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...
in the phrase *leisure or time*. Thus, there is a relatively homogeneous span of spectral material for most of the two syllables: -(sure 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 phrase; 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

<table>
<thead>
<tr>
<th>Fragment</th>
<th>Experiments 1</th>
<th>Experiments 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conor knew that bread and butter (are) both . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Frank thinks that sadness and anger (are) both . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Claire said that sour and bitter (are) both . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Zach knew that there (are) things . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>George thought my father and brother (are) like [good] . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Glenn thought his friend and neighbor (are) like plenty . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Ruth saw the maid and butler (are) at the top . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Rose knew that there (are) lamps . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>The company moved to (a) different . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Trent might get to (a) certain . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>Clay thinks that would be (a) good . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
<tr>
<td>The Smiths wouldn’t buy (a) Butterball . . .</td>
<td>. . .</td>
<td>. . .</td>
</tr>
</tbody>
</table>

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.
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). 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
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, 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 ($p < .01$), 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.

Acknowledgments

We thank Delphine Dahan, Sven Mattys, Arthur Samuel, and an anonymous reviewer for useful feedback on the manuscript. Also, we thank Victoria Hoover, Michael Tat, Andrea Hulme, Chris Heffner, and Claire Carpenter for help with data acquisition and data analysis.

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.

Funding

This work was supported by Grant DC004330 from the National Institute on Deafness and Other Communication Disorders to M.A.P. and by Grant 0847653 from the National Science Foundation Division of Behavioral and Cognitive Sciences to L.C.D.

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...
s 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 ($M_{\text{match}} = 94\%$, $M_{\text{mismatch}} = 95\%$), paired-samples $t(49) = 1.13, p = 0.27$. Although a significant difference was found in Experiment 2 ($M_{\text{match}} = 97\%$, $M_{\text{mismatch}} = 93\%$), 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


