Altering Context Speech Rate Can Cause Words to Appear or Disappear

Laura C. Dilley[1,2,3,4] and Mark A. Pitt[5]

[1]Department of Communicative Sciences and Disorders and [2]Department of Psychology, Michigan State University;
[3]Department of Psychology and [4]Department of Communication Sciences and Disorders, Bowling Green
State University; and [5]Department of Psychology, Ohio State University

Abstract

Speech is produced over time, and this makes sensitivity to timing between speech events crucial for understanding language. Two experiments investigated whether perception of function words (e.g., *or*, *are*) is rate dependent in casual speech, which often contains phonetic segments that are spectrally quite reduced. In Experiment 1, talkers spoke sentences containing a target function word; slowing talkers' speech rate around this word caused listeners to perceive sentences as lacking the word (e.g., *leisure or time* was perceived as *leisure time*). In Experiment 2, talkers spoke matched sentences lacking a function word; speeding talkers' speech rate around the region in which the function word had been embedded in Experiment 1 caused listeners to perceive a function word that was never spoken (e.g., *leisure time* was perceived as *leisure or time*). The results suggest that listeners formed expectancies based on speech rate, and these expectancies influenced the number of words and word boundaries perceived. These findings may help explain the robustness of speech recognition when speech signals are distorted (e.g., because of a casual speaking style).

The perception of spoken words is thought to depend largely on recovery of phonemic cues from frequency-specific (spectral) information (e.g., Marslen-Wilson & Welsh, 1978). Yet recognition of spoken words can be remarkably accurate when spectral cues are missing or severely distorted, as occurs, for example, in sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981), phase-vocoded speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), or auditory chimeras (Smith, Delgutte, & Oxenham, 2002). This fact suggests that temporal information is crucial for accurate spoken-word recognition. Despite compelling demonstrations, progress has been slow in understanding the role of speech timing (i.e., speech rate, speech duration, hierarchical rhythmic structure, etc.) in word recognition (Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003). In this study, we used natural, undegraded speech to show that entire words can seem to disappear or appear as a function of their context speech rate (i.e., the rate of all speech preceding and following the region of interest). The results have implications for understanding how adults and infants recognize spoken words and segment those words from speech, issues that are long-standing puzzles in the literature (e.g., Klatt, 1980).

We hypothesized that speech timing plays a decisive role in perceiving a word when the word's frequency spectrum shows substantial overlap or blending with the spectrum of adjacent words; this pervasive phenomenon is known as *coarticulation*. Coarticulation of adjacent words can sometimes be so severe that spectral information is insufficient to identify whether a given word is present in the speech stream, let alone where that word begins. This is especially true for short, high-frequency words, such as function words like *or* and *and* (Bell et al., 2003; Shockey, 2003).

We reasoned that when coarticulation of words is severe, the presence of a word could be conveyed by the duration of the blended phonemes relative to context speech rate. A typical case of heavy coarticulation is shown in Figure 1. The spectrum for the word *or*, spoken in its reduced form as *er*, blends almost totally with that of the preceding syllable *-sure*

Corresponding Author:
Laura C. Dilley, Michigan State University, Department of Communicative
Sciences and Disorders and Department of Psychology, 116 Oyer Building,
East Lansing, MI 43403
E-mail: ldilley@msu.edu
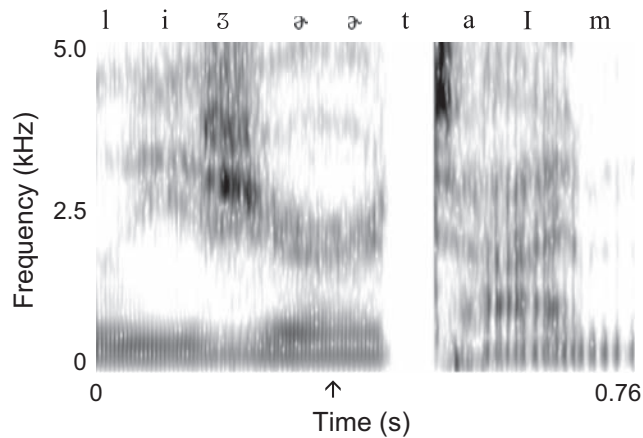
Fig. 1. Spectrogram illustrating heavy coarticulation of a function word in the phrase *leisure or time*. Phonemic content is shown as time-aligned International Phonetic Alphabet symbols at the top of the figure. The arrow on the *x*-axis indicates the approximate start of the function word *or*; note the utter absence of discontinuity marking the start of this word and the lack of clear cues differentiating the function word spectrally from the preceding syllable.

in the phrase *leisure or time*. Thus, there is a relatively homogeneous span of spectral material for most of the two syllables: -(s)*ure or*. We hypothesized that such a span is heard as two syllables because it is too long relative to the context speech rate to contain a single syllable. If this interpretation is correct, slowing down the context speech rate should make the span sound relatively shorter, like a single syllable, causing the function word to disappear and the phrase to be heard as *leisure time* instead of *leisure or time*.

One reason for thinking that speech rate might alter perception of a word's presence is that context speech rate affects the boundary between spectrally related phonemes (e.g., /p/ vs. /b/; see Liberman, Delattre, Gerstman, & Cooper, 1956; Miller & Liberman, 1979) and between singleton and geminate segments (Fujisaki, Nakamura, & Imoto, 1975; Pickett & Decker, 1960). We hypothesized that context speech rate could also affect the perceived presence of larger morphophonological units (i.e., words or syllables). This possibility stems from proposals concerning entrainment to temporal sequences; according to such proposals, an auditory event (e.g., a tone or syllable) of a given duration can be heard as corresponding to different rhythms (i.e., different numbers of "beats" or onsets). This correspondence would depend on the rate or rhythm of surrounding auditory events (e.g., Large & Jones, 1999; McAuley, 1995; Port, 2003; Povel & Essens, 1985; Saltzman & Byrd, 2000).

Rate normalization has been proposed as one mechanism behind speech-rate effects on phoneme boundaries (Miller & Liberman, 1979; Pisoni, Carrell, & Gans, 1983; Sawusch & Newman, 2000). Generalizing this account on the basis of entrainment, we hypothesized that listeners entrain to context speech and that this process affects the number of morphophonological units (words, syllables, segments) perceived in a given stretch of speech. According to this generalized-rate-normalization account, the lexical content (number of words)

in a spectrally ambiguous stretch of speech depends on the duration of that content relative to the surrounding speech rate, as well as other information, such as grammatical content. When a coarticulated stretch of speech is long relative to its surroundings, the listener should perceive a function word, because doing so is plausible given the rate cues, as well as higher-level information (e.g., semantic and syntactic context).

In two experiments, we tested whether the number of morphophonological units—here, the number of function words—is dependent on the duration of a given stretch of speech relative to the context speech rate, given grammatically viable contexts. For Experiment 1, we predicted that if context speech rate is made slow relative to a stretch of speech containing a function word—either by slowing down the context speech rate or by speeding up the stretch of speech itself—then that stretch of speech should be perceived as short and as containing fewer phonological units (i.e., fewer words). For Experiment 2, we predicted that if context events are made fast relative to a stretch of speech not containing a function word—either by speeding up the context speech rate or by slowing down the stretch of speech under analysis—then the stretch of speech should be perceived as relatively long and thus as containing an additional phonological unit (i.e., a function word that was never spoken).

Experiment 1
Method

Participants. Participants (*N* = 41) were young, American English speakers from the Midwest United States. All reported that they had normal hearing.

Materials. We constructed 50 sentences containing a critical function word embedded in a phonetic context expected to show heavy coarticulation with the function word; Table 1 lists fragments of these sentences used in experimental materials. Each sentence consisted of a grammatically acceptable beginning, whether or not the critical function word was present; that is, the span from the beginning of the sentence until just after the critical function word was grammatical, even if the function word was not present. For example, *Deena doesn't have any leisure or time* is a grammatically acceptable beginning for a sentence, even if the word *or* is missing.

Recordings of experimental stimuli were elicited from 29 speakers of American English from the Midwest United States. All but the last word of a sentence was presented on a computer screen in front of the speaker. The last word was presented only after the sentence had been erased for 1.5 s, at which point the speaker had to recite the sentence into a head-mounted microphone. Instructions emphasized accuracy in repeating the sentence verbatim; this procedure suggested that the experiment was investigating memory. Because they were not given any instructions pertaining to speech clarity, speakers

Table 1. Fragments Constituting the Speech Stimuli in Experiments 1 and 2

Taylor knew the principal and teacher (are) from Ohio . . .
Conor knew that bread and butter (are) both . . .
Frank thinks that sadness and anger (are) both . . .
Claire said that sour and bitter (are) both . . .
Chris said his mother and father (are) both . . .
Zach knew that there (are) things . . .
George thought my father and brother (are) like [good] . . .
Glenn thought his friend and neighbor (are) like plenty . . .
Ruth saw the maid and butler (are) at the top . . .
Rose knew that there (are) lamps . . .
The company moved to (a) different . . .
Trent might get to (a) certain . . .
Clay thinks that would be (a) good . . .
The Smiths wouldn't buy (a) Butterball . . .
Anne wanted to see (a) very funny . . .
It makes no sense to obey (a) petty . . .
It takes a lot of work to review (a) personal . . .
It costs a lot to tattoo (a) pink . . .
The boy wanted to glue (a) broken . . .
Dave asked how long it takes to repay (a) large . . .
Aspirin and other painkillers are (our) drugs . . .
The Murrays are (our) favorite . . .
The callers are (our) French contacts . . .
Mom said these are (our) gray gloves . . .
The accountants are (our) wise advisors . . .
Phil and Mary are (our) young cousins . . .
The leaves fell after (her) green . . .
The manager hid the candy before (her) six kids . . .
The sign was replaced after (her) black . . .
The message was clear after (her) blank . . .
Chris was very quick after (her) sharp . . .
The Perrys thought carefully after (her) wise advice . . .
The value went up after (her) rich neighbors . . .
People were offended after (her) rude . . .
The Smiths were shocked after (her) weird . . .
Deena doesn't have any leisure (or) time . . .
Anyone must be a minor (or) child . . .
Marty gave him a dollar (or) twenty last week . . .
George turned left at the river (or) bank . . .
Sally sold all her silver (or) jewelry last month . . .
Don must see the harbor (or) boats . . .
Fred would rather have a summer (or) lake . . .
Steve pitched the ball to center (or) left . . .
They promised him the future (or) aid . . .
Susan said those are (our) black socks . . .
Jake didn't vote for the member (or) constituent . . .
Jack reported trouble before (her) two children . . .
These documents are (our) fake . . .
Those tickets are (our) late entries . . .
These houses are (our) best . . .

Note: All fragments were grammatically acceptable beginnings of sentences. Parentheses indicate that the word was present in the Experiment 1 item but not in the Experiment 2 item, and square brackets indicate the reverse.
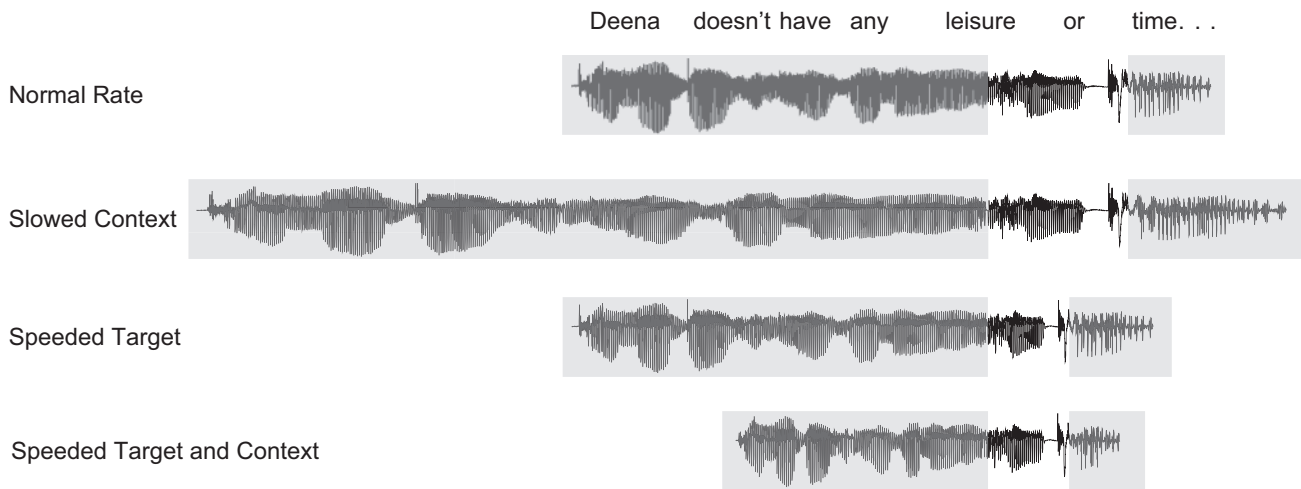
spoke naturally after adjusting to the task. An additional 70 sentences served as fillers intended to increase variety (e.g., length and structure). Visual presentation and audio recording of items were controlled by Presentation software (Version 12.1; Neurobehavioral Systems, Albany, CA).

We identified the 12 speakers who produced the fewest speech errors and disfluencies, as well as the fewest glottal onsets in critical function words, because continuous formant transitions across function words were desired; we used recordings from multiple talkers to increase the generalizability of the results across speakers. A single token recording was selected for each item using the following criteria: (a) The function words *are*, *or*, *our*, and *her* were spoken as [ɚ] and *a* was spoken as [ə], (b) the token recording showed continuous formant transitions across the critical function word plus the preceding syllable rhyme, and (c) the recording contained no hesitations or disfluencies.[1]

Fragments consisting of the grammatically acceptable beginnings of the sentences were taken from the token recordings and divided into target portion and context portion using spectrogram and waveform displays. The target corresponded to the critical function word plus the preceding syllable and the following phoneme (e.g., *-sure or t-*). The context corresponded to all speech preceding and following the target (e.g., *Deena doesn't have any lei- . . . -ime*).

Contiguous target and context regions were spliced out at zero crossings, where there was no energy in the waveform, and subjected to time manipulation using the Pitch-Synchronous Overlap and Add (PSOLA) algorithm in Praat software (Boersma & Weenink, 2002). Target and context regions were then recombined to create four conditions (Fig. 2); this method kept intact the spectral detail of the speech and altered only the speech rate. In the normal-rate condition, the entire fragment was presented at the spoken rate. In the slowed-context condition, the context was slowed through time expansion, and the target was presented at the spoken rate. (The target was thus acoustically identical to the target in the normal-rate condition). In the speeded-target condition, the target was speeded through time compression, and the context was presented at the spoken rate. In the speeded-target-plus-context condition, both the target and the context portions were speeded to the same degree through time compression. The time-compression factor was 0.6, and the time-expansion factor was 1.9. Each filler item was likewise speeded, slowed, or left unaltered in rate; the materials included approximately equal numbers of fillers at these three rates. All stimuli were then amplitude-normalized to 70 dB sound-pressure level (SPL).

Design and procedure. The experiment consisted of 120 trials, presented in a single session. Fifty trials contained experimental fragments of interest, and the remainder contained filler items. Participants were randomly assigned to one of four lists, with an approximately equal number of participants hearing each list. Each list contained 12 items in each of three

Deena   doesn't have   any   leisure   or   time. . .

Normal Rate

Slowed Context

Speeded Target

Speeded Target and Context

Fig. 2. Waveforms of a sample time-altered stimulus across the four conditions of Experiment 1. The sections of the waveform without background shading correspond to the target region, which consisted of the critical function word (*or*) plus the preceding syllable and the following phoneme (*-sure or t-*).

rate conditions and 14 items in the fourth rate condition. The pairing of items with conditions was counterbalanced across the four lists. Each participant heard each experimental fragment only once, in one of the four rate conditions.

The experiment began with 20 filler trials. The remaining trials were presented to all participants in the same random order. Participants were instructed to listen carefully to each fragment and to play it back as often as necessary to produce a veridical transcription of what they heard, typing their response using a computer keyboard. Stimuli were presented over studio-quality headphones at a comfortable listening level.

Results and discussion

The frequency of transcribing a function word in the target region was scored. Responses that did not include, at a minimum, a transcription of the target region plus the following syllable were discarded (6% of trials).[2] For remaining trials, function-word presence was coded as 1, and function-word absence was coded as 0.

Figure 3a shows that reports of the critical function word depended on the relative rate of the target and the context. In the normal-rate condition, function-word reports were quite high; the fact that they were not at ceiling is expected given that the speech was casually spoken. It is critical to note that a comparison of reports in the normal-rate condition and the slowed-context condition showed that merely slowing the context surrounding a function word caused the rate of function-word reports to drop by more than half, from 79% to 33%, even though the target regions containing the function word were acoustically identical. An equally dramatic reduction in function-word reports, relative to the normal-rate condition, was found in the speeded-target condition, in which the target region was speeded and the context was unaltered.

Function-word reports in the speeded-target-plus-context condition rebounded to close to their original levels in the normal-rate condition. That this mean did not reach the mean for the normal-rate condition is likely due to an overall drop in recognition accuracy associated with the significant compression factor (the compressed fragment was 60% of its original duration). A repeated measures one-way analysis of variance was significant by subjects, $F_1(3, 120) = 48.34$, $p < .001$, $\eta^2 = .55$, and by stimulus items, $F_2(3, 147) = 54.99$, $p < .001$. $\eta^2 = .53$. Post hoc two-tailed, paired-sample t tests with Bonferroni correction showed that all conditions differed significantly from one another in both by-subjects analyses and by-items analyses ($ps < .01$), except that the difference between the speeded-target condition and the slowed-context condition was not significant.

These results support the predictions of the generalized-rate-normalization account: Making the duration of a stretch of speech containing a function word fast relative to its context affected the number of morphophonological units perceived. This perceptual change was accomplished by either slowing down the context speech rate or speeding up the stretch of speech containing the target. That listeners could be induced to hear fewer morphophonological units implies that manipulating context speech rate induced listeners to hear fewer phonemes and fewer word boundaries than were actually spoken, a finding that has implications for word segmentation.

Experiment 2

A further test of the generalized-rate-normalization account is whether listeners can be made to hear more morphophonological units than were actually produced. This possibility was tested using fragments similar to those in Experiment 1, except for one minor (but crucial) change: The critical function word was never spoken. On the basis of the generalized-rate-normalization account, we predicted that if context events

a

b
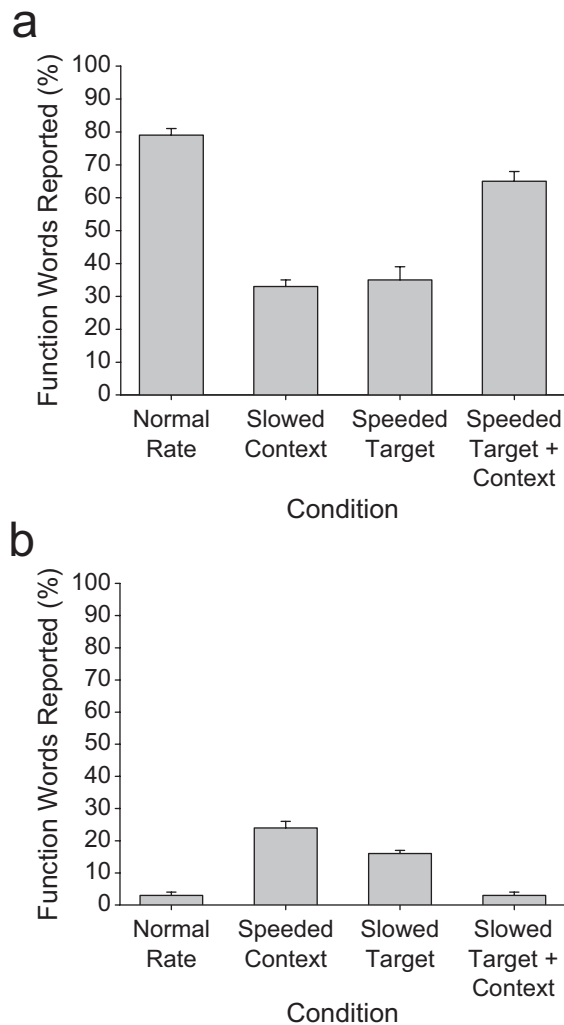
Fig. 3. Mean percentage of function words that participants reported hearing in (a) Experiment 1 and (b) Experiment 2, as a function of condition. In Experiment 1, the function words were spoken. In Experiment 2, the function words were not spoken. Error bars indicate standard errors of the mean.

were made fast relative to a stretch of speech that does not contain a function word—either by speeding up context speech rate or by slowing down the stretch of speech in question—then the stretch of speech would be perceived as relatively long and as containing more morphophonological units, even though those units were never spoken.

Method

Participants. Characteristics of the participants ($N = 69$) were the same as in Experiment 1.

Materials. We constructed sentences that had the same grammatical beginnings as in Experiment 1, but lacked the critical function word (e.g., *Deena doesn't have any leisure time*). Filler items were the same as in Experiment 1.

Recordings of experimental and filler sentences were obtained from 23 speakers using the elicitation task described in Experiment 1. From these recordings, we identified the 15 speakers who produced the fewest speech errors and disfluencies, and a single token recording was selected for each sentence.[3] Fragments consisting of the grammatically acceptable beginning of each sentence were spliced out of the recordings; each fragment was then divided into target and context portions. The target region was bounded by the same phoneme string as in Experiment 1, but the function word was not present (e.g., *-sure t-* in *Deena doesn't have any leisure time*). The context portion corresponded to all speech material preceding and following the target portion.

Four speech-rate conditions were created from the fragments. In the normal-rate condition, the entire fragment was presented at the spoken rate. In the speeded-context condition, the context was speeded, but the target was presented at the normal rate. In the slowed-target condition, the target was slowed, but the context was presented at the normal rate. Finally, in the slowed-target-plus-context condition, both the target and the context portions were slowed to the same degree. The time-compression factor was 0.6, and the time-expansion factor was 1.9. After alteration, target and context portions were concatenated in the proper order, and stimuli were amplitude-normalized to 70 dB SPL. The design and procedure were identical to those in Experiment 1.

Results and discussion

The frequency of transcribing a function word in the target region was scored. Responses that did not include, at a minimum, a transcription of the target region plus the following syllable were discarded from analysis (7% of trials).[4]

We obtained clear evidence that an alteration of speech rate can induce listeners to hear a function word (Fig. 3b). In the normal-rate condition (baseline), participants seldom (3% of the time) reported a function word in the target region—an expected finding because critical function words were never spoken. However, speeding the context surrounding the target caused an 8-fold increase in the rate of reporting a function word, even though the target was identical in the two conditions. Slowing the target resulted in a similar, 5-fold increase in the rate of hearing a function word that was never spoken. As in Experiment 1, when the context and the target were time-altered together (in this case, slowed), reports of the function word returned to the level found in the normal-rate condition. A repeated measures one-way analysis of variance with rate condition as the factor was significant by subjects, $F_1(3, 204) = 60.82$, $p < .001$, $\eta^2 = .47$, and by stimulus items, $F_2(3, 147) = 25.50$, $p < .001$, $\eta^2 = .34$. Post hoc tests showed that all conditions differed significantly from one another in both by-subjects analyses and by-items analyses ($ps < .01$), except for the normal-rate condition compared with the slowed-target-plus-context condition.

These results provide strong evidence for the generalized-rate-normalization hypothesis that context speech rate affects whether listeners perceive a word. In this experiment, listeners were made to perceive a function word that was never spoken. By implication, context speech rate affected the number of phonemes and word boundaries that listeners perceived. Thus, Experiment 2 replicated and extended the findings of Experiment 1.

General Discussion

The current experiments provide new insight into how timing information is used in speech perception. In Experiment 1, sentence fragments containing a critical function word were heard as having fewer such words when context speech rate was slowed. In Experiment 2, matched sentences in which the critical function word was never spoken were heard as containing function words when context speech rate was speeded. These experiments were based on the generalized-rate-normalization hypothesis, according to which the number of perceived morphophonological units in a stretch of speech depends on the duration of that stretch relative to the speech rate of the context in which it is embedded. These experiments indicate that listeners used context speech rate to help them decode spectrally ambiguous portions of the speech stream, and this process aided listeners in perceiving spoken words and in segmenting those words from the speech context.[5]

These studies are the first to show that context speech rate can modulate whether an entire word is perceived. The duration of a stretch of speech relative to its context speech rate also modulated the number of phonemes and implied word onsets perceived as present. These findings have implications for the important and unsolved problem of how infants and adults identify word onsets in connected speech (Cutler, Mehler, Norris, & Segui, 1983; Mattys, White, & Melhorn, 2005; Thiessen, Hill, & Saffran, 2005). Note that our speech-rate manipulations were several phonemes distant from the variably perceived function word; this contrasts with previous findings that speech-rate manipulations in the immediate vicinity of a to-be-perceived phoneme had an effect, but more distant manipulations typically had no effect (e.g., Sawusch & Newman, 2000).

These findings suggest that information on relative speech rate aids in interpreting ambiguous spectral cues and helps listeners identify and segment spoken words. How words and word boundaries are so robustly perceived when spectral cues are unclear remains poorly understood (Ernestus, Baayen, & Schreuder, 2002; Pitt, 2009). Our experiments suggest that word recognition depends in part on relative rate cues provided by speech context, and this study adds to a growing body of work showing that prosodic properties of speech context influence lexical recognition and word segmentation (Dilley, Mattys, & Vinke, 2010; Dilley & McAuley, 2008; Gout, Christophe, & Morgan, 2004; Salverda et al., 2003).

More generally, the results demonstrate the rapid and seamless integration of signal-based cues (spectral, temporal) and knowledge-based cues (syntactic, semantic) during spoken-word recognition. In this regard, our speech-rate phenomenon, particularly the results of Experiment 2, can be viewed as a temporal version of phonemic restoration, in which listeners readily restore phonemes in words whose acoustic evidence has been replaced by noise (Samuel, 2001). In phonemic restoration, sentential context biases perception, and such higher-level biases are likely at work in the effect we observed; they may be a precondition for the effect, for example.

Compared with studying spectral cues, studying how timing information is used in speech perception has proven challenging. The present results provide one answer to the puzzle of how reduced, spectrally attenuated syllables and words are recognized and segmented from continuous speech; rate normalization via temporal entrainment to speech rate provides a possible explanation for these findings. In the absence of clear spectral information, timing information becomes increasingly important in conveying the message intended by the speaker.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. To select similar speaking rates across talkers, we first determined the grand mean duration of the two syllables preceding the critical function word across all sentences produced by these speakers. After other selection criteria had been applied, a token recording of a given sentence was selected from the talker who produced the two syllables preceding the function word with a duration that was minimally different from the grand mean duration. This resulted in different talkers providing different numbers of experimental items ($M = 4.2$, range = 0–13, with 11 talkers represented in the final experimental stimulus set).

2. Results were identical when these responses were included.

3. The selection procedure for controlling speaking rate across sentences was identical to the procedure used in Experiment 1. This procedure resulted in different talkers providing different numbers of experimental items ($M = 3.3$, range = 0–8, with 13 talkers represented in the final set).

4. Results were identical when these responses were included.

5. A competing hypothesis that speech-rate mismatches per se were responsible for effects on function-word perception is untenable, for

several reasons. According to one version of this hypothesis, function words were not perceived when the rate across stimuli mismatched. However, in Experiment 2, more, not fewer, function words were perceived when the rate mismatched than when it matched. Thus, this hypothesis cannot account for the data across both experiments.

The results also do not support a weaker version of the hypothesis, that speech-rate mismatches cause reductions in general intelligibility associated with different illusory lexical percepts, depending on the veridical grammatical properties of fragments. To test this latter version of the hypothesis, we conducted additional analyses of transcription accuracy of phonemes in words preceding the critical function word in matching conditions (Experiment 1: normal rate, speeded target plus context; Experiment 2: normal rate, slowed target plus context) and mismatching conditions (Experiment 1: slowed context, speeded target; Experiment 2: speeded context, slowed target). Results revealed no difference in transcription accuracy between the matching and mismatching conditions in Experiment 1 (M_{match} = 94%, $M_{mismatch}$ = 95%), paired-samples $t(49)$ = 1.13, p = 0.27. Although a significant difference was found in Experiment 2 (M_{match} = 97%, $M_{mismatch}$ = 93%), paired-samples $t(49)$ = 6.21, $p < .001$, the size of the change (4%) was much smaller than the rise in rates of function-word reports (13%–21%). This result further suggests that such a rate effect cannot account for differences in function-word perception.

References

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

Boersma, P., & Weenink, D. (2002). Praat: Doing phonetics by computer (Version 4.0.26) [Computer software]. Retrieved August 6, 2010, from http://www.praat.org

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language-specific comprehension strategy. *Nature*, *304*, 159–160.

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

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

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

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

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M.A.A. Tatham (Ed.), *Auditory analysis and perception of speech* (pp. 197–219). London, England: Academic Press.

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

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

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

Liberman, A.M., Delattre, P., Gerstman, L., & Cooper, F.S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, *52*, 127–137.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions during word recognition in continuous speech. *Cognition*, *10*, 487–509.

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

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

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

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

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

Pitt, M.A. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language*, *61*, 19–36.

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

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

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

Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, *19*, 499–526.

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

Samuel, A.G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*, 348–351.

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

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

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

Smith, Z.M., Delgutte, B., & Oxenham, A. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87–90.

Thiessen, E.D., Hill, E.A., & Saffran, J.R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*, 53–71.