Rapid adaptation in online pragmatic interpretation of contrastive prosody

Chigusa Kurumada, Meredith Brown, Sarah Bibyk, Daniel, F. Pontillo, Michael K. Tanenhaus
{ckurumada,mbrown,sbibyk,dpontillo,mtan}@bcs.rochester.edu
Dartment of Brain and Cognitive Sciences, University of Rochester

Abstract

The realization of prosody varies across speakers, accents, and speech conditions. Listeners must navigate this variability to converge on consistent prosodic interpretations. We investigate whether listeners adapt to speaker-specific realization of prosody based on recent exposure and, if so, whether such adaptation is rapidly integrated with online pragmatic processing. We used the visual-world paradigm to investigate effects of prosodic cue reliability on the real-time interpretation of the construction "It looks like an X" pronounced either with (a) a H* pitch accent on the final noun, or (b) a contrastive L+H* pitch accent on *looks* and a rising boundary tone, a contour that can support a complex contrastive inference (e.g., It *LOOKS* like a zebra...(but it is not)). Eye-movements suggest that listeners process the L+H* on *looks* as an early cue to a contrastive interpretation. This effect, however, diminished when listeners had been exposed to the same speaker using the L+H* accent infelicitously (e.g., *Show me the blue square. Now, show me the BLUE circle*). We argue that the process of prosodic interpretations is modulated by the reliability of prosodic cue values, enabling listeners to navigate variability in prosody across speakers and contexts.

Keywords: Prosody, contrastive accent, pragmatic inference, eye-tracking, adaptation

Introduction

Successfully conveying an idea depends not only on what a speaker says, but also how she says it. Prosody – the tonal and rhythmic realization of speech – allows communication of pragmatic meanings and emotions that interact with the lexical and syntactic contents of an utterance (e.g., "*YOU* shouldn't say that" vs. "You shouldn't say *THAT*"). One long-standing issue in prosody research is how listeners map variable acoustic signals onto underlying prosodic representations. Prosodic features, such as pitch and duration, vary significantly across different speakers, populations, dialects, and contexts. For example, male voices generally have lower pitch than female voices, and speakers tend to use higher pitch when talking to a baby than to an adult. For listeners to map prosodic feature values onto more abstract prosodic representation (e.g., "high" tones), they therefore must take into account numerous situation-specific factors.

The lack of invariance between the acoustic signal and underlying linguistic representations is a more general problem in language comprehension. In studies on speech perception, it has been argued that listeners cope with this problem in two ways: by storing exemplars of speech signals (e.g., Goldinger, 1998; Pierrehumbert, 2001) and tracking statistical information about phonetic cue values (e.g., voice onset time (VOT)) in the input. Recent studies have proposed the idea that listeners can assess how reliably each cue predicts the underlying representations, and rapidly adapt their speech perception to more reliable cues in the input (Dell & Chang, 2013; Kleinschmidt & Jaeger, under review). For instance, listeners' categorization functions for /p/ and /b/ can shift after experiencing VOT distributions with more or less variance (Clayards et al., 2008). Recently attempts have been made to extend this logic to explain how listeners navigate syntactic variability to achieve robust and timely sentence processing (e.g., Fine et al., 2013; Kamide, 2012).

In the current study we evaluate the hypothesis that the human language comprehension system likewise deals with prosodic variability through sensitivity to statistics in the input. Specifically, we ask if listeners adapt their real-time prosodic interpretations to the reliability of prosodic cue values assessed in recent exposure. To this end, we investigate English speaker's interpretation of an intonation contour that is known to evoke a contrastive interpretation: the contrastive pitch accent (fall-rise: often annotated as L+H* in the ToBI convention (e.g., Silverman et al., 1992)) followed by a rising boundary tone (L-H%). This contour can signal a contrast between referents (e.g., We have pie$_{L+H*\ L-H\%}$ [but no cake]; Ward & Hirschberg, 1985) or predicates (e.g., Lisa HAD$_{L+H*}$ the bell$_{L-H\%}$ [but she no longer has one]; Dennison & Schafer, 2010).

This intonation contour has two properties that make it well-suited for investigating adaptation in prosody. First, online comprehension of the L+H* accent has been studied extensively and it has been shown to trigger immediate eye-movements to visually represented contrast items (e.g., Ito & Speer, 2008, Watson et al., 2008). For example, as soon as hearing L+H* on a color adjective (e.g., "Pick up a blue ball. Now, pick up a YELLOW$_{L+H*}$...") listeners fixate color-contrasted items that belong to the same object category as the previous referent. We can examine how recent exposure can modulate such rapid integration of the pitch accent. Second, while both the pitch accent and the boundary tone contribute to the contrastive meaning, their reliability may vary independently. In other words, some speakers may express a contrastive inference primarily through a pitch accent while others may rely more on a boundary tone. One way of navigating this variability would be to evaluate each prosodic representation independently and generalize the information selectively to the same type. In our study we test if lowering of the reliability of L+H* would apply specifically to L+H* in the future input, or it would lead to a down-weighting of prosodic information in general.

In our previous work (Kurumada, Brown et al., 2012, 2013), we embeded the L+H* – L-H% intonation contour in the English sentence *It looks like an X*. The L+H* accent was placed on the verb *looks*, followed by utterance final L-H% (Verb-focus prosody, Figure 1, right). We contrasted this with the same construction pronounced with a canonical ac-

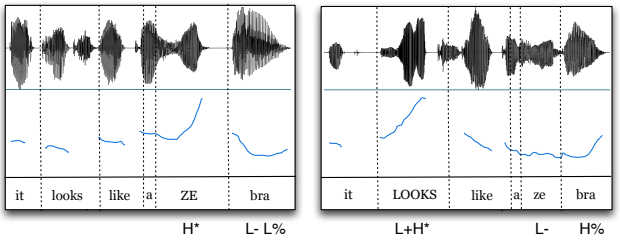
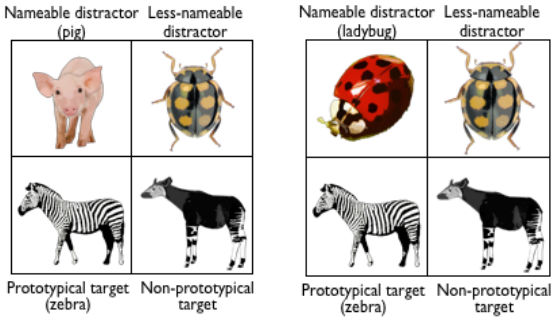


Figure 1: Examples of Noun-focus prosody (left) and Verb-focus prosody (right).



(a) 1-contrast display (b) 2-contrast display

Figure 2: Sample visual displays (Kurumada, Brown et al., 2013, and the main-task in the current experiemnt)

cent placement: H* accent falls on the final noun followed by L-L% (Noun-focus prosody, Figure 1, left). Verb-focus prosody evokes a contrastive interpretation (e.g., *It LOOKS like a zebra, but it is not one*) while Noun-focus prosody indicates that X is likely to be the identity of the referent. That is, these two contours trigger opposite pragmatic meanings depending on the prosodic features. We conducted a visual-world experiment (e.g., Tanenhaus et al., 1995) in which each four-picture display contained a visual contrast pair consisting of a prototypical and a non-prototypical picture of a noun (e.g., a zebra vs. a zebra-like animal) as well as one nameable and one less-nameable distractor (Figure 2). Participants clicked on pictures in response to utterances produced with Verb-focus and Noun-focus prosody.

In this previous experiment, we tested if listeners can generate the complex contrastive inference (e.g., it *looks like* a zebra but it is not) online and if they do so incrementally. We hypothesized that, if listeners make a contrastive inference as they are hearing the L+H* pitch accent in Verb-focus prosody, then they would immediately shift their attention to the non-prototypical target prior to the final noun (e.g., zebra). In particular, we predicted these anticipatory eye-movements to be observed only when there is a single contrast set, as in Figure 2 (a) (*1-contrast display*). In the absence of a contrastive pitch accent (Noun-focus prosody) and when there is more than one contrast set (as in Figure 2 (b), *2-contrast display*), gaze should not shift to the target until the onset of the noun. The results supported this prediction: as soon as L+H* was heard, looks to the non-prototypical member of the contrast set began to increase with the 1-contrast display.

In the present study, we added a between-subject pre-exposure task to this experiment to ask if the patterns of use of L+H* in the pre-exposure phase would affect the time-course of the intonation interpretation in the main experiment. It has been demonstrated that listeners suspend immediate contrastive inferences based on prenominal adjectives (e.g., a tall glass) when they were told that the speaker had an "impairment" that would cause language and social problems. The speaker also described objects erroneously, and used over-informative expressions (Grodner & Sedivy, 2011). Interestingly, the manipulations did not interfere with the speed and accuracy of reference resolution, suggesting that listeners adapted specifically to the speaker's unreliable use of prenominal adjectives as a cue to signal contrast. Extending the idea of Grodner & Sedivy's study, our experiment examines if listeners adapt their real-time contrastive inference through a brief exposure to a particular speaker's speech prosody. We do not provide any explicit instruction that calls attention to the speaker's pragmatic incompetency. Specifically, we ask the following two questions.

1. Will exposure to infelicitous uses of L+H* affect listeners' online (eye-movements) and offline (picture choice) responses to Verb-focus and Noun-focus prosody?

2. Will the reliability information apply selectively to L+H* or will it lead to a more general down-weighting of prosodic information?

Methods

Participants

47 students from University of Rochester were paid $10 to take part in the experiment. They were all native speakers of American English with normal or corrected-to-normal vision and normal hearing.

Stimuli

The experiment consisted of an pre-exposure task (12 items) and a main task (60 items). For the pre-exposure task, we used 27 geometric shapes defined by three shapes, colors, and patterns. We created 12 four-picture visual displays (illustrated in Figure 3-a), each of which was associated with a pair of simple instructions such as "Where's the blue circle? Now, where's the yellow circle?" We used four different carrier phrases "Where's the X", "Find me the X", "Point to the X", and "Show me the X". The second sentence in a pair always began with "Now".

The same speaker from Kurumada, Brown et al.,(2013) made multiple recordings of each pair of instructions with different prosodic contours. In the *High-reliability condition*, six of the 12 pairs were produced with felicitous L+H* on an adjective highlighting a feature contrast (e.g., "Where's the blue circle? Now, where's the YELLOW circle?"); three pairs were produced with felicitous L+H* on a contrasting

noun (e.g. "Where's the yellow triangle? Now, where's the yellow CIRCLE?"); and three pairs with no contrasting features were produced without contrastive prosody. In the *Low-reliability condition*, on the other hand, L+H* was used infelicitously. Six items were produced with L+H* on the wrong constituent; three items contained L+H* in the absence of contrastive features; and three items lacked L+H* despite the presence of contrasting features.

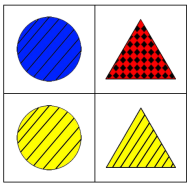
Auditory and visual stimuli for the main task were identical to those used in Kurumada, Brown et al. (2013). 16 imageable high-frequency nouns were embedded in the sentence frame *It looks like an X* and were recorded with Noun-focus and Verb-focus prosody. In addition, 44 filler items contained descriptions of a target picture (e.g., "Can you find the one with yellow spots?"). The target visual stimuli were pictures of the 16 target nouns and 16 pictures of visually similar but less common animals or objects (e.g. zebra vs. okapi; Figure 2). We refer to the picture from each pair that is more common (e.g. zebra) as the *prototypical* target picture, and the other (e.g. okapi) as the *non-prototypical* target picture.

As in our previous experiment, we used two types of visual displays: a) 1 target pair + 2 singletons (1-contrast display, Figure 2, left), and b) 1 target pair and 1 distractor pair (2-contrast display, Figure 2, right). Singletons in 1-contrast trials consisted of one easily nameable picture and one less-nameable picture to equate the complexity of the visual display across trials. In the eye-movement analysis we focused primarily on the trials with the 1-contrast display. As we noted above, it is only when there is a uniquely identifiable contrast set that Verb-focus prosody triggers anticipatory eye-movements to the non-prototypical target (Kurumada, Brown, et al., 2013). Nevertheless, we included 2-contrast displays to control for the task-specific contingencies between the prosody conditions and the target pictures. Without 2-contrast displays, listeners could form a simple association between the L+H* and the non-prototypical member of the contrast set in the display, bypassing a contrastive inference. In the statistical analyses reported below, we always included the display types as a factor.

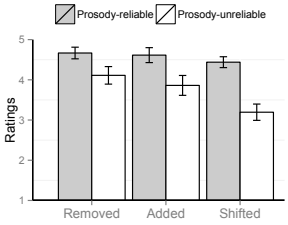
Pre-exposure task stimuli norming

To evaluate the suitability of the visual and audio stimuli for the pre-exposure task, we ran a norming study using Amazon Mechanical Turk, an online crowd-sourcing platform. 48 participants, all self-declared English native speakers, were randomly assigned to either the High-reliability or the Low-reliability condition. They were presented with the 12 items and were asked to click on the target picture referred to in each utterance. After each item, they were asked to rate on a 5-point scale how natural the intonation of the second sentence was (1 = extremely unnatural, 5 = perfectly natural). As expected, the sentences were perceived to be more natural in the High-reliability condition while the mean ratings were overall above average (Figure 3-b).

In addition, we asked the participants to provide comments on the input sentences. 12 participants (50%) in the Low-reliability condition pointed out some irregularity in the speech prosody. Seven out of these 12 participants reported that the accent placement was manipulated. The manipulation was thus salient enough to be noticed by some participants without being totally unnatural.



(a) Sample visual display of the pre-exposure task



(b) Naturalness ratings

Figure 3: Visual display and rating results in the pre-exposure task norming."Removed", "Added" and "Shifted" indicate the types of prosody manipulation applied to the items in the Low-reliability condition

Procedure

As in Kurumada, Brown et al., (2013), participants were first presented with a cover story in which a mother and a child were looking at a picture book, and the mother was helping the child to identify various objects and animals by verbally commenting on them. Each trial began with the presentation of a visual display containing four pictures. After 1 second of display preview, participants heard a spoken sentence over Sennheiser HD 570 headphones and clicked on a picture described by the sentence. Eye movements were tracked using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed after every fifth trial.

16 lists were constructed by crossing 1) the two exposure task conditions, 2) item presentation order, 3) the location of the prototypical and the non-prototypical items in the display, and 4) the prosodic contour (Noun-focus vs. Verb-focus).

Results and discussion

We analyzed three dependent measures to obtain converging evidence about the effect of the reliability manipulation: picture choices, response times, and proportions of fixations to pictures in visual displays. Each variable was assessed with multi-level generalized linear regression models implemented using the lmer function within the lme4 package in R (R Development Core Team, 2010; Bates et al. 2008). Data from three participants was excluded from the analysis due to technical problems during testing.

Picture choices

We first confirmed that participants selected the correct target picture in 97% of filler trials, indicating that participants did not have difficulty completing or attending to the picture selection task. We then analyzed their responses in the 16

critical trials to ask if participants encoded the visual contrast between the prototypical and non-prototypical targets and associated them with the two prosodic contours. In the High-reliability condition, participants selected the prototypical target picture 82.2% of the time with the Noun-focus prosody, but only 46.2% of the time with Verb-focus prosody. In the Low-reliability condition, participants selected the prototypical target picture 60.4% of the time with the Noun-focus prosody, and 44.1% of the time with Verb-focus prosody.

We constructed a multilevel logistic regression model with prosody condition, display type (i.e. 1-contrast vs. 2-contrast display), reliability manipulation and their interactions as fixed effects, and random intercepts and prosody slopes for participants and items[1]. As expected, prosody condition was a significant predictor of participants' picture choices ($\beta = -3.17$, $z = -16.391$, $p < .0001$). The interaction term between the prosody and reliability manipulation was also significant ($\beta = 0.89$, $z = -3.175$, $p < .01$). Thus participants' interpretations of Verb-Focus prosody were more or less consistent across the High-reliability and the Low-reliability conditions while Noun-focus prosody elicited more prototypical (expected) picture responses in the High-reliability condition.

Mouse clicking response times

To take a closer look at the effects of reliability manipulation on participants' picture selection behavior, we calculated the mouse-clicking response times (RTs) by subtracting the time at which the utterance ended from the time at which a picture was selected. We constructed a model including fixed effects of 1) prosody, 2) display type, 3) reliability manipulation, 4) response choice (prototypical vs. non-prototypical target picture) and 5) trial order on the log-transformed RTs. Neither the main effect of reliability manipulation nor its interaction with the prosody was a significant predictor of the RTs ($p > .9$, and $p > .8$ respectively). However, the three-way interaction between prosody (Verb-focus), reliability manipulation (High-reliability) and trial order was significant ($\beta = -6.635, t = -3.087, p < .01$). These results suggest that participants in the High- and Low-reliability conditions were overall equally fast in responding to the contrastive prosody while those in the High-reliability condition became faster in their responses to Verb-focus prosody over the course of the main task.

Eye-movements

Next we asked how the reliability manipulation affected the real-time interpretation of the contrastive prosody. As we mentioned earlier, our analysis focused on responses obtained with 1-contrast displays. We predicted that the results in the High-reliability condition would replicate the findings of Kurumada, Brown, et al.,(2013) in the following two ways: 1) The Verb-focus prosody, but not the Noun-focus prosody, would elicit more fixations to the target contrast set prior to the target noun; and 2) LOOKS$_{L+H*}$ would trigger anticipatory eye-movements to the non-prototypical target. Crucially, we predicted that these early effects of the L+H* accent would be reduced or eliminated completely in the Low-reliability condition.

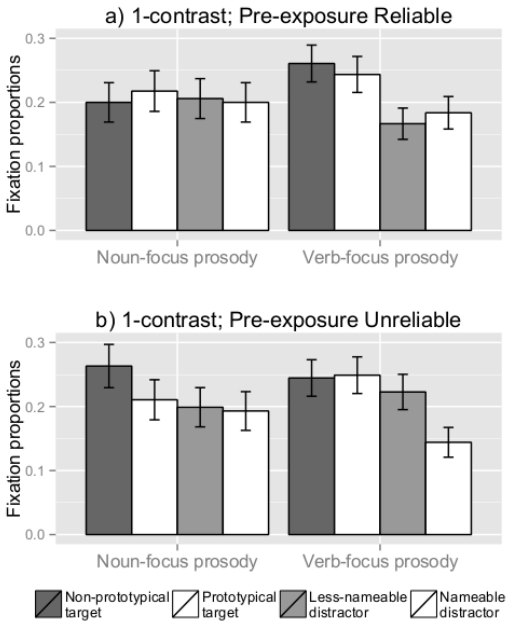


Figure 4: Mean fixation proportions to the target and distactor pictures averaged across the window of analysis. High-reliability condition (top) and Low-reliability condition (bottom). Error bars represent standard errors of the mean.

Our statistical analysis focused on data points sampled within a 200ms window beginning at 200 ms after the onset of "looks". We chose this window for two reasons. First, this allows us to equate the number of samples taken from Noun-focus and Verb-focus prosody conditions. We did not define our analysis window according to word boundaries because the pronunciation duration of "looks (like)" is significantly longer in Verb-focus prosody than in Noun-focus prosody. Second, this analysis window roughly corresponds to "looks like" in Noun-focus condition and "looks" in the Verb-focus prosody. Therefore, we can safely assume that the eye-movements in this time window are not affected by the segmental information of the final noun (e.g., zebra). The only meaningful information from the speech signal that can guide visual search is the prosodic contour.

Figure 4 plots proportions of fixations to the target and distractor pictures averaged across the current window of analysis. As predicted, in the High-reliability condition (top panel), Verb-focus prosody elicited more fixations to the target contrast set (i.e., prototypical and non-prototypical target pictures) while Noun-focus prosody did not give rise to such a bias. On the other hand, in the Low-reliability condition (bottom panel), participants fixated the non-prototypical tar-

[1]P-values for the fixed effects were calculated from F statistics of type 3/type 1 hypotheses using the package LmerTest (http://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf). Random effects were removed stepwise if the full model failed to converge.

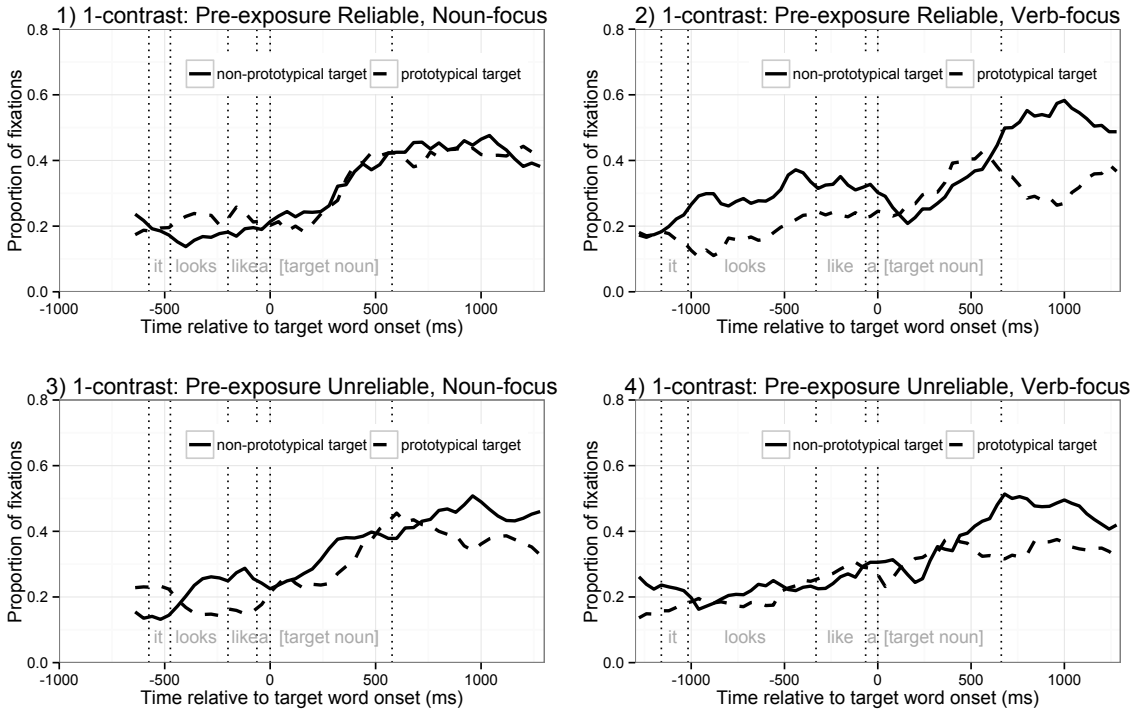


Figure 5: Mean fixation proportions to target pictures with 1-contrast displays. x-axis shows the time relative to the noun onset. Vertical dotted lines indicate the average time points of segment boundaries.

get picture *regardless* of the prosody conditions. Additionally, Verb-focus prosody elicited eye-movements to the less-nameable distractor pictures as well as to the target contrast set. Overall, the effect of the prosodic contours is less pronounced in the Low-reliability condition.

We constructed linear mixed-effects regression models to examine the effects of prosody condition (Noun-focus vs. Verb-focus), display type (1- vs. 2- contrast), exposure task manipulation (High-reliability vs. Low-reliability), and standardized trial order on logit-transformed ratios of fixations to the target contrast set vs. distractor pictures. The final model included random intercepts and slopes for prosody and display type, by participants and items. The model suggested that the predicted three-way interaction between prosody, display type and reliability manipulation was marginally significant ($\beta = 0.244$, $t = 1.632$, $p < .09$). With 1-contrast display, participants in the High-reliability condition were more likely to look at the target contrast set prior to the target noun.

Next, we examined fixation proportions to the non-prototypical target vs. all pictures in the display. The predicted three-way interaction between prosody, display type and reliability manipulation was significant ($\beta = 0.02$, $t = 1.986$, $p < .05$): participants were significantly less likely to fixate the non-prototypical target in response to Verb-focus prosody when they had been exposed to infelicitous uses of L+H*s in the pre-exposure task. Neither the main effect of the trial order nor its interaction with the prosody was a significant predictor ($p > .8$ and $p > .9$ respectively). The absence of an order effect indicates that the effects observed are not driven by task-specific strategies developed during this experiment.

Figure 5 plots the fixation proportions to the prototypical and non-prototypical target pictures in response to Noun-focus and Verb-focus prosody. In the High-reliability condition (top row), Verb-focus prosody triggered anticipatory eye-movements to the non-prototypical targets. Remarkably, this effect became prominent immediately after the onset of "*LOOKS*". Since the current stimuli were recorded by a native speaker attempting to produce most natural sounding contours, it is likely that the initial segment "it" carries prosodic information that allows listeners to predict upcoming prosodic continuations. In the Low-reliability condition (bottom row), however, such immediate effects diminished. In response to Noun-focus prosody, listeners showed numerical trend towards fixating the non-prototypical target while no such trend was observed in the High-reliability condition or in our previous study without a pre-exposure task (Kurumada, Brown et al., 2013). Taken together, the results suggest that the brief exposure to infelicitous uses of L+H* biased participants against making prosody-contingent anticipatory eye-movements prior to the final noun.

Conclusion

The current results provide evidence of rapid and implicit prosodic adaptation. The time course of the online interpretation of Verb-focus prosody was modulated by the reliability

of prosodic cue values manipulated in the pre-exposure task. Critically, the effect generalized across constructions: participants seemed to have learned how likely the contrastive accent (L+H*) would indicate a contextual contrast in the pre-exposure task and applied this knowledge to a new construction in the main-task. Furthermore, participants in the Low-reliability condition still converged on the contrastive interpretation after hearing the L-H% boundary tone at the end of the sentence, which was not manipulated in the exposure task. This indicates that participants selectively downweighted L+H* as a cue to a contrastive interpretation rather than discounting prosody entirely. This complements Grodner & Sedivy's (2011) findings, and illuminates the flexibility of the adaptation mechanism that allows robust prosodic interpretations.

The present study also suggests a way to reconcile some conflicting findings in previous studies on the time course of prosodic interpretations. Dennison & Schafer (2010) used the same prosodic contour (L+H* L-H%) as Verb-focus prosody and reported that participants typically suspended making a contrastive inference until after they had heard both L+H* and the sentence-final L-H% boundary tone. This result is at odds with other studies, including ours, in which the contrastive accent was processed incrementally. This may be at least partly due to a within-subject prosodic manipulation Dennison & Schafer applied. In their study, listeners heard L+H* in conjunction with multiple types of boundary tones. In addition, some instances of L+H* did not convey a contrastive inference. In light of the current results, we would argue that the participants in Dennison & Schafer's study might have learned not to make immediate use of the information provided by L+H*, which did not reliably signal contrast in the task environment. The rapid and implicit learning mechanism demonstrated in this study can thus provide a productive framework for understanding how listers optimize their online pragmatic interpretations of prosody by leveraging reliable, and discounting unreliable, information in the input.

Acknowledgements

Thanks to members of MTan Lab, Anne Pier Salverda, and T. Florian Jaeger for valuable discussion, and to Chelsea Marsh and Olga Nikolayeva for support with participant testing. This research was supported by NICHD grants HD27206 and HD073890 (MKT), a JSPS post-doctoral fellowship (CK), and an NSF graduate research fellowship (MB).

References

Bates, D.M., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108(3),* 804-809.

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

Dell, G., & Chang, F. (2013). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B.*

Dennison, H. Y., & Schafer, A. (2010). Online construction of implicature through contrastive prosody. *Speech prosody 2010 conference.*

Fine, A. B., Jaeger, T. F., Farmer, T. A, and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*. DOI: 10.1371/journal.pone.0077661.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, 105(2), 25179.

Grodner, D. J., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. In E.Gibson and N. Pearlmutter (eds.), *The processing and acquisition of reference*. Cambridge MA: MIT Press, 239-72.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *JML, 58,* 541–573.

Kamide, Y. (2012). Learning individual talkers structural preferences. *Cognition,* 124(1). 66-71.

Kleinschmidt, D., & Jaeger, T. F. (under review). Robust Speech Perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Prosody and pragmatic inference: It looks like speech adaptation. *Proceedings of the 34th Conference of the Cognitive Science Society.*

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. & Tanenhaus, M. K. (2013). Incremental processing in the pragmatic interpretation of contrastive prosody. *Proceedings of the 35th Conference of the Cognitive Science Society.*

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.) *Frequency effects and the emergence of lexical structure*. John Benjamins, Amsterdam. 137-157.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., & Price, P., et al. (1992). ToBI: A standard for labeling English prosody. In *International conference on spoken language processing* Banff.

Watson, D., Gunlogson, C., & Tanenhaus, M. (2008). Interpreting pitch accents in onlinecomprehension: H* vs L+H*. *Cognitive Science, 32,* 1232-1244.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of Fall-Rise intonation. *Language*, 61(4), 747-776.