Real-time expectations based on context speech rate can cause words to appear or disappear

Meredith Brown (mbrown@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268, Rochester, NY 14627-0268

Laura C. Dilley (ldilley@msu.edu)
Department of Communicative Sciences & Disorders, Michigan State University
116 Oyer, East Lansing, MI 48824

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268, Rochester, NY 14627-0268

Abstract

To test predictions of a forward modeling framework for spoken language processing, we characterized effects of context speech rate on the real-time interpretation of indefinite noun phrases using the visual world paradigm. The speech rate of sentence material distal to the onset of the noun phrase was manipulated such that the segments surrounding the determiner *a* in singular noun phrases had a faster speech rate than the surrounding context and the segments surrounding the onset of plural noun phrases had a relatively slow rate. These manipulations caused listeners to fail to perceive acoustically present determiners and to falsely perceive determiners not present in the signal. Crucially, fixations to singular and plural target pictures revealed effects of distal speech rate during the real-time processing of target expressions, strongly suggesting a locus in perceptual expectations. These results set the stage for quantitative tests of forward models of spoken language processing.

Keywords: Speech rate; prosody; expectations; speech processing; visual world paradigm

Introduction

Expectation-based approaches in which perceptual input is evaluated with respect to internally generated forward models provide compelling and increasingly influential explanations of phenomena in the perception and motor control literatures (e.g. Jordan & Rumelhart, 1992; Kawato, 1999). For example, DIVA, an influential model of speech production, incorporates a forward model component that predicts the auditory signal likely to result from a particular configuration of articulators within the vocal tract (Guenther & Micci Barreca, 1997). The forward model component accounts for the speed and efficiency with which the system can control speech movements, given the relatively slow mechanisms by which acoustic feedback influences speech production.

Similar forward modeling approaches may also provide a promising explanatory framework for spoken language comprehension. As in the domain of motor control, forward modeling of the perceptual attributes of upcoming speech provides a compelling explanation for the remarkable speed and efficiency of speech perception and spoken word recognition in the face of considerable variability. This is a particularly attractive feature of expectation-based approaches to higher-level language comprehension as well, which propose a central role for expectations in processes such as syntactic comprehension and lexical processing (e.g. Levy, 2008; Altmann & Kamide, 1999). We propose that comprehension also involves developing expectations about the acoustic realization of upcoming speech, conditioned on various contextual factors such as prosodic phrasing, speech rate, discourse history, and speaker-specific characteristics. Previous work suggests that these expectations are best characterized as probability distributions, with listeners representing not only the expected form of a spoken word given the set of contextual conditioning factors, but also a measure of the variance or uncertainty of their estimate (Clayards et al., 2008; Levy et al., 2009). The degree of congruence between these perceptual expectations and the incoming acoustic signal then contributes to the differential activation of competing lexical alternatives. Finally, when perceptual expectations are incongruent with the actual realization of a word, listeners should update their beliefs about the cues that condition their perceptual expectations, resulting in adaptation that more closely aligns listeners' expectations with the characteristics of the signal in context.

Speech prosody generates acoustic regularities that are likely to foster expectations about the acoustic realization of upcoming speech sounds, including pitch and temporal characteristics that listeners perceive as patterning. This perceived patterning has been shown to influence real-time spoken language processing. For example, manipulations of pitch and duration early in an utterance influence the interpretation of cues to prosodic structure several syllables downstream (Dilley & McAuley, 2008; Dilley et al., 2010; Brown et al., 2011). The distal locus of these effects suggests that they are rooted in listeners' expectations about the acoustic-phonetic realization of upcoming segments.

Speech rate is particularly likely to systematically influence listeners' expectations about upcoming material. Speech sounds are interpreted relative to the global speech rate of an utterance, affecting perceived phoneme distinctions such as voicing contrasts (e.g. Miller, 1987; Reinisch et al., 2011). Therefore, listeners must evaluate the incoming signal with respect to a speaker's estimated rate. Dilley and Pitt (2010) demonstrated that effects of context speech rate scale up to the perceived rate of articulation of larger constituents, in-

cluding syllables and words. They found that when the speech rate of a function word (e.g. *or* in the phrase *leisure or time*) is increased relative to the rate of the surrounding context by either speeding up the function word or slowing down the surrounding context, participants are less likely to report hearing a function word in a sentence transcription task. Conversely, when the speech rate of segments surrounding a location in which a function word would be licensed grammatically (e.g. *leisure time*) is slowed down relative to the surrounding context, participants are more likely to report hearing a function word within the relatively slow portion of speech, effectively hallucinating having heard this item.

The findings of Dilley and Pitt (2010) suggest that listeners rapidly entrain to the rate of an utterance and develop speech rate expectations that influence the perceived number of morphophonological constituents within a spectrally ambiguous stretch of speech of a certain duration. However, the task used in this study does not directly address the prediction of the forward modeling account that the observed speech rate effects have an expectation-based locus. Indeed, it is possible that the sentence elicitation task itself contributed to the observed effects by engaging the production system and inviting explicit comparison of the perceived utterance with different alternative parses. Without time course information, it is unclear whether the observed effects of speech rate on the appearance and disappearance of words are based on perceptual expectations or post-perceptual reinterpretation of the input.

Experiment overview

As a first step toward testing the predictions of the forward model, we used the visual world paradigm (Tanenhaus et al., 1995) to assess the effect of context speech rate on the interpretation of indefinite singular and plural noun phrases. In these noun phrases, the presence or absence of the plural morpheme *–s* was phonemically ambiguous, due to the presence of a sibilant-initial word following the target expression. In addition, the speech rate of sentence material distal (i.e. non-local) to the onset of the target expression was manipulated such that the segments surrounding the determiner in singular noun phrases had a faster speech rate than the distal context and the segments surrounding the onset of plural noun phrases had a slower speech rate than the distal context. These distal context manipulations were predicted to bias listeners to fail to perceive acoustically present determiners within singular noun phrases and to falsely perceive determiners prior to plural noun phrases, thereby shifting whether the noun phrases were judged to be singular or plural. We hypothesized that listeners' expectations about the acoustic-phonetic realization of upcoming segments within an utterance would be the source of these perceptual effects, and that manipulating the speech rate of material distal to the onset of the target expression would therefore influence fixations to pictures of singular vs. plural referents during the real-time processing of the noun phrase.

Methods

Participants

Thirty-two native English speakers from the University of Rochester participated in the visual world experiment. All participants had normal hearing and normal or corrected-to-normal visual acuity.

Materials

The speech stimuli used in the experiment were 24 grammatical declarative sentences containing a target noun phrase consisting of an adjective and plural noun (e.g. *(a) brown hen(s)*). Each target expression was preceded by at least six syllables of utterance context ending in a vowel or rhotic consonant (e.g. *The Petersons are looking to buy*), a phonetic context in which a high degree of coarticulation with the determiner *a* would be expected. The target expression was followed by at least two additional syllables, beginning with a sibilant-initial word (e.g. *soon*) to increase participants' reliance on the determiner, rather than the presence or absence of plural *–s*, as a cue to number.

	preceding context	determiner region	target expression
singular			
no manipulation	1480	381	598
distal manipulation	2812	381	1137
plural			
no manipulation	1480	313	598
distal manipulation	888	313	359

Table 1: Mean durations of each sentence region by target expression type and condition. For each recording, duration of the determiner region was identical between distal- and no-manipulation versions of each recording.

Spoken sentences containing singular and plural versions of the target expression were elicited from 12 speakers and recorded using a Marantz PMD 660 digital recorder sampling at 44.1 kHz in a sound-attenuated booth. A singular and plural version of two critical items were selected from each speaker's recordings. The pitch synchronous overlap-and-add algorithm was then used to create two resynthesized versions of each recording (Moulines & Charpentier, 1990), in addition to two versions not discussed here. In the *distal-manipulation condition*, the speech rate of sentence material distal to the potential location of a determiner was altered; in the *no-manipulation condition*, the speech rate of the utterance was unaltered. For singular expressions, the distal speech rate manipulation involved temporally expanding the utterance context preceding and following the determiner region (the region beginning with the word preceding the determiner and ending with the following phoneme, e.g. *buy a b-*) by a factor of 1.9. This manipulation resulted in the

determiner region having a faster speech rate than the surrounding context. Likewise, the distal speech rate manipulation for plural expressions involved temporally compressing the determiner region by a factor of .6, to slow the rate of the determiner region relative to the surrounding context and thereby encourage the perception of an acoustically absent determiner. The mean duration across each region of the stimuli in each condition is provided in Table 1.

To reduce the salience of the singular-plural ambiguity in the critical items, 18 singular and 18 plural filler items were included in which the number of the target expression was more clearly signaled (e.g. *that rusty knife*). Of these filler items, one third were temporally compressed by a factor of .6 and one third were expanded by a factor of 1.9.

Procedure

Each trial began with the presentation of a visual display containing four clip-art pictures. Two pictures depicted singular and plural versions of the target expression. The other two were singular and plural versions of a different object whose labels were phonologically unrelated to the target word. To ensure visual similarity between singular and plural pictures, the two versions of each picture were created by manipulating a single picture, either by duplicating and superimposing copies of a single target object or by isolating a single object within a picture of multiple target objects.

After 500 ms of display preview, participants heard a spoken sentence over Sennheiser HD 570 headphones. Their task was to click on the picture referred to in the sentence. Participants were not given feedback on their performance, and incorrectly selected a distractor picture on fewer than .5% of all critical trials. Throughout the study, eye movements were tracked and recorded using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed after every fifth trial.

Two sets of three lists were constructed by randomizing picture positions and trial order, dividing the experiment into three blocks, and rotating through permutations of the blocks. Within each list, an equal number of items were assigned to each of four singular and four plural conditions. The pairing of items with conditions was counterbalanced across participants, and each participant encountered each sentence only once. All lists started with four filler items to familiarize participants with the referent identification task.

Analyses

Response choices were analyzed with separate multilevel logistic regression models for singular and plural items. Condition, the duration of the sibilant *-s* in each recording, and their interactions were included as fixed effects, and random intercepts and slopes were included for participants and items. For the measure of sibilant duration, we used the duration of the sibilant in each recording prior to manipulation, to minimize collinearity between sibilant duration and condition and to account for known effects of rate normalization on the relation between phoneme duration and segmentation (e.g. Miller, 1987). The duration of the sibilant was standardized by subtracting the mean value and dividing by the standard deviation. The final model was selected by removing fixed and random effects stepwise and comparing each smaller model to the more complex model using the likelihood ratio test (Baayen, Davidson, & Bates, 2008).

Growth curve analysis was used to evaluate the proportions of fixations to singular and plural target pictures over time by condition (Mirman et al., 2008). This analysis method is a variant of multi-level regression modeling that has emerged as an alternative to other statistical methods used to evaluate effects of experimental manipulations in the visual world paradigm. In growth curve analysis, the proportions of fixations to a picture are first aggregated by participants or by items for each condition at each time point sampled within the region of interest. Orthogonal power polynomial terms are then fit to the resulting fixation proportion curves to model variations in curve parameters (e.g. the intercept and slope) that can be attributed to the independent variable(s) and to participant-wise or item-wise variation. Because it explicitly models the trajectory of change in proportions of fixations over time, growth curve analysis is a more appropriate and temporally sensitive analysis technique than traditional analysis of variance approaches, which frequently entail the violation of assumptions about the independence and distribution of data points and the loss of fine-grained temporal detail.

Separate growth curve analyses were conducted for fixations to singular and plural target pictures in response to singular and plural tokens, with the no-manipulation condition used as the reference condition in all analyses. For analysis of fixations during the processing of the target noun phrase, data were aggregated and analyzed by participants and by items (indicated throughout as B_1 and B_2, respectively). Each model used a third order polynomial to capture the generally sigmoidal shape of the curves.

For analysis of fixations during the processing of the target noun phrase, we characterized the effects of condition on the shape of the fixation curve by adding to the base model the effects of condition on the intercept and linear term, which represent the overall mean curve height and the slope of the curve. The onset of the adjective was used as the reference point for consistency across analyses of singular and plural conditions. Because items varied in the extent to which the adjective in the target expression biased listeners toward the target picture relative to the distractor picture, the onset of signal-driven fixations relative to the start of the adjective was determined by visually evaluating the point of divergence between the mean proportion of fixations to either version of the target picture and the mean proportion of fixations to either version of the distractor picture, averaged across conditions. Averaging across singular and plural target pictures and across conditions permitted unbiased evaluation of the mean point of divergence (Barr, 2008), estimated to be 300 ms after the onset of the adjective. Fixation curves were therefore analyzed between 300 and 1500 ms after adjective onset.

Results

Responses in the picture selection task

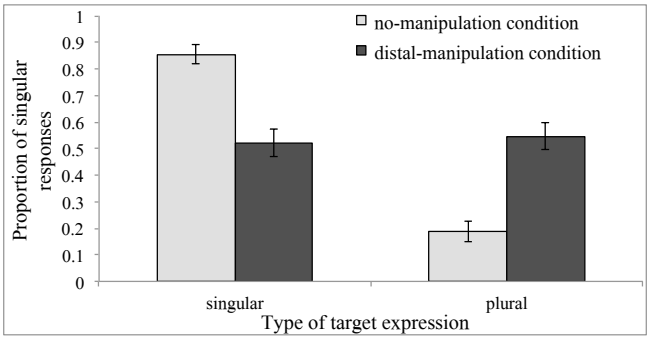

Figure 1: Proportions of correct picture selections.

Results from the picture selection task are shown in Figure 1. The regression model of responses for singular items revealed significant effects of condition. Fewer singular target pictures were selected in the distal-manipulation condition than in the no-manipulation condition (B=2.12, z=5.24, $p<.0001$). Analysis of responses to plural items showed the opposite effect, with more singular target pictures selected in the distal-manipulation condition than in the no-manipulation condition (B=-2.09, z=-5.45, $p<.0001$). For both models, the duration of the sibilant and by-participants and by-items random slopes did not contribute significantly to variance in response data and were not included in the final model.

These data show a pattern similar to those in off-line response choice data obtained by Dilley and Pitt (2010) in a sentence transcription task using a wider variety of function words. Increasing the relative speech rate of the determiner by slowing the rate of the distal context decreased the likelihood of listeners perceiving the determiner, whereas speeding up the distal context biased listeners to perceive a determiner that was not present in the signal. In addition, these results confirmed that the sibilant following the target noun was acoustically ambiguous, causing listeners to base their judgments primarily on the perception of the determiner.

Fixations during processing of target expression

Figure 2 shows the ratio of mean fixation proportions for singular target pictures to the sum of mean fixation proportions for both singular and plural target pictures, averaged across the window starting at 300 ms after the onset of the adjective and ending 300 ms after the offset of the target expression. The pattern of results was similar to the effects observed in responses in the picture selection task. During the processing of singular expressions, the proportion of fixations to the singular picture was higher overall in the no-manipulation condition than in the distal-manipulation condition. Fixations during the processing of plural expressions showed the opposite effect. The proportion of fixations to the singular picture was higher in the distal-manipulation condition than in the no-manipulation condition. This pattern of results supported

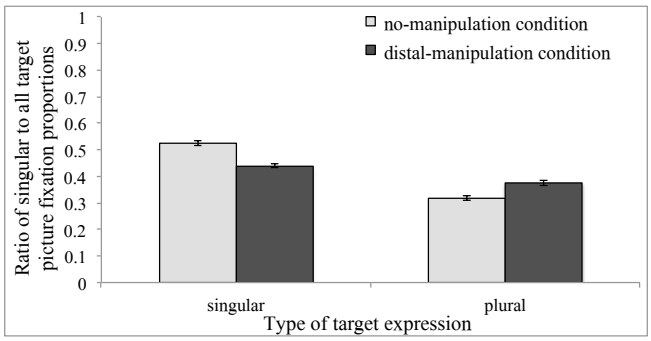

Figure 2: Ratio of mean proportions to singular target pictures to the sum of mean proportions to singular and plural target pictures between 300 ms following adjective onset and 300 ms following the offset of the target expression.

the prediction that effects of distal speech rate manipulation on listeners' judgments would manifest during real-time comprehension.

Figure 3 (left) provides more detailed information about the time course of fixations to singular and plural pictures during the processing of singular items. Growth curve analysis revealed that distal manipulation had a significant negative effect on the intercept term for fixations to the singular picture (B_1=-.16, t_1=-5.33, $p<.0001$; B_2=-.15, t_2=-4.36, $p<.0005$) and a significant positive effect on the intercept term for fixations to the plural picture (B_1=.07, t_1=2.19, $p<.05$; B_2=.07, t_2=1.79, $p<.1$), although this effect was marginal in the items analysis. Relative to the no-manipulation condition, the distal-manipulation condition elicited a lower overall proportion of fixations to the singular picture and a higher overall proportion of fixations to the plural picture. Further, the linear term showed significant effects of distal manipulation on fixations to both the singular picture (B_1=-1.23, t_1=-4.07, $p<.0001$; B_2=-1.18, t_2=-4.28, $p<.0001$) and the plural picture (B_1=.87, t_1=2.84, $p<.0001$; B_2=.79, t_2=19.83, $p<.0001$), suggesting that the rate of change was less steep overall for fixations to singular pictures and steeper for fixations to plural pictures in the distal-manipulation condition.

Figure 3 (right) shows the time course of fixations to singular and plural target pictures during the processing of plural items. For the most part, analyses of fixations during the processing of plural items yielded a pattern of results similar to that obtained for singular items. Relative to the no-manipulation condition, the intercept term was significantly higher in the distal-manipulation condition than in the no-manipulation condition for fixations to singular pictures (B_1=.08, t_1=3.22, $p<.005$; B_2=.08, t_2=3.21, $p<.005$), and lower for fixations to plural pictures (B_1=-.14, t_1=-4.43, $p<.0001$; B_2=-.14, t_2=-4.26, $p<.0005$). Further, the linear term for fixations to the singular picture differed significantly across conditions (B_1=.81, t_1=3.65, $p<.0005$; B_2=-.79, t_2=-19.83, $p<.0001$), but the linear term for fixations to the plural picture did not (B_1=.15, t_1=.45, $p>.1$; B_2=.23, t_2=.70, $p>.1$).

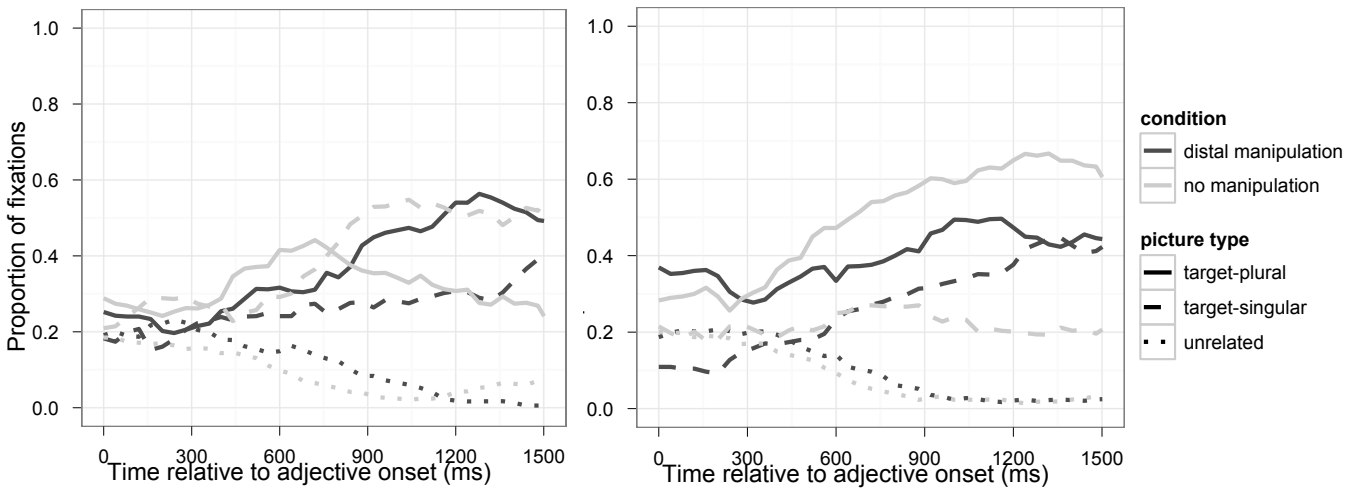

Figure 3: Proportions of fixations to pictures in response to singular (left) and plural (right) target noun phrases. Black and grey lines represent distal- and no-manipulation conditions; solid and dashed lines represent fixations to plural and singular pictures.

In summary, participants were less likely overall to fixate plural pictures and more likely to fixate singular pictures in the distal-manipulation condition than in the no-manipulation condition. This result demonstrates that manipulating the speech rate of an utterance distal to the potential location of a determiner influences whether a determiner is perceived during the real-time processing of indefinite noun phrases, consistent with the predictions of the forward modeling account.

Discussion

When speech distal to the onset of an indefinite determiner at the onset of a singular noun phrase was temporally expanded, listeners were less likely to perceive the determiner and were more likely to select a plural picture as the referent of the noun phrase. Conversely, when speech distal to the onset of a plural noun phrase was temporally compressed, listeners were more likely to select a singular picture, suggesting that the distal context manipulation induced the perception of a determiner that was not acoustically present in the signal. The absolute speech rate of the determiner region of each item was identical across conditions, demonstrating that it was the speech rate of the determiner relative to its surrounding context, rather than its absolute rate, that drove these effects.

Crucially, fixations to singular and plural pictures revealed that effects of distal speech rate occurred during the real-time processing of the target expression, strongly suggesting a locus of the effect in perceptual expectations. The observed time course of speech rate effects demonstrates that listeners entrain to the overall rate at which speech sounds are articulated within an utterance and expect this speech rate to persist in upcoming material, biasing them to expect relatively long stretches of speech to contain more morphophonological constituents and relatively short stretches of speech to contain fewer. Indeed, expectations based on context speech rate may be a powerful cue in the online interpretation of natural speech, which is generally readily interpretable despite frequently degraded or ambiguous spectral cues to segmental content.

According to a forward modeling account of spoken language processing, the language processing system includes a component that integrates relevant contextual information from a variety of sources, such as speech rate and speaker identity, to predict the acoustic-phonetic attributes of different lexical alternatives. These expectations crucially influence listeners' perception of the incoming acoustic signal and the relative activation of competing lexical alternatives. Further, mismatch between expectations and the signal is predicted to result in feedback that continuously updates the probabilistic links between contextual factors and expected outcomes.

Although most of the growing body of work demonstrating fine-grained sensitivity to acoustic detail has not been framed in terms of perceptual expectations (e.g. Gow, 2001; Hawkins, 2003; Salverda et al., 2003; though see McMurray et al., 2011), such findings are congruent with this framework. Particularly suggestive are effects of acoustic detail being conditioned by context manipulations (e.g. Salverda et al., 2007; Sumner & Gafter, 2011) and by speaker-specific characteristics (e.g. Kraljic et al., 2008; Creel et al., 2008).

Our findings set the stage for further tests of predictions of the forward modeling account. For example, if perceptual expectations about upcoming speech are represented probabilistically, the variance of the distribution of expected acoustic forms should influence the magnitude of expectation-based effects on perception. Moreover, incongruence between expectations and the actual realization of a word in context should result in perceptual adaptation bringing expectations more in line with relevant characteristics of the signal (e.g. speaker-specific accent or idiosyncratic prosody). To test these predictions, we are manipulating the relative speech rate not only of the determiner region, but also of segments surrounding the sibilant following the target expression.

Conclusions

This study demonstrates that perceptual expectations have sufficiently strong effects on perception to make words appear within and disappear from the signal during real-time language processing. Forward modeling of the acoustic realization of upcoming speech may be a crucial mechanism enabling listeners to cope with and exploit variability in the input, particularly when spectral cues are degraded or ambiguous. These findings establish distal speech rate manipulation as a suitable experimental paradigm for testing further predictions of the forward modeling account.

Acknowledgments

This research was supported by an NSF predoctoral fellowship (MB), NSF grant BCS-0847653 (LCD), and NIH grants HD27206 and DC0005071 (MKT). We gratefully acknowledge Dana Subik for participant recruitment and testing, Anne Pier Salverda for valuable discussions, and Chelsea Marsh for assistance with stimulus creation.

References

Altmann, G., & Kamide, Y.(1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474.

Brown, M., Salverda, A., Dilley, L., & Tanenhaus, M.(2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review*, *18*, 1189–1196.

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

Creel, S., Aslin, R., & Tanenhaus, M. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*, 633–664.

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., & McAuley, J. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*, 291–311.

Dilley, L., & Pitt, M.(2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*, 1664–167.

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

Guenther, F., & Micci Barreca, D. (1997). Neural models for flexible control of redundant systems. In P. Morasso and V. Sanguineti (eds.), *Self-organization, Computational Maps and Motor Control* (pp. 383–421). Amsterdam: Elsevier-North Holland.

Hawkins, S.(2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405.

Jordan, M., & Rumelhart, D.(1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*, 307–354.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*, 718–727.

Kraljic, T., Samuel, A., & Brennan, S.(2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*, 332–338.

Levy, R.(2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K.(2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*, 21086–2109.

McMurray, B., & Jongman, A.(2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–246.

Miller, J.(1987). Rate-dependent processing in speech perception. In A. Ellis (ed.), *Progress in the psychology of language* (pp. 119–157). London: Erlbaum Associates.

Mirman, D., Dixon, J., & Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*, 475–494.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 666–688.

Reinisch, E., Jesse, A., & McQueen, J.(2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 978–996.

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

Salverda, A., Dahan, D., Tanenhaus, M., Crosswhite, K., Masharov, M., & McDonough, J.(2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, *105*, 466–476.

Sumner, M., & Gafter, R.(2011). Integrating frequency, formality and phonology in the perception of spoken words. Talk presented at the 85th annual meeting of the Linguistic Society of America, Pittsburgh, PA.

Tanenhaus, M., Spivey–Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.