Psychological Science OnlineFirst, published on June 6, 2014 as doi:10.1177/0956797614533705

Research Article

Psychological Science
1–8

DOI: 10.1177/0956797614533705
pss.sagepub.com

Long-Term Temporal Tracking of Speech Rate Affects Spoken-Word Recognition

Melissa M. Baese-Berk[1], Christopher C. Heffner[2], Laura C. Dilley[1], Mark A. Pitt[3], Tuuli H. Morrill[1], and J. Devin McAuley[4]

[1]Department of Communicative Sciences and Disorders, Michigan State University; [2]Department of Hearing and Speech Sciences, University of Maryland, College Park; [3]Department of Psychology, Ohio State University; and [4]Department of Psychology, Michigan State University

Abstract

Humans unconsciously track a wide array of distributional characteristics in their sensory environment. Recent research in spoken-language processing has demonstrated that the speech rate surrounding a target region within an utterance influences which words, and how many words, listeners hear later in that utterance. On the basis of hypotheses that listeners track timing information in speech over long timescales, we investigated the possibility that the perception of words is sensitive to speech rate over such a timescale (e.g., an extended conversation). Results demonstrated that listeners tracked variation in the overall pace of speech over an extended duration (analogous to that of a conversation that listeners might have outside the lab) and that this *global* speech rate influenced which words listeners reported hearing. The effects of speech rate became stronger over time. Our findings are consistent with the hypothesis that neural entrainment by speech occurs on multiple timescales, some lasting more than an hour.

Keywords

speech perception, word segmentation, entrainment, speech rate

Received 10/19/13; Revision accepted 4/8/14

A long-standing principle in neuroscience is that the nervous systems of humans and other animals adapt to the statistical (e.g., distributional) properties of sensory stimulation in the environment (Barlow, 1961). This principle is reflected at all levels of the nervous system, from single neurons to networks and behavior (e.g., Yang & Shadlen, 2007), and has been identified as essential to the survival of organisms (Geisler & Diehl, 2002). A growing body of evidence demonstrates that humans detect statistical regularities in a variety of modalities, including vision (Fiser & Aslin, 2002), hearing (speech stimuli—Saffran, Aslin, & Newport, 1996; nonspeech stimuli—Saffran, Johnson, Aslin, & Newport, 1999), and touch (Conway & Christiansen, 2005).

Distributional characteristics have been shown to be important for acquiring and understanding spoken language. Saffran et al. (1996, 1999) demonstrated that infants and adults use transitional probabilities between syllables to determine the boundaries between spoken words, a process commonly described as statistical learning. Statistical learning has also been shown to influence the development of phonological categories, as well as more fine-grained levels of phonetic perception (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Maye, Weiss, & Aslin, 2008).

One unresolved question is the extent to which listeners track the distributional characteristics of temporal information in speech. In previous studies of timing in speech, researchers have most frequently investigated short timescales. Temporal cues are known to influence phoneme perception (Liberman, Delattre, Gerstman, & Cooper, 1956). Furthermore, several studies have demonstrated that changes in speech rate influence phoneme identification (Miller, 1981; Miller, Aibel, & Green, 1984;

Corresponding Author:
Laura C. Dilley, Department of Communicative Sciences and Disorders, Oyer Center, 1026 Red Cedar Rd., Michigan State University, East Lansing, MI 48824
E-mail: ldilley@msu.edu

Summerfield, 1981). Examining speech intelligibility more broadly, Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) found that spoken-word recognition is remarkably robust even when spectral frequency information in speech is severely degraded (but temporal information is spared). Work by Greenberg and his colleagues (Greenberg & Arai, 2001; Silipo, Greenberg, & Arai, 1999) suggests that frequency information contained in the amplitude-modulation spectrum is critical for intelligibility (for a discussion of the relative importance of timing and spectral information in perception, see Elliott & Theunissen, 2009). Although some research has shown that listeners' perceptions of spoken words can change over time with exposure to time-compressed speech (e.g., Dupoux & Green, 1997), the timescales over which timing information is relevant to speech perception have not been specified.

Listeners' perception of spoken words is influenced by the timing information within the short-term speech context (i.e., about 300 ms to 3 s, or the approximate length of a phrase; see Dilley & Pitt, 2010). Grammatical, or "function," words, such as *or, a, not*, and *and*, are very important for the meaning of a sentence; however, such words have highly variable acoustic realizations and can blend fully with surrounding words, which makes them difficult to distinguish on the basis of spectral information (Shockey, 2003). For example, the word *or* can often be spoken as "errr"; thus, the phrase *leisure or time*, for example, can be confused with *leisure time*. In our prior work (Dilley & Pitt, 2010), we constructed grammatical sentences with and without such function words (e.g., *Deena didn't have any leisure or time* and *Deena didn't have any leisure time*) and presented listeners with these sentences. For the sentences with the function words, we selected recorded stimuli in which the function word was acoustically blended with its context. The sentences with the function words were presented both with and without the short-term context around the function word (i.e., the *distal context*, about 300 ms to 3 s) digitally slowed relative to the rest of the sentence. The acoustic information in the function words and their distal context were held constant across these versions. Listeners in the slowed-distal-context condition reported hearing the function words significantly less often than did listeners in the normal-rate condition, even though the words themselves and their immediate context were acoustically identical.

To explain the effect of distal temporal information on speech perception, we proposed an extension of entrainment accounts in basic auditory perception (Dilley & Pitt, 2010). According to this proposal, the context speech rate generates expectations in a listener about the tempo of events within a given stretch of acoustic material, and these expectations lead to perception of different numbers of syllables or words. Our proposal is consistent with the idea that neural oscillations, which are pervasive in the brain (Schroeder & Lakatos, 2009), are entrained by speech input (Ahissar et al., 2001; Luo & Poeppel, 2007). In their recent review of the literature on entrainment of neural oscillations in humans, Peelle and Davis (2012) highlighted our prior work (Dilley & Pitt, 2010) as an example of how distal speech rate influences perception. They suggested that neural entrainment may drive this effect, which in turn influences the recognition of later words in an utterance.

In the nonspeech auditory domain, results from several studies support the notion of stimulus-driven entrainment over both short and long timescales (Barnes & Jones, 2000; Jones & McAuley, 2005; Large & Jones, 1999; McAuley & Jones, 2003). Jones and McAuley (2005) found distortions in perceived duration that emerged over the course of the experimental session; these results are consistent with the idea that listeners tracked the average pace of the auditory stimuli they experienced (see also McAuley & Miller, 2007). In the present study, we investigated whether similar systematic shifts might be observed in speech perception as a function of the pace of speech in a listener's environment. Specifically, given our previous findings (Dilley & Pitt, 2010), we were interested in whether the speech rate over the course of an hour-long experimental session (the *global-context speech rate*) would have an influence on word recognition above and beyond the influence of the immediate (distal) temporal context.

To address this question, we manipulated both global-context speech rate (i.e., the experiment-wide speech rate) and distal-context speech rate (i.e., the speech rate within an utterance) and examined how this rate information influenced perception of words over time. If, as we expected, listeners used the global-context speech rate in addition to the distal-context speech rate in identifying spoken words, speech perception would change as a function of the global context rate. Such findings would suggest that learning distributional (i.e., statistical) properties of the global speech rate is important for word recognition. We also expected that the effect of global speech rate should increase over time, which would also indicate learning of the distribution of the global pace of speech. Such long-term changes in speech perception would demonstrate a temporal sensitivity in spoken language not previously realized. Furthermore, such changes would be consistent with an entrainment hypothesis, suggesting that the perceptual system entrains both locally and globally.

Method

Participants

Participants (n = 55) were adult native speakers of American English (mean age = 20.7 years). Following procedures in our earlier study (Dilley & Pitt, 2010),

8 participants were excluded from analysis because of their low accuracy in transcribing nontarget portions of the utterances. We recruited 15 participants per condition, because this is the number of participants that has previously showed the distal-speech-rate effect (Dilley & Pitt, 2010; Heffner, Dilley, McAuley, & Pitt, 2013). All participants reported that they had normal hearing.

Materials

Experimental materials were chosen from among those we elicited for the prior study (Dilley & Pitt, 2010). Each target utterance had a spectrally ambiguous portion that included a reduced critical word (*or, a, her, our*, or *are*), which resulted in an ambiguous number of words in the utterance. All utterances could lead to a grammatical percept with or without the critical word. Filler utterances that had no ambiguity with regard to the number of words were also included.

Multiple versions of each experimental utterance were created by modifying the speech rate of the distal-context portion (i.e., the portion of the utterance that was more than one syllable away from the spectrally ambiguous potential word boundary), following our prior stimulus design (Dilley & Pitt, 2010). By defining the bounds of the target region as including one syllable before and after the spectrally ambiguous word boundaries, we controlled coarticulatory and timing cues immediately adjacent to the ambiguous word boundary, and thus these cues could not be responsible for any observed effects of the speech-rate manipulation on segmentation or word recognition.

We created several distal-speech-rate conditions in which the duration (and thus the speech rate) of the context portion of the utterance was the same or was expanded. We used the Pitch-Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2014) to process all of the segments in the entire distal-context portion of each utterance. In one condition (1.0), the duration of the distal-context portion was not changed. For the other conditions (1.2, 1.4, 1.6, and 1.8), we altered the duration of the segments in the entire distal-context portion. For example, in the 1.4 distal-speech-rate condition, the duration of the distal-context portion was multiplied by a factor of 1.4, which resulted in a duration that was 140% of the utterance's original duration. Each phoneme within the modified portion of the utterance was expanded by a factor of 1.4. In our prior work (Dilley & Pitt, 2010), we used both context-expanded and context-compressed stimuli; however, because the effect was most robust when the context was expanded, we chose to examine only context-expanded stimuli in the present study.

Although the distal speech rate thus varied across the different versions of each experimental utterance, the proximal acoustic information (i.e., the acoustic information in the syllables immediately surrounding the spectrally ambiguous word boundary, including the rate of speech) was identical in all versions of the utterance. The speech rate was not normalized to a single rate across utterances before these manipulations were performed. Therefore, tokens assigned a particular duration multiplier were presented at a range of spoken rates, corresponding to a distribution of speech rates across the tokens. The speech rates of stimuli in the present study were well approximated by a normal distribution (M = 5.64 syllables per second, SD = 1.07). The duration modification simply shifted the mean and scaled the standard deviation by the factor in the duration multiplier.

Filler stimuli were also assigned a duration multiplier, which changed the speech rate of these utterances in their entirety. These modifications were performed in Praat software (Boersma & Weenink, 2014).

Design

Participants were assigned to one of three global-speech-rate conditions, which were defined by the statistical distribution (i.e., the mean) of the duration multipliers applied to the distal speech context in the utterances they heard: 1.2 (n = 16), 1.4 (n = 15), or 1.6 (n = 16). The number of participants per group was uneven because of the excluded participants. Within each of these global-speech-rate conditions, participants listened to items at three distal speech rates: one that was faster than the mean distal rate, one that was slower than the mean distal rate, and one that was equal to the mean distal rate. Table 1 shows the three distal-speech-rate conditions in each global-speech-rate condition. For example, in the 1.2 global-speech-rate condition, participants were exposed to distal speech rates produced by multipliers of 1.0 (faster than the mean rate), 1.2 (the mean rate), and 1.4 (slower than the mean rate).

Procedure

Participants listened to the utterances and were asked to type what they heard. During the course of the experiment, each participant heard 189 utterances, 126 filler stimuli and 63 experimental stimuli. Assignment of utterances to distal speech rates was counterbalanced across items and participants so that in each global-speech-rate condition, each utterance was presented to the same number of participants at each of the three distal speech rates. Items were divided into three equal blocks (i.e., 63 utterances per block). The experiment was self-paced and lasted approximately 1 hr.

Analysis

We analyzed participants' transcriptions to determine whether they reported hearing the critical function word

Table 1. Duration Multipliers Used to Create the Faster, Average, and Slower Relative Distal Speech Rates in Each of the Global-Speech-Rate Conditions

Global-speech-rate condition	Duration multiplier				
	1.0	1.2	1.4	1.6	1.8
1.2	Faster	Average	Slower	—	—
1.4	—	Faster	Average	Slower	—
1.6	—	—	Faster	Average	Slower

in the spectrally ambiguous portion of each utterance. Responses that reflected inattention to the phonemic properties of the speech around the function word (approximately 5%) were not included in the analysis, as in our prior work (Dilley & Pitt, 2010). Responses were analyzed using mixed-effects logistic regression and model comparison to identify the random and fixed factors that best fit the data. The resulting model included participants and items as random factors. Global speech-rate condition (1.2, 1.4, or 1.6), relative distal speech rate (slower, average, and faster; see Table 1), and block (1, 2, or 3) were included in the model as fixed factors. Each factor was Helmert coded so that each level could be compared with the subsequent levels and multiple comparisons among factor levels could be conducted without loss of power (Barr, Levy, Scheepers, & Tily, 2013). Across the global-speech-rate conditions, each condition was compared with the slower condition or conditions; within each global-speech-rate condition, each level of relative distal speech rate was compared with the slower level or levels, and each experimental block was compared with previous blocks.

Results

Figure 1 shows the percentage of critical function words that were reported in each of the three global-speech-rate conditions at each of the three relative distal speech rates, collapsed across blocks. As in our previous work (Dilley & Pitt, 2010), listeners reported hearing the function word less often at slower rates of distal speech. Furthermore, it is clear that listeners reported hearing the function word less often as the global speech rate decreased (i.e., as the duration multiplier increased), even when the distal speech rate (i.e., the acoustic token) was identical.

The logistic mixed-effects regression model revealed that performance differed significantly across the three relative distal speech rates (i.e., slower, average, and faster; faster vs. slower rate: $\beta = 1.32$, $z = 12.25$, $p < .0005$; average vs. slower rate: $\beta = 0.93$, $z = 7.79$, $p < .0005$). The logistic mixed-effects regression model also showed that

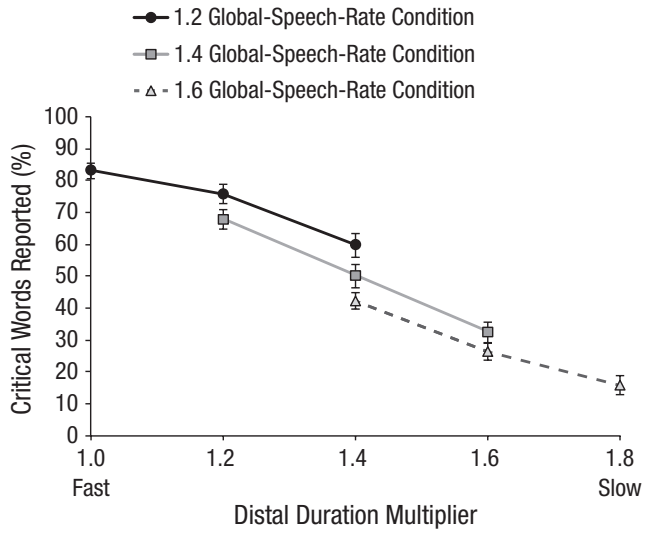

Fig. 1. Percentage of critical words that were reported as a function of distal-speech-rate condition for each global-speech-rate condition, collapsed across blocks. Error bars indicate ±1 *SE*.

performance differed significantly across the global-speech-rate conditions (1.2 vs. 1.4 and 1.6: $\beta = 1.73$, $z = 8.28$, $p < .0005$; 1.4 vs. 1.6: $\beta = 0.91$, $z = 3.60$, $p < .0005$).

Thus, the long-term statistical properties of the global speech rate influenced listeners' speech perception, as indexed by the rate at which the critical words were reported across the global-speech-rate conditions. In other words, speech perception changed as a function of the global-speech-rate context experienced over the experimental session.

We also examined whether there was evidence of increased tracking of the distribution of the global pace of speech over the length of the experimental session (Fig. 2), which an entrainment account would predict. In the first block, listeners demonstrated a distal-speech-rate effect but not a global-speech-rate effect. The global-speech-rate effect emerged in Blocks 2 and 3 (i.e., the lines for the different global-speech-rate groups become more separated in the center and right panels of Fig. 2). Furthermore, block did not appear to influence the effect of relative distal rate.

These patterns were borne out in the model analysis. Adding block to the regression model significantly improved the fit of the model according to a log-likelihood test, $\chi^2(29) = 52.31$, $p < .0005$. Although the overall effect of block was not significant ($z < 1$), there were three significant interactions between global speech rate and block—1.2 (vs. 1.4 and 1.6) × Block 1 (vs. Blocks 2 and 3): $\beta = 0.95$, $z = 3.56$, $p < .003$; 1.4 (vs. 1.6) × Block 1 (vs. Blocks 2 and 3): $\beta = 0.68$, $z = 2.30$, $p < .02$; and 1.2 (vs. 1.4 and 1.6) × Block 2 (vs. Block 3): $\beta = 0.60$, $z = 2.29$, $p < .02$. The remaining interaction, 1.4 (vs. 1.6) × Block 2

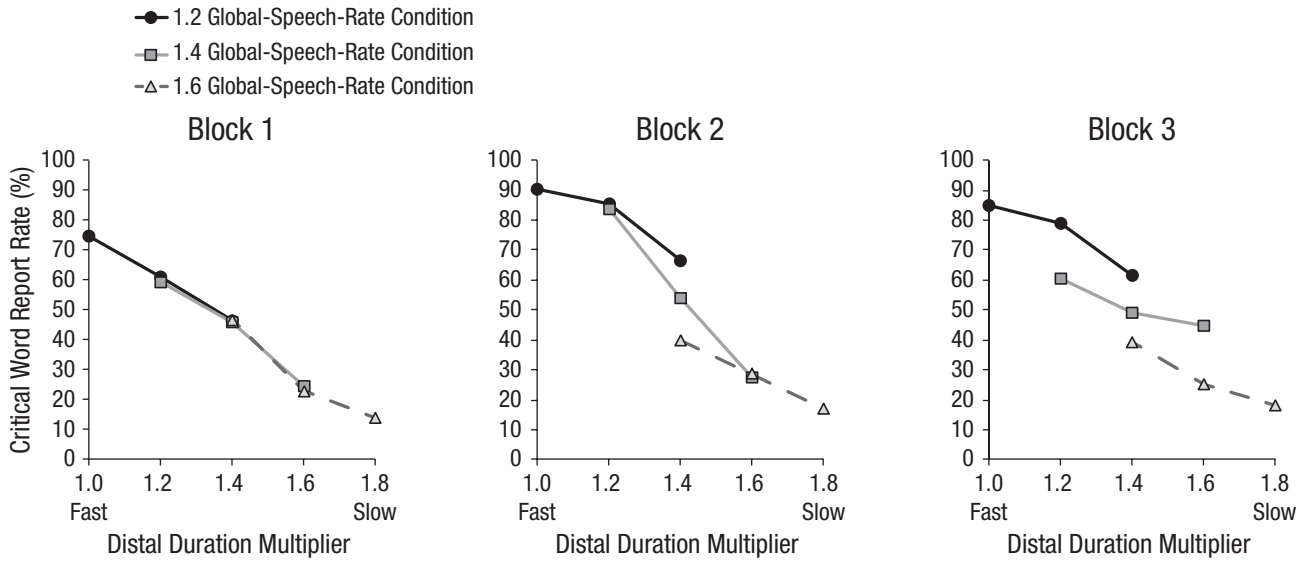

Fig. 2. Percentage of critical words reported as a function of the distal duration multiplier for each of the global-speech-rate conditions. Results from each block are presented separately.

(vs. Block 3) was marginally significant, $\beta = 0.52$, $z = 1.73$, $p < .08$.

These interactions reveal increased learning of distributions of temporal information in speech over time, as evidenced by the fact that listeners' speech perception was more influenced by the statistics of global-context speech rate in later blocks (i.e., after more exposure) than in earlier blocks. Adding the interaction between block and relative distal rate did not improve model fit significantly ($\chi^2 < 1$). This suggests that the effect of global speech rate, as a function of block, is above and beyond the effect of relative distal speech rate first demonstrated in our previous study (Dilley & Pitt, 2010). Taken together, these results demonstrate that statistical learning of long-term temporal properties of speech influences word perception and that this learning increases over time.

General Discussion

The present investigation is the first to examine whether listeners track speech rate over long timescales, such as those associated with extended conversations, and whether such information leads to differences in word perception for the same acoustic material. We took advantage of recent work showing that the relative distal speech rate influences how many words listeners hear in acoustically identical segments of speech. We investigated whether spoken-word recognition is influenced not just by the relative distal speech rate (i.e., within an utterance), but also the global (average) speech rate that listeners experience over the course of an experiment. There were two main findings: Listeners' sensitivity to the global speech rate (i.e., the distribution of distal speech rates) influenced how frequently key spoken words were reported, and these effects became stronger over the course of the experiment (i.e., about an hour).

In the present study, we examined the distal-speech-rate effect specifically with slowed context speech rates. In our prior work (Dilley & Pitt, 2010), we demonstrated distal-speech-rate effects for both slowed and speeded contexts. Although we chose to focus the present examination on slowed speech rates, we expect that the influence of global speech rates would extend to speeded speech rates as well. The bulk of our results thus far demonstrate that the perception of function words is influenced by the distal speech rate. However, other research from our lab suggests that the distal-speech-rate effect is also present in the case of reduced syllables in content words (Baese-Berk et al., 2013). That is, the perception of words such as *tear* and *terror* is similarly influenced by distal speech rate. We anticipate that the effect of global speech rate would also extend to reduced syllables in content words.

Although it is clear that statistical learning of spoken language is a behavioral result of auditory exposure in adults and infants, the mechanisms underlying statistical learning are not well-specified (Romberg & Saffran, 2010). The present results are consistent with the view that neural entrainment is the underlying mechanism in some forms of human statistical learning and sensitivity to distributional patterns in the environment. Because

neural oscillations underlie entrainment (Peelle & Davis, 2012), these results are consistent with a hypothesis that oscillations support tracking the pace of speech on multiple timescales. The observed effects of global speech rate on spoken-word recognition are in line with results from a wide array of other work on adaptation and perceptual learning in speech.

This body of work has demonstrated that the distribution of a variety of phonetic properties over the course of an experiment can influence perception of fine-grained phonetic detail (Norris, McQueen, & Cutler, 2003) and that this effect lasts for a significant period of time (Kraljic & Samuel, 2005). For example, exposure to ambiguous phonetic tokens in specific contexts results in shifts in perception of a phonetic continuum. Furthermore, this perceptual restructuring for speech is influenced by global knowledge about the speaker (Kraljic, Brennan, & Samuel, 2008), and information in the environment that is completely unrelated to the speech stimuli (e.g., the presence of a stuffed toy in the room) has also been shown to influence speech perception (Hay & Drager, 2010). Taken together with previous research, our current results demonstrate that many types of global information, including segmental, prosodic, and nonlinguistic information, can influence perception of local events in speech. This suggests that models of speech perception and word recognition must be updated to account for the influence of global information in the acoustic and physical environment.

Researchers have investigated perceptual adaptation to speech and have demonstrated that exposure to time-compressed speech yields better comprehension of such speech (Adank & Devlin, 2010; Dupoux & Green, 1997). The present findings are the first to show that learning of time-altered speech engenders temporal expectancies over long timescales in a manner that captures the statistics of global speech rate. Unlike previous studies of perceptual learning of time-compressed speech, in which exposure has led to monotonic increases in comprehension accuracy, our study shows that the distribution of speech rates in the input can predict word perception. The speech stimuli used here always contained a critical function word that was not time compressed or acoustically altered in any way; this word either was or was not perceived by the listener, and a higher rate of reporting the critical word corresponded to better accuracy. Thus, the present results are the first to reveal that increased accuracy in speech perception is, in part, a direct function of the statistical properties of the speech rates to which a listener is exposed.

These findings are consistent with proposals that entrainment of neural oscillations by speech input is crucial for word recognition (Giraud & Poeppel, 2012; see Peelle & Davis, 2012, for a review). Prior studies of neural correlates of human speech perception have examined neural responses across nonidentical acoustic stretches of speech (Adank & Devlin, 2010; Ahissar et al., 2001), an approach that does not allow for direct comparisons of the influence of context speech rate. However, the behavioral paradigm used here, in which word recognition was examined for acoustically identical stretches of speech, has the potential to permit examination of neural correlates of human speech perception with much more specificity than has been possible in previous research. Given the present results, we would expect to see distinct patterns of temporal dynamics of neural phase locking in the different global-speech-rate conditions used here. If the neurophysiological response to identical tokens changed as a function of the context in which the tokens were heard, this would be strong evidence that global-context speech rate influences perception at a neural level.

Conclusions

The present study is the first to demonstrate that listeners track the pace of speech over an extended duration analogous to that of a conversation that listeners might have outside the lab. Moreover, the fact that the observed effect of global speech rate on what words listeners report hearing becomes stronger over time is consistent with the behavioral effects of statistical learning. These findings, together with those in our prior work (Dilley & Pitt, 2010), show that both local and more global aspects of speech timing combine to influence listeners' perceptions of the number of units (e.g., words, syllables, segments) in a given stretch of acoustically identical speech. These results suggest an important new factor to be taken into account in theories of spoken-word recognition: the interaction between the local timing information of an utterance and the more extended temporal context in which the utterance is embedded. The contextual effects of speech rate reported here are consistent with recent proposals that neural oscillations are entrained by speech rate in a manner that is critical for successful spoken-language processing (Ahissar et al., 2001; Giraud & Poeppel, 2012; Peelle & Davis, 2012). In this view, the present findings suggest that individuals track temporal information in speech on both shorter timescales (on the order of seconds) and longer ones (on the order of minutes or hours), and that these effects are consistent with proposals of neural entrainment.

Author Contributions

C. C. Heffner, L. C. Dilley, T. H. Morrill, and J. D. McAuley contributed to the study design. C. C. Heffner performed the testing and collected data. M. M. Baese-Berk, T. H. Morrill, and J. D. McAuley analyzed and interpreted the data. M. M. Baese-Berk,

L. C. Dilley, and J. D. McAuley drafted the manuscript, and M. A. Pitt provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

Portions of the data presented here were first presented at Perspectives on Rhythm and Timing, July 2012, at the University of Glasgow.

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 National Science Foundation Faculty Early Career Development (CAREER) Program Grant BCS 0874653 (to L. C. Dilley) and by a Provost Undergraduate Research Initiative award from the Michigan State University College of Social Science (to J. D. McAuley and C. C. Heffner).

References

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*, 1124–1132. doi:10.1016/j.neuroimage.2009.07.032

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences, USA*, *98*, 13367–13372. doi:10.1073/pnas.201400998

Baese-Berk, M. M., Dilley, L. C., Henry, M., Vinke, L., Banzina, E., & Pitt, M. A. (2013). Distal speech rate influences lexical access [Abstract]. *Abstracts of the Psychonomic Society*, *18*, 191.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, *41*, 254–311. doi:10.1006/cogp.2000.0738

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.3.70) [Computer software]. Retrieved from http://www.praat.org/

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*, 804–809. doi:10.1016/j.cognition.2008.04.004

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: General*, *31*, 24–39. doi:10.1037/0278-7393.31.1.24

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

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 914–927.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), Article e1000302. Retrieved from http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000302

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458–467. doi:10.1037/0278-7393.28.3.458

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*, 419–448. doi:10.1098/rstb.2001.1055

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*, 511–517. doi:10.1038/nn.3063

Greenberg, S., & Arai, T. (2001). The relation between speech intelligibility and the complex modulation spectrum. In P. Dalsgaard, B. Lindberg, H. Benner, & Z. Tan (Eds.), *EUROSPEECH 2001 Scandinavia, Proceedings of the 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event* (pp. 473–476). Aalborg, Denmark: Aalborg University Press.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*, 865–892. doi:10.1515/ling.2010.027

Heffner, 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*, 1275–1302.

Jones, M. R., & McAuley, J. D. (2005). Time judgments in global temporal contexts. *Attention, Perception, & Psychophysics*, *67*, 398–417.

Kraljic, T., Brennan, S., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*, 54–81.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

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

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

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*, 1001–1010. doi:10.1016/j.neuron.2007.06.004

Maye, J., Weiss, D., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*, 122–134.

McAuley, J. D., & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 1102–1125. doi:10.1037/0096-1523.29.6.1102

McAuley, J. D., & Miller, N. S. (2007). Picking up the pace: Effects of global temporal context on sensitivity to the tempo of auditory sequences. *Attention, Perception, & Psychophysics*, *69*, 709–718.

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

Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception & Psychophysics*, *35*, 5–15. doi:10.3758/BF03205919

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, Article 320. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fpsyg.2012.00320/full

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906–914. doi:10.1002/wcs.78

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*, 9–18.

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*. Hoboken, NJ: Wiley-Blackwell.

Silipo, R., Greenberg, S., & Arai, T. (1999). Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations. In G. Olaszy, G. Németh, & K. Erdőhegyi (Eds.), *Proceedings of the 6th European Conference on Speech Communication and Technology* (pp. 2687–2690). Budapest, Hungary: Budapest University of Technology and Economics.

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

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*, 1075–1080.