguity and either rate manipulation (all χ^2<1, *p*>.5). Results therefore supported the hypothesis that distal speech rate influences listeners to hear different phonological structures involving different numbers of phonetic segments and reassignment of phonetic material into phonological "slots."

Experiment 2b

Taken with the results of Experiment 1, the findings in Experiment 2a suggest that distal speech rate may be a powerful factor in perception of lexical content and word segmentation, and that its effects are not limited to function words or contexts with specific locations of word boundaries. Instead, the results support the view that the effects of distal speech rate are much broader, influencing the integration of multiple types of cues in speech processing in a variety of conditions. However, previous results examine only one side of lexical access and spoken word recognition, asking whether listeners identify an ambiguous region as one lexical item or another. It is unclear whether distal speech rate influences timed lexical perception during spoken word recognition, in addition to influencing identification after an entire sentence has been heard and a final parse has been determined. Therefore, in Experiment 2b, we use a word-monitoring task to examine whether distal speech rate also influences perceptual processing before a final parse is achieved in lexical access and spoken word recognition.

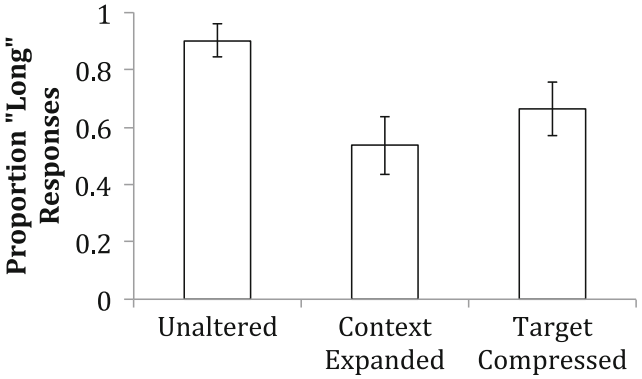

Fig. 3 Proportion of "long" responses (e.g., *see us*) in each of the three distal rate conditions. Error bars represent two standard deviations of the mean

Methods

Participants

Participants (*n*=20; age range = 18–40 years, mean = 20.3 years, SD = 4.8 years, 14 female) were native American English speakers with no reported speech or hearing difficulties. No individual participated in any of the other experiments presented here.

Materials

The materials used in the present study were the stimuli from the Unaltered Rate and Context-Expanded rate conditions in Experiment 2a. A total of 72 fillers were also included in the experiment. To obscure the fact that experimental items involved verb targets, with ambiguities involving a potential embedding of a vowel-initial word, targets in a number of filler items were verbs and/or vowel-initial words that did not involve potential embeddings.

Procedure

As in Experiments 1 and 2a, participants heard items presented in one of the two distal rate conditions. However, instead of transcribing the sentence, participants were presented with two words on the right and left of the screen and were asked to indicate which of the two options they heard via the computer keyboard using two keys labeled "RIGHT" and "LEFT" (corresponding to the buttons they were supposed to press if they

[2] It should be noted that a model with a fully specified random effect structure (i.e., including all distal rate conditions) did not converge. We attempted analyses first removing correlations between the random effects, which did not converge. Similarly, a model with interactions between the distal rate conditions and location of the ambiguity in the random effect structure also did not converge, nor did a model that included both distal rate condition comparisons in the random effect structure. However, given the fact that this relatively complex model did converge, we believe our results are robust.

selected the item on the right or left of the screen, respectively). One word was always the target word that was actually presented in the sentence (e.g., *appear*) and the other was a foil (e.g., *peer*). In cases where the ambiguity resulted in two content words (e.g., *peer/appear)*, they monitored for whether they heard the longer parse (e.g., *appear*) or the shorter parse (e.g., *peer).* In cases where the item contained a short phrase as one of the two options, they listened for the second half of the "long" version if the long version was a short phrase (e.g., *us* for the *cease/see us* item) versus the single word in the "short" version (e.g., *ceases* for the *cease/see us* item). Note that, as in Experiment 2a, the "long" version was always the intended target, and the signal actually contained the extra syllable (i.e., *see us* was always produced). Therefore, accuracy can be defined as whether the participant detected the target.

Fillers were semantically and structurally similar to experimental items. In the case of fillers, one of the two words shown on the screen had occurred in the filler item ("target"), while the other had not ("foil"). Most foil options in fillers were either phonologically related to the target (e.g., *paired* when the target was *pair*) or semantically related to the whole sentence (e.g., *done* for the filler *Did Louis cook the hot dogs well enough?* – with target *enough*). The target words displayed on the screen for filler trials were a mixture of content and function words.

In addition to accuracy, reaction time was measured from the end of the target for each trial. Participants were allowed to respond at any time during the trial. Reaction time was calculated from the onset of the target word. Since the proximal acoustics of the target word were identical across distal rate conditions, there was no confound between the proximal acoustics and the reaction time.

Analysis

We conducted two mixed-effect model analyses. For accuracy, we conducted a logistic mixed effect regression (correct responses coded as 1, incorrect responses coded as 0) with distal rate condition as a fixed factor and random intercepts for subject and item, and random slopes for distal rate condition by subject and item. For reaction time, we conducted a linear mixed effect regression with distal rate condition, accuracy, and their interaction as fixed factors, random intercepts for subject and item, and random slopes for distal rate condition by subject and item. Model comparison was used to determine significance of each factor.

Results

Accuracy

Figure 4 shows the average proportion of correct responses for the word-monitoring task. It is clear that participants were substantially less accurate in the Context-Expanded rate condition. The results of the mixed model analysis support these observations. Distal rate condition was a significant predictor of accuracy (β = -2.099, *s.e.* = 0.482, z = -4.357, χ^2 = 51.293, $p < .001$). As in Experiments 1 and 2, distal speech rate influences participants' perception of ambiguous stretches of speech.

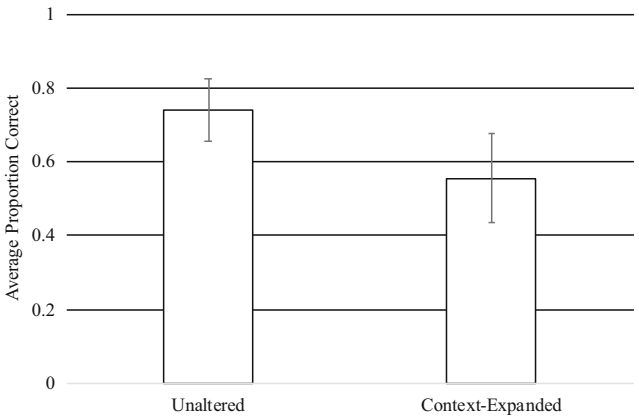

Fig. 4 Proportion of "long" responses (e.g., *see us*) in each of two experimental distal rate conditions. Error bars represent two standard deviations of the mean

Reaction time

Figure 5 shows the reaction time for each distal rate condition when the responses were correct and incorrect. Overall, it is clear that the reaction times are longer for the Context-Expanded Condition compared to the Unaltered Rate Condition. This is particularly true for correct responses.

The results of the mixed effects regression support the observation that reaction times are longer for the Context-Expanded Condition compared to the Unaltered Rate Condition. Accuracy was a significant predictor of reaction time (β = -765.5, *s.e.* = 374.6, t = 8.066, χ^2 = 35.56, $p < .004$), with incorrect responses being slower than correct

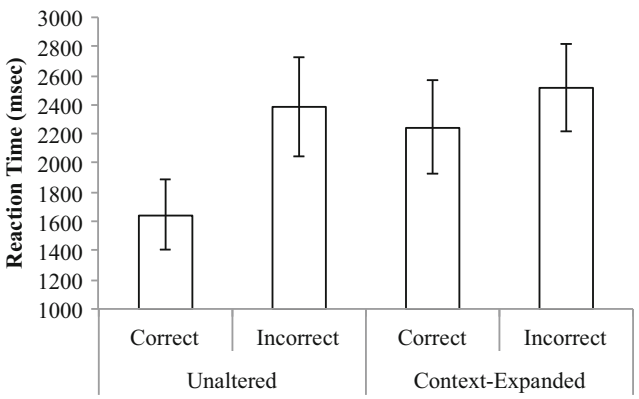

Fig. 5 Reaction time in each of two experimental distal rate conditions, including correct and incorrect responses. Error bars represent two standard deviations of the mean

responses. Further, distal rate condition was a significant predictor of reaction time ($\beta = 62.3$, *s.e.*= 335.2, $t = 7.802$, $\chi^2 = 35.39$, $p < .004$). Interestingly, the interaction between distal rate condition and accuracy was not significant ($p > .25$). Taken together, these results support the view that distal speech rate influences on-line lexical perception and word segmentation in timed and untimed perceptual processing, helping to resolve phonetically ambiguous weak syllables lacking clear amplitude envelope cues to onset. Participants hearing Context-Expanded tokens were less accurate and slower to respond, even in the cases where they did so accurately. This finding is particularly interesting because in the cases of the Context-Expanded Condition, listeners had more time to process the context, which might have led to the expectation that they might be faster at responding to targets. However, they were considerably slower in this distal rate condition, as predicted by a view that distal speech rate resulted in the same acoustic material sounding like a less "good" example of the target word. The results suggested that listeners utilized the rate information to resolve this ambiguity in exactly the ways that were predicted if word recognition is sensitive to distal speech rate in timed lexical processing. These results are compatible with previous studies demonstrating the influence of distal rate on word segmentation (e.g., Breen, Dilley, McAuley, & Sanders, 2014; Brown, Dilley, & Tanenhaus, 2012) using a different experimental paradigm. Thus, the results provide further support that distal speech rate is a powerful factor in word segmentation, assisting with perception of weak syllables lacking clear amplitude envelope cues.

General discussion

The present studies tested whether distal speech rate could induce changes in whether listeners heard one or two syllables for sonorant stretches of speech in a wide array of lexical and phonological structures. We tested the hypothesis that distal speech rate – a factor that has previously been shown to cause listeners to hear or not hear a monosyllabic function word (e.g., Dilley & Pitt, 2010) – could influence whether listeners heard a reduced syllable in a much wider variety of lexical contexts and materials than have been previously tested. Of note, we tested whether distal speech rate could influence perception of content words (such as *form* vs. *forum*), and/or cause phonological restructuring of a "chunk" of speech (such as *guide* vs. *guy had* [ga # d]). Given that casual, everyday speech often involves substantial phonetic reduction and ambiguity (Ernestus & Warner, 2011; Johnson, 2004; Shockey, 2008), such a demonstration would be a powerful proof-in-concept that provided further evidence of a role of distal speech rate in recovery of a speaker's intended words during typical conversational dynamics. The results of the three studies presented here provide firm evidence that distal speech rate influences spoken word recognition in a much wider set of contexts than has previously been reported. In particular, we found that distal speech rate caused changes to perception of content words – e.g., *form* vs. *forum* – and that it could cause reorganization of speech material into very different imputed phonological structures involving different numbers of syllables and word boundaries (e.g., *guide* vs. *guy had* [ga # d], *I align* vs. *I line*, etc.)

Previous studies of the distal speech rate effect have utilized materials containing function words, which can disappear perceptually when the distal speech rate surrounding a stretch of speech containing the function word is manipulated to be relatively slow compared to the rate of the stretch of speech itself (cf. Dilley & Pitt, 2010). Previous research has not eliminated the possibility, however, that the distal speech rate effect was particularly influential for function words, since such words play a distinctive role in syntactic constructions and typically carry a different functional load than do content words. The present results, however, suggest that, while function words may be more likely to be phonetically reduced, distal speech rate influences speech perception more broadly. The effect manifests not only when a function word could be perceived or not perceived, but also when the identity of a content word is ambiguous (e.g., *form* vs. *forum*). Further, distal speech rate influences the number of syllables perceived in a variety of phonetic contexts, even when there is not a word boundary on one side of the ambiguous region. This suggests that the effect demonstrated here and in previous studies is not limited to word segmentation involving function words or specific phonetic contexts, but is a more general effect that influences spoken word recognition and speech segmentation.

These effects are particularly interesting because they provide a segmentation and lexical recognition mechanism for detecting syllables with otherwise subtle acoustic evidence of their presence (i.e., in cases where acoustic discontinuities that might otherwise provide a phonetic cue to a word onset are weak or missing). Many syllables begin with consonants, which usually generate localized acoustic discontinuities in the speech signal (e.g., amplitude reduction or frication noise; Stevens, 2000). The temporal pattern of these consonantal discontinuities is faithfully reflected in the dynamics of neural oscillations, which are coupled to the amplitude envelope of speech during comprehension (Keitel, Gross & Kayser, 2018; Alexandrou, Saarinen, Kujala, & Salmelin, 2018; Kösem et al., 2018; Zoefel, Archer-Boyd, & Davis, 2018). Adults listeners are particularly attuned to consonants for lexical search as onsets (Havy, Serres, & Nazzi, 2014; New and Nazzi, 2014), consistent with the fact that consonants are a statistically reliable indicator of lexical distinctiveness cross-linguistically (Oh, Coupé, Marisco, & Pellegrino, 2015). However, many other syllables that start with vowels or

semivowel consonants, i.e., liquids (/l, r/) or glides (/j, w/), thus will show more or less continuous acoustic transitions relative to surrounding vowel sounds (Wright, 2004; Zhou, Espy-Wilson, Boyce, Tiede, Holland & Choe, 2008). Notably, many vowel-initial syllables optionally begin with glottal onsets in English, which generate a spectrotemporal discontinuity and a salient increase in amplitude at vowel onset, but a great many do not show any such discontinuity (Dilley et al., 1996; Dilley et al., 2017; Redi & Shattuck-Hufnagel, 2001). Even when reduced syllables are spoken with minimal observable acoustic evidence of their onsets (e.g., minimal amplitude envelope reduction at onset), statistical distributions of temporal cues to a reduced syllable are nevertheless largely separated from those of such cues in utterances that truly lack an extra syllable (e.g., *one small step for a man* vs. *one small step for man*; Baese-Berk et al., 2016, Dilley et al., 2017). While syntactic and semantic information play a role in recognizing reduced pronunciation variants (Eisner & McQueen, 2018; Ernestus, Baayen, & Schreuder, 2002), such information cannot explain the robustness of human speech perception. Thus, there is a need for identifying further statistically reliable cues that assist listeners with recovery of the lexical items comprising a speaker's intended message.

The present results showed that distal speech rate contributes to perception of syllables across a range of lexical types and phonetic contexts. These results effectively demonstrate that word boundaries can be "imputed" – "heard" or not "heard" – in the middle of a highly sonorant stretch of speech as a function of the distal context speech rate. This seems to imply that potentially *any* stretch of sonorant material – including single vocalic nuclei and/or unitary sonorant consonants like /ɹ/, /l/, /j/, and /w/, is somewhat ambiguous with regard to whether it contains a word boundary. This observation – that distal speech rate can potentially lead listeners to hear or not hear a word boundary in a sonorant stretch of speech (where an onset might have been realized with weak envelope cues or none at all) – raises the question of when and why a search for a new word boundary is initiated or not. We propose that distal speech rate actively limits lexical searches for otherwise possible alternative parses, and that distal rate information would be informative in a statistical sense about the temporal properties of the *intended* upcoming structures. Previous work has demonstrated that not all possible parses are entertained throughout lexical selection. For example, Norris, Cutler, McQueen, and Butterfield (2006) showed that embedded words were rapidly inhibited during lexical access and competition. This suggests that some mechanism must limit when a search is initiated, and we propose that distal rate is a potential candidate for a limiting mechanism.

The emerging picture of distal speech rate effects is that of very different implied parses of a given stretch of acoustic material – i.e., different numbers and imputed locations of syllabic, lexical, and/or phonetic boundaries. For a variety of reasons, this picture is consistent with the sort of real-time gradient upweighting and downweighting of lexical alternatives that is posited to occur under data explanation and predictive coding approaches (e.g., Norris, McQueen, & Cutler, 2016; Pickering & Garrod, 2013; Tavano & Scharinger, 2015).

Crucially, Brown and colleagues recently have demonstrated evidence for such dynamic upweighting and downweighting of lexical candidates (Brown, 2014; Brown, Dilley & Tanenhaus, 2012, 2014). Using stimuli involving ambiguities between indefinite singular and plural forms such as *see a raccoon swimming* versus *see raccoons swimming*, Brown and colleagues showed that gradient changes to distal speech rate resulted in gradient changes to the proportion of looks to a picture of a singular referent (e.g., one raccoon) versus a plural referent (e.g., two raccoons). Further, distal speech rate both to the left and to the right of target material modulated listeners' consideration of lexical alternatives. Such experiments provided further evidence of gradation in consideration of lexical alternatives as a function of gradient modifications to distal speech rate (Heffner et al., 2013).[3]

Further consistent with a data explanation or predictive coding view, recent evidence supports the view that top-down information from syntactic and semantic context is necessary in order for listeners to use distal speech rate to disambiguate proximal phonetic material into words. Pitt et al. (2016) degraded distal context in different ways (e.g., noise vocoding, sine-wave speech, replacement with tone sequences). They found that temporal cues in distal context were insufficient to convey rate information that resulted in a change in perception of a function word, except in distal rate conditions where the context could be interpreted as an intelligible speech signal.[4] Intelligible contexts – degraded speech precursors and intelligible sinewave speech – permitted distal speech rate to influence whether a reduced function word was heard or not, but unintelligible signals (speech or non-speech) did not. This finding highlights the fact that the distal speech rate effect involves top-down, knowledge-driven expectations about timing properties of upcoming speech, rather than being strictly a bottom-up cue. The importance of signal intelligibility can be understood to translate

[3] As noted by an anonymous reviewer, our current work could provide additional evidence for this gradation, given that our expansion and compression rates were slightly different factors (i.e., a comparable expansion rate for the context expanded condition as compared to the target compressed condition would actually be 1.66, not 1.4). However, given the lack of a significant difference between the target compressed and context expanded conditions, we caution over-interpretation of any numerical trends in our data.

[4] Note that this finding contrasts with previous work in reconciling phoneme level ambiguity (e.g., Bosker, 2017), which demonstrates that the intelligibility of the precursor does not influence resolution of lower-level (e.g., phoneme) ambiguities in the same way it influences lexical and syllabic level ambiguities. Understanding this distinction should be considered in future research.

to a significant role for top-down predictions about possible candidate lexical items as a means of supporting the usage of low-level, bottom-up temporal cues for recovery of reduced forms. Given the heterogeneity of lexical structures that are influenced by distal speech rate, as shown by this study in connection with others, the distal speech rate effect thus appears to share many similarities with classic studies of phoneme restoration, which also show significant top-down influences (e.g., Samuel, 1981, 1996; Warren, 1970; Warren & Sherman, 1974).

Further, the emerging picture is that distal speech rate may contribute in different ways to perception and segmentation of phonetic units, depending on the nature of potential ambiguity in the signal. The distal effects on lexical perception demonstrated by Dilley and colleagues (e.g., Dilley & Pitt, 2010) – which has involved ambiguities in the number of syllables (not just function words) that are heard – is surprisingly robust, critically *so long as* the context is intelligible so as to support top-down predictions (Pitt et al., 2016). This "disappearing word effect" further exhibits what might be called a "dose-response" relationship – a gradient change in distal rate manipulation corresponds to a gradient change in the likelihood of hearing an extra syllable, item variability notwithstanding (Brown, Dilley & Tanenhaus, 2014; Heffner et al., 2013; see also Bosker & Ghitza, 2018, for a similar result with phoneme-level effects). By contrast, effects of distal speech rate on perception of units smaller than the syllable – specifically, phonemes – appear considerably more variable in size and reliability of effect. Historically, distal rate effects on phoneme-level perception, e.g., for consonants like /g/ vs. /k/ differing in voice-onset time, have been recognized to be fairly fragile, with generally small effects on phonetic boundary shifts (Kidd, 1989; Summerfield, 1981; Wade & Holt, 2005). Recent studies have generalized knowledge of distal rate effects on speech perception by considering a wider array of phoneme-level perceptual ambiguities, both ambiguities in type of phoneme, e.g., / / versus / :/ in Dutch (Bosker, 2017), and number of phonemes, e.g., *Canadian oats* versus *Canadian notes* (Heffner et al., 2017; Reinisch, Jesse, & McQueen, 2011; Reinisch & Sjerps, 2013), and singleton/geminate contrasts (Mitterer, 2018) and have demonstrated effect sizes more comparable to those seen the "disappearing word" cases. While variability in study designs obviously raises caveats to any generalizations across these studies – including in the type of dependent measure used (e.g., temporal boundary shifts vs. proportion of responses indicating an extra unit) – it tentatively appears to be the case that distal rate effects on *phoneme*-level perceptual ambiguities (vowel-length contrasts, phoneme count, singleton/geminate) may be less robust than distal rate effects on syllable-level perceptual ambiguities (*leisure or time*, *I align* vs. *I line*, *guide* vs. *guy had*).

To explain this apparent difference in robustness of distal rate effects, we point to likely differences in degree of statistical dependency between context distal rate and syllable-level ambiguities versus phoneme-level ambiguities. A number of recent studies have focused on predictability and duration for a given potentially reduced lexical item (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gahl, Yao, & Johnson, 2012; Jaeger, 2010; Kuperman, Pluymaekers, Ernestus, & Baayen, 2007; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013), but there has been less attention to statistical dependencies between rate of units in context and those in the to-be-parsed ambiguous chunk associated with constraints on a speaker of planning a connected utterance (Dell, 2013; Dilley et al., 2017; Indefrey & Levelt, 2004; Krivokapic, 2007, 2014). Recent work has shown high statistical dependency between distal context speech rate and proximal syllable duration (Baese-Berk, Dilley, Schmidt, Morrill, & Pitt, 2016; Dilley et al., 2017). Further, the consistent range of spectrotemporal modulation in the amplitude envelope further supports the hypothesis that the brain is specialized for a biologically limited range of temporal information supporting the combinatorics of human language (Ding et al., 2017). However, there is little reason to believe that distal context speech rate would reliably predict statistically the durations of individual consonants, given the variability in lexical structures associated with the unfolding utterance. Thus, we propose as a hypothesis for future testing that differences in the informativity of distal speech rate for different levels of *proximal* linguistic structure – vis à vis (proximal) *syllables* versus *phonemes* – can explain differences in the robustness of distal rate effects on syllable versus phoneme level phonetic perception. For example, directly examining the statistical relationships between distal rate and syllable versus phoneme level production would help hone predictions that could be used in perception testing.

Relatedly, the fact that distal speech rate effect emerges in a variety of phonetic contexts, as demonstrated here, is particularly interesting because of the special status that has been afforded to onsets as a result of previous studies of speech perception and production (Allopenna, Magnuson, & Tanenhaus, 1998; Kessler & Treiman, 1997; Marslen-Wilson & Zwitserlood, 1989; Wilshire, 1999). Many models of speech perception have persisted in affording word onsets a special status (Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978). The TRACE model (McClelland & Elman, 1986) and SHORTLIST (Norris, 1994) propose that both onsets and rhymes are quite important for perception. However, it is important to note that even these models cannot account for the types of "misperceptions" seen in our current studies, wherein the percept does not necessarily match with either the onset or the rhyme (e.g., *cease* and *see us* have quite different syllable structures, in addition to differing in the number of words intended by the speaker). Instead, it seems, models of

spoken word recognition should account for overall phonetic match, including a possible normalization for speech rate, when determining the candidates for activation and ultimate identification. The Neighborhood Activation Model (Luce & Pisoni, 1998) predicts that words will be activated as a result of their global similarity to the target; however, NAM calculates similarity over words with the same number of segments. Further, the NAM does not address how supra-segmental features such as speech rate may influence spoken word recognition and lexical segmentation when more than one parse is available to the listener. These results suggest that current models of spoken word recognition may require modification to account for the results presented here.

One intuitive explanation for the effects seen here, and in other studies examining distal speech rate, is that of expectation, as modulated by context. Prediction and expectation have long been known to influence speech perception, broadly. For example, Ganong (1980) demonstrated that the perception of an ambiguous phoneme could be modulated by whether that phoneme was presented in a real word or a non-word. That is, in cases where listeners might expect to hear a real word, their perception of an ambiguous phoneme was modulated by this expectation. Timing also clearly influences expectations about upcoming information, including word boundaries and syllable. Several previous studies have examined how listeners resolve lexical embedding during perception. That is, listeners must be able to differentiate between a word like *hamster* and the word *ham,* which is embedded in the carrier word. These embedded words phonetically match the first syllable of the carrier word, thus they provide substantial competition for the carrier word under most models of spoken word recognition. However, previous studies have suggested that the amount of competition is actually much smaller than might be expected (Davis, Gaskell, & Marslen-Wilson, 1998; Davis et al., 2002). In fact, Zhang and Samuel (2015) demonstrated that embedded words prime carrier words only under very specific circumstances (e.g., the embedded word comprises a large part of the carrier word). Norris, Cutler, McQueen, and Butterfield (2006) suggest that embedded words can be rapidly suppressed, presumably to limit search options during lexical retrieval. One could imagine that cases such as those examined in the present studies are analogous to embedded words. That is, *cease* shares many phonemes with *see us*, so it is possible that both are highly active, and suppression is necessary to determine which option is the correct target. However, what is the mechanism that drives that suppression, or limits the search possibilities?

It is possible that timing information is one mechanism by which a search among multiple possibilities is constrained. Salverda, Dahan, and McQueen (2003) asked how listeners may be able to determine quickly whether the target word is a carrier word or an embedded word by examining the acoustic properties of embedded and carrier words. They demonstrate that the monosyllabic word carries fine-grained phonetic detail that listeners use to predict an upcoming word boundary. Specifically, they suggest that speakers lengthen the final segment before a prosodic boundary (Kingston & Beckman, 1990; Turk & Shattuck-Hufnagel, 2000), and listeners use this information in perception to predict upcoming boundaries, and disambiguate the target from potential competitors.

Because timing information is critical in spoken word recognition and word segmentation, and is used to disambiguate competitors in other contexts, it is possible that timing-related expectation and prediction play a critical role in the distal speech rate effect shown here. If listeners predict not only what material the speaker may say next, but also how much information they might produce, using distal speech rate would be a very useful cue. That is, if a listener can predict, using the rough number of syllables produced in the previous stretch of time, the number of syllables or phonemes that might be produced in an upcoming utterance, prediction and expectation could play a very important role in providing disambiguating among multiple segmentation options. Perhaps, in addition to utilizing semantic and syntactic likelihood (Staub & Clifton, 2006), listeners also utilize speech rate to make predictions. The results from Heffner et al. (2013) suggest that the listeners utilize speech rate even when other cues to segmentation are present, and that listeners integrate speech rate with many other cues.

An integration of such cues would suggest that the speech recognition system is exquisitely sensitive to a number of cues, each helping to constrain the search parameters, allowing for the system to quickly eliminate irrelevant options that may result from a partially ambiguous signal. If one imagines that each vowel offglide or midpoint may be the start of a syllable, resulting in ambiguous options for parsing, the number of options the system may consider could be astronomical. Therefore, multiple cues must be used in conjunction to reduce the number of options considered by the system. In the case of distal speech rate effects, as listeners search for syllable onsets, entrainment to a distal speech rate may help constrain the search space, greatly reducing the number of ambiguous options considered during recognition. That is, if a listener has entrained to a particular speaking rate that might predict an onset in an ambiguous region (i.e., a faster speaking rate), the listener may determine that the most likely parse of this ambiguous region includes a syllable onset. Similarly, if they are entrained to a speaking rate that would suggest no such onset (i.e., a slower rate), they may determine that the most likely parse does *not* include a syllable onset.

The results of the present study suggest that the distal speech rate effect is not limited to word segmentation. This resonates with a broader view of language processing, promoted by Dahan and Magnuson (2006), who describe broad conceptions of spoken word recognition as a piece of a larger puzzle encompassing speech perception, recognition, word

segmentation, and processing more generally. While some models of speech segmentation can account for the results of previous studies (e.g., Mattys et al., 2007; Mattys & Melhorn, 2007), most current models of spoken word recognition cannot account for interactions, integration, and reconciliation of myriad factors such as speech rate that may also influence how ambiguous phonetic material is understood, and how this processing impacts word segmentation, recognition, and parsing more broadly. In the present studies, we demonstrate that the distal speech rate effect is rather general, and should be integrated into models of spoken word recognition from the broader viewpoint as promoted by Dahan and Magnuson (2006).

In summary, these results support the view that distal speech rate affects word segmentation and lexical perception, influencing perception of the number of syllables regardless of word class (i.e., content vs. function words) and in a wide range of positions relative to word boundaries. This suggests that distal speech rate may be a useful cue for effective parsing of casual, reduced speech, which contains highly variable pronunciations. Distal speech rate could cause listeners to reorganize the same stretch of acoustic material into different phonological structures involving different numbers of phonetic segments, different imputed locations of word boundaries, and restructuring of phonetic material into phonological "slots." The results of the present study add to a growing body of work suggesting that contextual information, including distal speech rate, is critically important for restricting the possibilities we consider when listening to speech. However, very few models of spoken word recognition or speech perception consider the role of such contextual information, especially in the timing domain. The case presented here is particularly interesting for development of future models and theories of spoken word recognition, because the results do not fit squarely in the domains of speech perception, spoken word recognition, or word segmentation, as they are typically defined. Instead, the work here falls into a broader view of language understanding, with distal speech rate influencing multiple aspects of understanding.

Acknowledgements This work was partially supported by an NSF Faculty Early Career Development (CAREER) Award and NSF grant BCS 1431063 to Laura C. Dilley and by a University of Oregon Faculty Research Award to Melissa M. Baese-Berk.

Appendix: Target Stimuli Lists

Experiment 1

Dan has to pet the bear/bearer today.
Deb has to peck the care/carer today.
Don has to tick the dear/dearer today.
Hal has to tack the ware/wearer today.
Jen has to pack to own/Owen today.
Jess has to pick the king/keying today.
Jim has to tick the form/forum today.
Jon has to tip the scene/seeing today.
Ken has to pack the line/lion today.
Kim has to peck the warn/warren today.
Liz has to kick the stair/starer today.
Nan has to pit the yearn/urine today.
Pam has to tack the pair/parer today.
Pat has to kick the ping/peeing today.
Rob has to tip the heir/error today.
Ron has to pit the hoard/horror today.
Tim has to pep the line/lion today.
Tom has to pep the poor/pourer today.

Experiment 2 (information in parentheses corresponds to the two alternatives displayed on the screen for Experiment 2b for experimental items)

Based on this new study, today I prove/approve the results. (prove/approve)

Being a nosy person, I peer/appear around the corner. (peer/appear)

Do you think it's wise to place/play us music on the piano. (place/us)

Each weekday morning, I rise/arise from bed for work. (rise/arise)

For these reasons I pose/oppose new arguments to the senate. (pose/oppose)

He should go ahead and bite/buy it instead of waiting. (bite/it)

I believe that the guide/guy had directed them outside. (guide/had)

I hate moving paintings so much, I tack/attack them angrily on the wall. (tack/attack)

It's bad to make yourself scarce/scare us without warning. (scarce/us)

It's easy to note/know it very well just by reading. (note/it)

Jan thought he might cease/see us talking in the theater. (cease/us)

Our goal is to seize/see his power before he does. (seize/his)

The doctor will wait/weigh it today for further lab analysis. (wait/it)

The people who I press/oppress into service hate me. (press/oppress)

To be friendly, I know point/appoint this fellow to the senate. (point/appoint)

With this new rule, I mend/amend our club's constitution. (mend/amend)

With this notice, I sign/assign him to your custody. (sign/assign)

When I refinish a room, I line/align it with the wallpaper. (line/align)

References

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018). Cortical tracking of global and local variations of speech rhythm during connected natural speech perception. *Journal of Cognitive Neuroscience*, 1–16.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439.

Baese-Berk, M. M., Dilley, L. C., Schmidt, S., Morrill, T. H., & Pitt, M. A. (2016). Revisiting Neil Armstrongs Moon-Landing Quote: Implications for Speech Perception, Function Word Reduction, and Acoustic Ambiguity. *PloS one*, *11*(9), e0155975.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-Term Temporal Tracking of Speech Rate Affects Spoken-Word Recognition. *Psychological Science*, *25*(8), 1546–1553.

Baese-Berk, M. M., Morrill, T. H., & Dilley, L. C. (2016). Do non-native speakers use context speaking rate in spoken word recognition?. In *Proceedings of the 8th International Conference on Speech Prosody (SP2016)* (pp. 979-983).

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, *30*(6), 644–663.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92–111.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. Available at www.praat.org. Acessed 21 Nov 2018.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. Attention, Perception & Psychophysics, 79, 333-343. https://doi.org/10.3758/s13414-016-1206-4.

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience, 33*(8), 955-967.

Breen, M., Dilley, L. C., McAuley, J. D., & Sanders, L. D. (2014). Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation. *Language, Cognition and Neuroscience*, *29*(9), 1132-1146.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539-571.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

Brown, M. (2014). *Interpreting Prosodic Variation in Context* (Unpublished doctoral dissertation). University of Rochester, Rochester, NY.

Brown, M., & Kuperberg, G. R. (2015). A hierarchical generative framework of language processing: Linking language perception, interpretation, and production abnormalities in schizophrenia. *Frontiers in Human Neuroscience*, *9*, 643.

Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2014). Probabilistic prosody: Effects of relative speech rate on perception of (a) word (s) several syllables earlier. In *Proceedings of the 7th International Conference on Speech Prosody, Dublin, Ireland, May 20–23* (pp. 1154-58).

Bürki, A., Fougeron, C., Gendrot, C., & Frauenfelder, U. H. (2011). Phonetic reduction versus phonological deletion of French schwa: Some methodological issues. *Journal of Phonetics, 39*(3), 279-288.

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

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

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In M. J. Traxler & M. A. Gernsbacher (Eds.), Handbook of Psycholinguistics (pp. 249-283). Amsterdam: Academic Press

Davidson, L. (2006). Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica, 63*, 79-112.

Davis, M. H., Gaskell, M. G., & Marslen-Wilson, W. (1998). Recognising Embedded Words in Connected Speech: Context and Competition. In *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997* (pp. 254–266). London: Springer London.

Davis, M. H, Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218–244.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515-535.

Dell, G. S. (2013). Cascading and feedback in interactive models of production: A reflection of forward modeling?. *Behavioral and Brain Sciences*, *36*(4), 351-352.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457-2465.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of vowelinitial syllables as a function of prosodic structure. *Journal of Phonetics*, 24, 423444.

Dilley, L. C., Arjmandi, M. K., & Ireland, Z. (2017). Spectrotemporal cues for perceptual recovery of reduced syllables from continuous, casual speech. *Journal of the Acoustical Society of America, 141*(5), 3700.

Dilley, L. C., Morrill, T. H., & Banzina, E. (2013). New tests of the distal speech rate effect: examining cross-linguistic generalization. *Frontiers in Psychology*, *4*.

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

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews, 81*, 181–187.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85*, 761-768.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1976). Further observations on the role of silence in the perception of stop consonants.

Journal of the Acoustical Society of America, 59, S40. doi: https://doi.org/10.1121/1.2002677

Drijvers, L., Mulder, K., & Ernestus, M. (2016). Alpha and gamma band oscillations index differential processing of acoustically reduced and full forms. *Brain and Language*, *153*, 27-37.

Eisner, F., & McQueen, J. M. (2018). Speech Perception. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Language and Thought*, *3*, 1.

Ernestus, M., Baayen, R. H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, *81*(1-3), 162-173.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics, 39*(SI), 253-260.

Farris, M. C., & Barshi, I. (2013). *Misunderstandings in ATC communication: Language, cognition, and experimental methodology*. Ashgate Publishing, Ltd

Fougeron, C., & Steriade, D. (1997). Does the deletion of French schwa lead to neutralization of lexical distinctions? In *Proceedings of Eurospeech* (Vol. 2, pp. 943-946).

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789-806.

Ganong, W. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language*, *44*(3), 325-349.

Gow Jr, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*(1), 133-159.

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

Havy, M., Serres, J., & Nazzi, T. (2014). A consonant/vowel asymmetry in word-form processing: Evidence in childhood and in adulthood. *Language and Speech*, *57*(2), 254-281.

Heffner, C. C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes, 28*(9), 1275-1302.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964-988.

Hillenbrand, J. M., & Houde, R. A. (1996). Role of F0 and amplitude in the perception of intervocalic glottal stops. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1182-1190

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101-144.

Jaeger, T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29-54).

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology, 16*(3), e2004473.

Kemps, R., Ernestus, M., Schreuder, R., & Baayen, R. H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, *90*, 117–127.

Kessler, B., & Treiman, R. (1997). Syllable Structure and the Distribution of Phonemes in English Syllables. *Journal of Memory and Language*, *37*(3), 295–311.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 736.

Kingston, J., & Beckman, M. E. (Eds.). (1990). Lengthenings and shortenings and the nature of prosodic constituency. In *Papers in Laboratory Phonology Volume 1, Between the Grammar and Physics of Speech* (pp. 152–178). Cambridge: Cambridge University Press.

Kohler, K. J. (1998). The disappearance of words in connected speech. *ZAS Papers in Linguistics*, *11*, 21-33.

Kohler, K. J. (2006). Paradigms in experimental prosodic analysis: from measurement to function. *Methods in Empirical Prosody Research*, (3), 123-152.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology, 28*(18), 2867–2875.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics, 35*(2), 162–179.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1658), 20130397.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261-2271.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2017). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*, 1-8.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6-14.

Liberman, A. M., Delattre, Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *52*(2), 127-137.

Lisker, L., & Abramson, A. S. (1967, 1970). *The voicing dimension: some experiments in comparative phonetics*. Paper presented at the Proceedings of the 6th International Congress of Phonetic Sciences, Prague.

LoCasto, P., & Connine, C. M. (2002). Rule-governed missing information in spoken word recognition: Schwa vowel deletion. *Perception & Psychophysics, 64*(2), 208-219.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words; the neighbourhood activation model. *Ear and Hearing*, *19*, 1–36.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313-318.

Manuel, S. Y., Shattuck-Hufnagel, S., Huffman, M. K., Stevens, K., Carlson, R., & Hunnicutt, S. (1992). *Studies of vowel and consonant reduction*. Paper presented at the 1992 International Confernce on Spoken Language Processing, University of Alberta: Edmonton, Canada.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953-978.

Mattys, S. L., & Melhorn, J. F. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. *Journal of the Acoustical Society of America*, *122*(1), 554–567.

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*(4), 960–977.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500.

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

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

McQueen, J. M. (1998). Segmentation of Continuous Speech Using Phonotactics, *Journal of Memory and Language, 39*(1), 21-46.

Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception and Psychophysics, 35*(1), 5–15.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Attention, Perception & Psychophysics*, *25*(6), 457–465.

Mitterer, H. (2018). The singleton-geminate distinction can be rate dependent: Evidence from Maltese. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 9*(1).

Morrill, T. H., Baese-Berk, M., Heffner, C., & Dilley, L. C. (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic Bulletin and Review*, *22*(5), 1451-1457.

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. *Cognition*, *131*(1), 69–74.

New, B., & Nazzi, T. (2014). The time course of consonant and vowel processing during word recognition. *Language, Cognition and Neuroscience*, *29*(2), 147-157.

Niebuhr, O., & Kohler, K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, *39*(3), 319-329.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2006). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology*, *53*(2), 146–193.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition, and Neuroscience, 31*(1), 4-18.

O'Dell, M., Nieminen, T. (2018) Distal rate effect for Finnish epenthetic vowels. Proc. 9th International Conference on Speech Prosody 2018, 646-650. https://doi.org/10.21437/SpeechProsody.2018-131.

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, *53*, 153-176.

Olasagasti, I., Bouton, S., & Giraud, A. L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex*, *68*, 61-75.

Park, H., Thut, G., & Gross, J. (2018). Predictive entrainment of natural speech through two fronto-motor top-down channels. *bioRxiv*, 280032.

Patterson, D. J., LoCasto, P., & Connine, C. M. (2003). A corpus analysis of schwa vowel deletion frequency in American English. *Phonetica, 60*, 45-68.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 320.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329-347.

Pickett, J. M., & Decker, L. R. (1960). Time factors in perception of a double consonant. *Language and Speech*, *3*, 11–17.

Pierrehumbert, J. and D. Talkin, (1991) Lenition of /h/ and glottal stop. Papers in Laboratory Phonology II, Cambridge Univ. Press, Cambridge UK. 90-117

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, *34*(4), 314–322.

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics*, *39*(3), 304-311.

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics*, *78*(1), 334-345.

Poellmann, K., Bosker, H. R., McQueen, J. M., & Mitterer, H. (2014). Perceptual adaptation to segmental and syllabic reductions in continuous spoken Dutch. *Journal of Phonetics*, *46*, 101-127.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*(3), 331–350.

R Development Core Team (2014). R: A language and environment for statistical computing. Vienna, Austria. <http://www.R-project.org>. Accessed 5 Aug 2018.

Ravignani, A., Honing, H., & Kotz, S. A. (2017). The evolution of rhythm cognition: Timing in music and speech. *Frontiers in Human Neuroscience*, *11*, 303.

Redi, L. & Shattuck-Hufnagel, S. (2001). Variation in realization of glottalization in normal speakers. *Journal of Phonetics, 29*, 407-429. doi: https://doi.org/10.1006/jpho.2001.0145

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397-1415.

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

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101-116.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 612-637.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension*Cognition, 90*(1), 51–89.

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes?. *Journal of Experimental Psychology: General*, *125*(1), 28.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Attention, Perception & Psychophysics*, *62*(2), 285–300.

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96-109.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133*(1), 140–155.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

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

Shockey, L. (2008). Sound patterns of spoken English. Wiley, Hoboken.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130.

Staub, A., & Clifton, C., Jr. (2006). Syntactic prediction in language comprehension: Evidence from either… or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425.

Stevens, K. N. (2000). *Acoustic phonetics*. Cambridge: MIT Press.

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

Tavano, A., & Scharinger, M. (2015). Prediction in speech and language processing. *Cortex*, *68*, 1-7.

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375-400.

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

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

Van de Ven, M., & Ernestus, M. (2017). The role of segmental and durational cues in the processing of reduced words. *Language and Speech, 61*(3), 358–383.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374-408.

Wade, T., & Holt, L. L. (2005). Effects of later-occurring nonlinguistic sounds on speech categorization. *Journal of the Acoustical Society of America*, *118*(3), 1701–1710.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392-393.

Warren, R. M., & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. *Attention, Perception, & Psychophysics*, *16*(1), 150-156.

Wilshire, C. E. (1999). The "Tongue Twister" Paradigm as a Technique for Studying Phonological Encoding. *Language and Speech*, *42*(1), 57–82.

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriad (Eds.) *Phonetically based phonology* (pp. 34–57).

Zhang, X., & Samuel, A. G. (2015). The activation of embedded words in spoken word recognition. *Journal of Memory and Language, 79*, 53–75.

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, *123*(6), 4466–4481.

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology*, *28*(3), 401-408.