Metrical expectations from preceding prosody influence spoken word recognition

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

Anne Pier Salverda (asalverda@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

Two visual world experiments tested the hypothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 showed that phonemically overlapping words with different initial stress patterns compete for recognition. Experiment 2 further demonstrated that fundamental frequency and syllable timing patterns across material preceding the target word can influence the relative activation of competing alternatives with different initial stress patterns. The activation of alternatives with initial stress was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word. These findings suggest that expectations about the acoustic realization of an utterance include information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues. These results are interpreted as support for expectation-based forward models in which acoustic information in the speech stream is interpreted based on expectations created by prosody.

Keywords: Prosody; spoken word recognition; lexical stress; visual world paradigm; expectations; lexical competition

Introduction

A growing body of work indicates that expectations about the acoustic realization of the phonemes and prosody of a spoken sentence influence how listeners initially interpret incoming acoustic-phonetic cues during spoken language processing. For example, manipulations of pitch and duration early in an utterance influence the interpretation of cues to prosodic and morphophonemic constituency several syllables downstream (Dilley & McAuley, 2008; Dilley et al., 2010; Brown et al., 2011; Dilley & Pitt, 2010). However, little is known about the types of representations that contribute to these perceptual expectations. The present study investigates whether perceived prosodic and metrical patterning across preceding portions of an utterance can influence listeners' expectations about the metrical organization of upcoming material, modulating their interpretation of proximal cues to lexical stress and therefore influencing the activation of potential lexical candidates.

Lexical stress is a key contributor to sentence-level prominence patterns and rhythmicity. Listeners are sensitive to segmental and suprasegmental cues to stress during spoken word recognition (Cutler, Dahan & van Donselaar, 1997). Although vowel quality is the most potent stress cue to influence lexical processing in English, other suprasegmental cues such as duration also distinguish stressed from unstressed syllables, and judgments about these suprasegmental stress cues are modulated by surrounding prosody in off-line tasks (Niebuhr, 2009). Cues to stress may influence not only the recognition of particular words, but also the process of segmenting the speech stream more generally. For example, listeners are more likely to misperceive phrases like "she's a must to avoid" as "she's a muscular boy" than they are to mishear "in closing" as "enclosing", suggesting that listeners preferentially posit word boundaries prior to prominent syllables (Cutler & Butterfield, 1992). This metrical segmentation strategy is substantiated by the distribution of stressed syllables within the English lexicon: approximately 90% of content words in conversational English have initial stress (Cutler & Carter, 1987).

Perceived metrical patterning is a potentially powerful source of expectations in speech perception. Speech prosody often exhibits characteristics that listeners perceive as patterning (Couper-Kuhlen, 1993; Pierrehumbert, 2000). For example, listeners tend to hear stressed syllables in English as perceptually isochronous, i.e., as occurring at regular intervals (e.g., Lehiste, 1977). In addition, previous work using non-linguistic auditory stimuli (e.g. sequences of alternating tones) has demonstrated that pitch, temporal, and/or amplitude patterning in distal (i.e. non-local) auditory context can influence the processing of proximal material (e.g. the perceived relative prominence of high vs. low tones; Woodrow, 1911; Thomassen, 1982). The tendency for speakers to use recurring sequences of pitch accents within an intonational phrase (Couper-Kuhlen, 1993; Pierrehumbert, 2000) may likewise contribute to the perceived metrical structure across syllables in an utterance.

We conducted two visual world experiments to test the hypothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 verified that phonemically overlapping words with different initial stress patterns compete for recognition. Experiment 2 further demonstrated that fundamental frequency and syllable timing patterns across material preceding the target word can influence the relative activation of competing alternatives with different initial stress patterns. The activation of alternatives with initial stress was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word, and vice versa. These findings suggest that expectations about the acoustic realization of upcoming speech include information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues during spoken word recognition.

Experiment 1

The main goal of Experiment 1 was to establish that phonemically overlapping words with different initial stress patterns compete for recognition. We demonstrate that materials containing a target word with relatively neutral segmental and suprasegmental stress cues on the initial syllable elicit initial activation of both initially-stressed and initially-unstressed lexical alternatives.

Methods

Participants The participants were 16 students from the University of Rochester. All participants were native English speakers with normal hearing and corrected-to-normal visual acuity, and received $7.50 for their participation in the study.

Materials The 24 speech stimuli used in the experiment were grammatical declarative sentences containing either a word with initial strong-weak stress (e.g. *jury*) or a phonemically overlapping word with initial weak-strong stress (e.g. *giraffe*). The initial syllables for each related pair of words were produced with as close to the same vowel quality and pronunciation as possible. Each stimulus began with at least two disyllabic words with initial primary stress, followed by one or two monosyllabic words (e.g. *Heidi sometimes saw*). This distal context material was followed by another monosyllabic word (e.g., *that*) followed by the target word. Whereas the preceding context for each item was the same for both SW and WS target words, the sentence material following the target word differed to maximize semantic coherence (e.g. *the jury leaving the courthouse* vs. *the giraffe in the city zoo*).

To discourage participants from noticing the stress pattern manipulation, we included 48 filler items for which the visual display contained two pictures whose labels had a different phonological relation (e.g. words with onset-embedded competitor words, like *antlers* and *ant*). For half of the filler items, one of the phonologically related words was mentioned in the utterance. In addition, an equal number of SW, WS, and monosyllabic target words were used in the filler items. Eight filler items were identity-spliced between the first and second syllables of the target word, whereas the rest were spliced at some point prior to the target word.

The first author recorded multiple tokens of each sentence as WAV files at 44.1 kHz, producing each sentence with minimal F0 excursions and slight F0 declination. Each recording was split into two halves at the end of the first syllable of the target word, at a point in the waveform with an amplitude of zero. Identity- and cross-spliced versions of the item containing the SW target word were created by splicing together the last half of a SW recording either with the first half of another SW recording or with the first half of a WS recording. Likewise, identity- and cross-spliced versions of the item containing the WS target word were created by splicing the last half of a WS recording together with the initial SW and WS sentence fragments used to create the SW items.

Procedure The study was divided into three phases. In the first phase, each of the 304 clip-art pictures used in the visual world experiment was presented to the participant in the center of the display, with its label printed underneath. Each picture-label pair was presented for a minimum of 3 seconds. Participants proceeded at their own pace through the set of picture-label pairs by pressing the space bar.

The visual world experiment began immediately following the picture-label exposure phase. Each trial started with the presentation of a visual display containing four pictures, two of which corresponded to the phonologically related SW and WS words on critical trials. The remaining two distractor images were selected such that they were distinct from the two potential target pictures with respect to visual and semantic properties and the phonological properties of their labels. After 500 ms of display preview, participants heard a spoken sentence, and their task was to click on the picture that was referred to in the sentence. They were not given feedback on their performance during the experiment. 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 every five trials.

Immediately after the completion of the visual world experiment, we assessed participants' ability to generate the appropriate labels for both members of each stress-alternating word pair. Participants named each of the 48 associated pictures, presented in a randomized sequence, and their responses were recorded. Responses were scored as correct if they preserved the phonemes and stress pattern across the initial two syllables (e.g. *jury*, *jury box*, and *jury members* were all considered correct, but *jurors* was not).

For the visual world experiment, two lists were constructed by randomizing the positions of the images on the screen within each trial and pseudorandomizing the order of trials within the list. Within each list, half of the experimental trials had SW target words and half had WS target words, and of these, half were identity-spliced and half were cross-spliced.

The assignment of items to each of the four conditions was counterbalanced across participants, for a total of eight lists. Six practice trials were included at the start of the experiment to familiarize participants with the picture selection task.

Analyses For statistical analysis, proportions of fixations to the target, competitor, and distractor pictures on experimental trials were averaged across the window starting at 200 ms after the onset of the target word and ending 750 ms later, i.e., 200 ms after the mean offset of the target word. The mean proportions were then transformed using the empirical logit function (Cox, 1970). Effects of target word type and splicing condition on logit-transformed fixation proportions were analyzed in a multilevel linear regression model. Fixed effects included picture type, target word type, splicing condition, trial number, and interactions between these factors. Random effects included intercepts and slopes for participants and items. Trial number was standardized by subtracting the mean value and dividing by the standard deviation. To select the final model, effects were removed stepwise and each reduced model was compared to the more complex model using the likelihood ratio test (Baayen et al., 2008).

Results and discussion

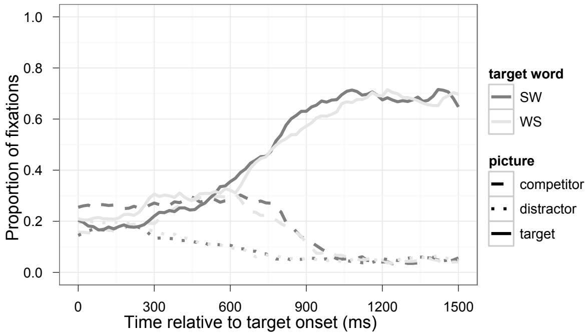

Figure 1: Proportions of fixations to target, competitor, and distractor pictures in Experiment 1. Line color denotes target word type (SW vs. WS); line texture denotes picture type.

Overall, participants found the referent identification task easy, and clicked on the incorrect picture on less than 3% of all experimental trials. These trials were excluded from further analysis. In addition, we excluded trials for which the target or competitor picture was not correctly named by the participant in the third phase of the experiment (18.9%), since their performance on this task revealed whether they associated the intended names with the pictures.

Figure 1 shows the proportion of fixations to the target, competitor, and distractor pictures as a function of condition starting at the onset of the target word. As expected, hearing the initial sounds of a WS word elicited transient competition from a phonemically overlapping SW competitor, starting at approximately 250 ms after the onset of the target word. The proportions of target and competitor fixations were roughly equivalent until approximately 600 ms after target word onset, when fixations began to converge on the target picture. Crucially, hearing the initial sounds of a SW word also elicited competition effects from WS competitors with approximately the same magnitude and time course.

Multilevel linear regression analyses confirmed the prediction that words with initial WS stress initially compete for recognition with phonemically overlapping words with initial SW stress. The logit-transformed proportion of fixations to the distractor pictures was significantly lower than the transformed proportions of fixations to the target (*B*=.59, *SE*=.11, *t*=5.16, $p<.0001$) and competitor (*B*=.29, *SE*=.08, *t*=3.70, $p<.0005$) pictures, after taking into account by-participant and by-item random intercepts, by-participant random slopes for picture type, and by-item random slopes for the interaction between picture type and splicing condition. Neither the target word stress pattern nor trial number contributed significantly to model fit, suggesting that competition effects were similar for SW and WS words and were stable across the experiment. Neither factor was included in the final model. Further, fixed effects of splicing condition did not contribute significantly to variance in fixation proportions. Taken together, these findings indicate that, for our materials, the competition between phonemically overlapping words with different initial stress patterns was similar for SW and WS target words.

These results verified that phonemically overlapping words with different initial stress patterns compete for recognition, and suggested that statistically-based heuristics or biases to interpret lexically stressed syllables as the onsets of content words do not dominate processing, at least when corresponding weak and strong syllables have similar vowel quality and segmental pronunciation. The overall similarity of lexical competition effects in identity- vs. cross-spliced items further suggested that we succeeded in creating items with relatively neutral segmental and suprasegmental cues to the lexical stress of the initial syllable.

Experiment 2

The goal of Experiment 2 was to characterize the effects of preceding prosody on listeners' initial interpretation of different-stress cohort pairs. In this experiment, the acoustic characteristics of preceding portions of the utterance distal to the target word were manipulated, leaving the acoustic realization of the target word and its immediately surrounding context unchanged. Syllables in the distal context with lexical or sentence-level stress were manipulated to have the same relative F0 level (either low or high with respect to surrounding syllables) and to be roughly isochronous. We hypothesized that this prosodic manipulation would bias listeners to expect upcoming syllables with similar pitch and timing characteristics as preceding prominent syllables to be lexically stressed, and conversely to expect upcoming syllables with different pitch and timing characteristics to be unstressed.

Methods

Participants The participants were 32 students from the University of Rochester who met the same inclusion criteria

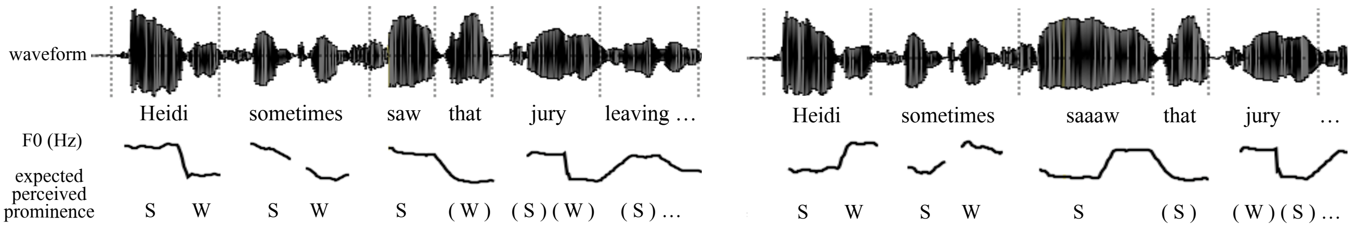
waveform
Heidi sometimes saw that jury leaving ...
F0 (Hz)
expected perceived prominence
S W S W S (W) (S) (W) (S) ...
Heidi sometimes saaaw that jury ...
S W S W S (S) (W) (S) ...

Figure 2: Example item in SW-biasing (left) and WS-biasing (right) conditions. The fundamental frequency (F0) and timing of stressed syllables within preceding distal context were manipulated to bias listeners to perceive the initial syllable of the target word as stressed or unstressed. Acoustic properties of the target word and its proximal context were the same across conditions.

as for Experiment 1.

Materials The stimuli used in Experiment 2 were created by manipulating the prosody of the sentence material preceding the target word in each of the identity- and cross-spliced items used in Experiment 1, using the pitch synchronous overlap-and-add algorithm (Moulines & Charpentier, 1990). All distal F0 manipulations involved removing non-vocalic pitch points and shifting the remaining pitch points within each context syllable up by 35 Hz (for high syllables) or down by 25 Hz (for low syllables). This manipulation preserved the natural microvariation and F0 declination of the original recording while imposing salient periodic alternations onto the F0 contour. F0 manipulations were performed for all syllables through at least the second syllable following the offset of the target word. Whether the first syllable of the target word had low or high F0 varied between items. Temporal manipulations involved compressing or expanding the rime of the first monosyllabic word within the preceding context such that the duration of the fifth intervocalic interval (i.e. the interval between vowel onsets in the fifth and sixth syllables) either matched the mean duration of the following two intervocalic intervals or was twice the mean duration of the preceding four intervocalic intervals, following Dilley and McAuley (2008). Similar prosodic manipulations were performed on filler items.

Two versions of each item were created with different acoustic characteristics across the distal context preceding the target word but the same acoustic characteristics across its proximal context (i.e. the preceding adjacent syllable) and all subsequent material (Figure 2). In the *SW-biasing condition*, syllables in the distal context with lexical and/or sentence-level stress (e.g. *Heidi sometimes saw...*) were manipulated to have the same relative F0 level as the initial syllable of the target word (e.g. high, cf. Figure 2), and the duration of the fifth syllable was manipulated such that the timing of the sequence of prominent syllables was approximately isochronous with the timing of the first syllable of the target word. Whether the fifth syllable was shortened or lengthened to accomplish this regularity depended on the structure of the preceding context: It was shortened when the distal context ended in one monosyllabic word and lengthened when it ended in two. The SW-biasing context was predicted to bias listeners to perceive the initial syllable of the target word as lexically stressed (consistent with e.g. *jury* rather than *giraffe*). This bias was predicted to increase the initial proportion of fixations to the SW competitor when the target word began with an unstressed syllable, and to decrease the initial proportion of fixations to the WS competitor for SW target words.

Conversely, in the *WS-biasing condition*, the F0 of stressed syllables in the distal context was manipulated to have the opposite relative F0 level as the initial syllable of the target word (e.g. low, cf. Figure 2), and the duration of the fifth syllable was manipulated such that the timing of the sequence of prominent syllables was approximately isochronous with the timing of the syllables immediately preceding and following the first syllable of the target word. The WS-biasing context was predicted to bias listeners to perceive the initial syllable of the target word as unstressed, and therefore to increase the initial proportion of fixations to the WS competitor for target words beginning with a stressed syllable and to decrease the proportion of fixations to the SW competitor for target words beginning with an unstressed syllable.

Procedure The procedure was the same as Experiment 1.

Analyses Proportions of fixations to target, competitor, and distractor pictures on experimental trials were computed and averaged across the same time window as in Experiment 1. Effects of word stress, preceding prosody, and splicing condition on logit-transformed fixation proportions were analyzed using multilevel linear regression. Fixed effects and random slopes included picture type, target word type, distal prosody condition, splicing condition, trial number (i.e. the position of the item in the sequence encountered by the participant), the initial F0 of the target word (low vs. high), and interactions between factors. Since we were primarily interested in the effects of distal prosody on the relative proportions of fixations to the target and competitor pictures, fixations to the distractor pictures were not included in the analysis. Trial number was included in the analyses due to recent work suggesting that listeners rapidly adapt to the reliability of prosodic cues, particularly in counterbalanced experimental designs (e.g. Kurumada et al., to appear; Brown et al., under review). A full explication of prosodic adaptation effects, however, is beyond the scope of the present paper.

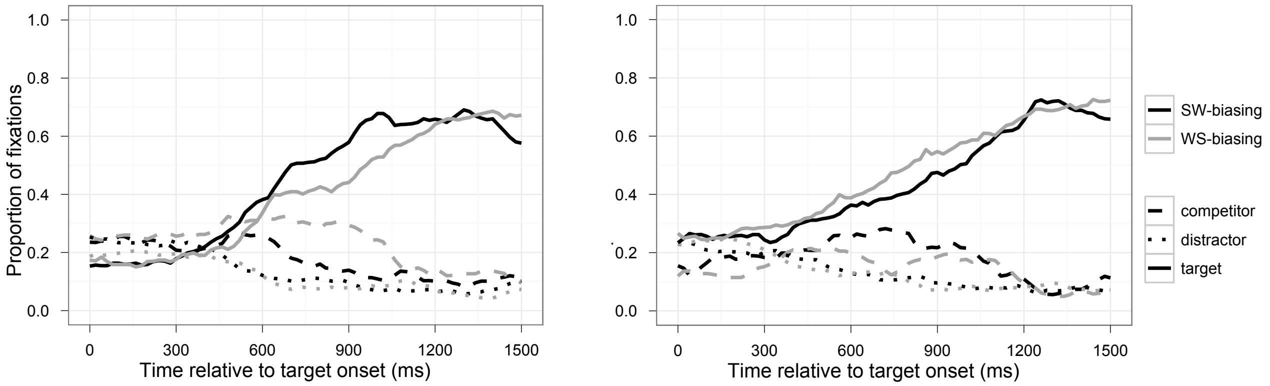

Figure 3: Proportion of fixations to target, competitor, and distractor pictures upon hearing SW (left) and WS (right) target words in SW- vs. WS-biasing contexts in Experiment 2. Line color denotes target word type; line texture denotes picture type.

Results and discussion

Trials in which participants clicked on the incorrect picture (<2%) or for which participants generated an incorrect label for the target or competitor picture in the post-test (29.1%) were excluded from analysis. Figure 3 shows the proportion of fixations to the target, competitor, and distractor images as a function of condition starting at the onset of the target word. Starting at around 350 ms after the onset of SW target words (Figure 3, left), SW-biasing prosody resulted in higher proportions of target fixations and lower proportions of competitor fixations, with effects persisting for approximately 800 ms. For WS words (Figure 3, right), SW-biasing prosody had opposite effects, with relatively low proportions of target fixations and relatively high proportions of competitor fixations.

	B	*SE*	*t*	*p*
intercept	-.65	.09	-7.20	<.0001
target picture	.37	.12	3.11	<.005
WS-biasing prosody	.17	.09	1.91	<.1
trial number	-.06	.03	-2.06	<.05
target picture x WS-biasing prosody	-.29	.12	-2.39	<.05

Table 1: Parameters of the final multilevel regression model of logit-transformed fixation proportions in response to SW target words in Experiment 2. The model included by-item intercepts and slopes for picture type and prosody condition.

Multilevel linear regression analyses revealed not only the predicted three-way interaction between picture type, distal prosody condition, and target word stress pattern (B=.50, SE=.25, t=2.02, p<.05), but also a significant interaction between these three factors and trial number (B=-.46, SE=.17, t=-2.73, p<.01). To investigate the source of this significant four-way interaction, separate analyses were conducted for SW and WS words. Analysis of fixations during the processing of SW words revealed a significant interaction between target picture and prosody condition (Table 1). Participants were more likely to fixate the competitor picture in the WS-biasing condition than in the SW-biasing condition (B=.17, SE=.08, t=2.13, p<.05); target fixations did not differ significantly by condition. Although trial number had a significant main effect, it did not enter into any significant interactions.

Analysis of fixations during the initial processing of WS words revealed not only a significant interaction between target picture and prosody condition, but also a significant three-way interaction between target picture, prosody condition, and trial number (Table 2). We explored this interaction by fitting two additional models to the data, with trial number centered one standard deviation above and below the mean. These models revealed that the interaction between target picture and prosody condition was significant early in the experiment (B=.56, SE=.18, t=3.18, p<.005). In the WS-biasing condition, target fixations were significantly higher (B=.32, SE=.13, t=2.49, p<.05), whereas competitor fixations were significantly lower (B=-.24, SE=.10, t=-2.46, p<.05). However, the interaction between target picture and prosody condition was not significant late in the experiment.

General discussion

Results from two visual world experiments supported the hypothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 showed that initially unstressed words compete for recognition with phonemically overlapping words with initial stress, and vice versa. Experiment 2 further demonstrated that F0 and syllable timing patterns across material preceding the target word influence the relative activation of competing SW and WS alternatives. The activation of SW alternatives was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word, while the activation of WS alternatives was higher when preceding stressed syllables had suprasegmental characteristics dissimilar to the initial syllable of the target word.

These findings show that expectations about the acoustic realization of upcoming material within an utterance include

	B	*SE*	*t*	*p*
intercept	-.61	.06	-10.28	<.0001
target picture	.30	.10	3.05	<.005
WS-biasing prosody	-.10	.08	-1.20	>.1
trial number	-.10	.06	-1.60	>.1
target picture x WS-biasing prosody	.23	.12	1.96	<.05
target picture x trial number	.16	.08	1.93	<.1
WS-biasing prosody x trial number	.14	.08	1.70	<.1
target picture x WS-biasing prosody x trial number	-.33	.12	-2.82	<.005

Table 2: Parameters of the final multilevel regression model of logit-transformed fixation proportions in response to WS target words in Experiment 2. The model included random intercepts and picture type slopes for participants and items.

information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues during spoken word recognition. This indicates, in turn, that cues to lexical stress in sEnglish are not restricted to word-internal cues such as a syllable's F0, duration and amplitude, but can also include sentence-level patterning. This observation suggests that expectations from a variety of sources influence listeners' interpretation of suprasegmental stress cues.

The observation that prosodic expectations influence the interpretation of suprasegmental stress cues raises questions about the mechanisms by which various sources of contextual information are integrated with the incoming signal. Our findings are congruent with forward-modeling approaches in which perceptual input is evaluated with respect to internally generated hypotheses about the acoustic-phonetic realization of upcoming material. Forward models have been fruitfully explored within influential theories of motor control (e.g. Wolpert et al., 1995; Guenther & Micci Barreca, 1997), and may likewise provide a promising explanatory framework for aspects of spoken language understanding.

Acknowledgments

This research was supported by a 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.

References

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.

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

Brown, M., Salverda, A. P., Gunlogson, C., & Tanenhaus, M. K. (under review). Rapid integration of prosodic and discourse cues during spoken-word recognition.

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction.* Amsterdam: John Benjamins.

Cox, D. R. (1970). *The analysis of binary data.* London: Chapman and Hall.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*, 218–236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.

Cutler, A., Dahan, D. , & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201.

Dilley, L., Mattys, S. L., & 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. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*, 291–311.

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

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.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (to appear). Pragmatic interpretation of contrastive prosody: It *looks like* speech adaptation. *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*, 253–263.

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

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica*, *66*, 95–112.

Pierrehumbert, J. (2000). Tonal elements and their alignment. In M. Horne (ed.), *Prosody: Theory and experiment* (pp. 11–36). Dordrecht: Kluwer Academic Publishers.

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

Thomassen, J. M. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America*, *71*, 1596–1605.

Wolpert, D. M., Gharamani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*, 1880–1882.

Woodrow, H. (1911). The role of pitch in rhythm. *Psychological Review*, *18*, 54–77.