Pragmatic interpretation of contrastive prosody: It *looks like* speech adaptation

Chigusa Kurumada
kurumada@stanford.edu
Dept. of Linguistics
Stanford University

Meredith Brown
mbrown@bcs.rochester.edu
Dept. of Brain and Cognitive Sciences
University of Rochester

Michael K. Tanenhaus
mtan@bcs.rochester.edu
Dept. of Brain and Cognitive Sciences
University of Rochester

Abstract

Drawing on insights from recent work on phonetic adaptation, we examined how listeners interpret prosodic cues to two opposing pragmatic meanings of the phrase "It looks like an X" (e.g., "It looks like a zebra (and it is one)" and "It LOOKS like a zebra (but its actually not)". After establishing that different prosodic contours map onto these meanings (Experiment 1), we demonstrated that prosodic interpretation is shifted by inclusion of another alternative (Experiment 2); the reliability a speaker's use of prosody to signal pragmatic alternatives (Experiment 3); and most importantly by the distribution of cue values along a continua (Experiment 4). We conclude that listeners derive linguistically meaningful categories from highly variable prosodic cues through rational inference about assumptions that are shared in the conversational context and adaptation to distributional characteristics of prosodic cues.

Keywords: Prosody, contrastive accent, Gricean pragmatics, speech adaptation, rational inference

Introduction

In a famous scene in the movie *Taxi Driver*, Robert DeNiros character repeatedly utters, *You talkin' to me*. As he changes pitch contours, the intended meaning of the utterance shifts from a question to a challenge. As the example illustrates, prosody carries information about a speakers intentions. However, the acoustic features of prosodic alternatives, as well as the mappings between prosodic patterns and intended meanings vary considerably across speakers. For example, although rising boundary tones can distinguish between questions and assertions in many languages, many questions in fact do not end with a rising boundary tone. Also speakers who use "up-talk" often end assertions with a rising boundary tone. Likewise, a pitch accent preceded by an initial drop (fall-rise: often annotated as L+H* in the ToBI convention (Silverman et al., 1992)) can signal the presence of contrast. However, characterizing the acoustic properties that signal this contrast in natural speech is far from straightforward. L+H* and a simple rise (H*) have overlapping interpretations that are highly dependent on utterance context (Ito & Speer, 2008; Watson, Tanenhaus, & Gunlogson, 2008).

How, then, do listeners navigate the lack of invariance in prosodic cues to pragmatic meaning? We propose that listeners solve the variability problem for prosody in the same way as they solve the variability problem for phonetic features, namely by *adaptation*. Just as prosodic contours vary according to both random and systematic factors, phonetic features of speech contain massive variability, which presents a challenge to listeners who are to derive discrete phonemic categories. It has been suggested that the speech perception system deals with the lack of invariance in two ways. One is to store separate, speaker-, group-, and context-specific representations of tokens from the same categories (Goldinger, 1998). The other is to adapt phonetic categories to the distributional characteristics of the acoustic input. For example, Clayards, Tanenhaus, Aslin, and Jacobs (2008) provided evidence that listeners adapt their perceptual categories according to the mean and the variance of a bimodal distribution along a VOT continuum (e.g. /b/-/p/).

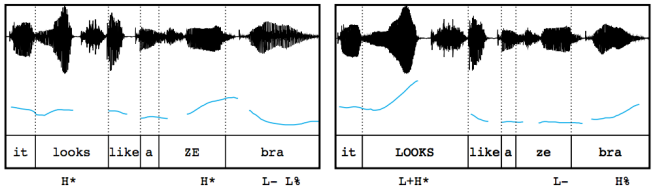

Figure 1: Waveforms (top) and pitch contours (bottom) of the utterance "It looks like a zebra". The affirmative interpretation *It is a zebra* is typically conveyed by the pattern on the left, while the negative interpretation *It is not a zebra* is conveyed by the pattern on the right.

Importantly, the way listeners integrate different kinds of information in speech perception is compatible with the hypothesis that they make rational inferences (Kleinschmidt & Jaeger, 2011). Listeners can weight different kinds of information according to their reliability, adjust phonetic categories based on more reliable information, and ignore deviation from the expected patterns when there is an ad-hoc source for the unfamiliar pronunciations (Kraljic, Brennan, & Samuel, 2008).

The current study draws on these insights to explore the hypothesis that listeners navigate prosodic heterogeneity by adapting their interpretations through rational inference. We focus on the construction "It looks like an X", which can evoke different pragmatic meanings depending on its prosodic realization. A canonical accent placement (as illustrated in Figure 1, left panel, henceforth *noun-focus prosody*) typically elicits an affirmative interpretation (e.g. *It looks like a zebra and I think it is one*). When the verb "looks" is lengthened and emphasized with a contrastive accent (L+H*) and the utterance ends with a L-H% boundary tone (Figure 1, right, *verb-focus prosody*), it can trigger a negative interpretation (e.g. *It LOOKS like a zebra but its actually not one*; see also Dennison & Schafer, 2010).

We explored the adaptation hypothesis in four rating experiments. Experiment 1 established that listeners systematically derive different pragmatic interpretations based on noun- and verb-focus prosodic contours. Experiments 2 and 3 demonstrated that pragmatic interpretations are systemati-

cally modulated by speaker-specific use of particular prosodic contours in different linguistic contexts and the reliability with which a speaker signals pragmatic contrasts prosodically. In Experiment 4, we exposed listeners to affirmative- and negative-interpretation tokens with different distributions of constituent duration and fundamental frequency (f0) values, sampled from a continuum of noun- and verb-focus prosodic contours. Consistent with the adaptation hypothesis, listeners' judgments shifted according to the distributional properties of the input. Taken together, our results provide novel evidence that listeners make optimal use of speaker and context-specific information to derive pragmatic meaning from contrastive prosody.

Experiment 1

We elicited listeners' interpretations of "It looks like an X" in two types of rating tasks to establish that the proposed prosodic contours result in different pragmatic inferences.

Methods

Participants We used an online crowd-sourcing platform (Amazon's Mechanical Turk) for the experiment. We posted 65 separate HITs (Human Intelligence Tasks: experimental tasks for participants to work on) and received 63 HITs from distinct individuals. Participants were all self-reported native speakers of American English. They received $0.80 for completing the task and the mean task duration was 11 minutes.

Stimuli 24 imageable high-frequency nouns were embedded in the sentence frame "It looks like an X". Two tokens of each item with noun-focus and verb-focus prosodic patterns were recorded by a native speaker of American English.

Procedure Participants were first presented with a cover story in which a school teacher described animals and objects in an encyclopedia with pictures that were not directly accessible to his students. In response to a question from a child about what he saw on the page, the teacher said, "It looks like an X" (e.g., *It looks like a zebra*). The participants' task was to judge whether the teacher was referring to a picture of the target noun (e.g., a zebra) or something else.

For each item, participants first heard one of the two target prosodic patterns, and rated how likely it is that the teacher is looking at a picture of the target noun or a picture of something else. Likelihood was indicated using a 100-point scale (0 = something else, 100 = a picture of the noun referent). Next, they heard the same item produced with the other prosodic pattern and answered a 3-alternative forced choice question: If the teacher had used the second intonation pattern, the likelihood of the picture depicting an exemplar of X (e.g. a zebra) would be (a) greater, (b) the same, or (c) less.

Results and Discussion

Figure 2a plots responses in the 100-point-scale rating task. Participants rated items with noun-focus prosody higher than those produced with the verb-focus prosody, indicating that they were more likely to derive the affirmative interpretation

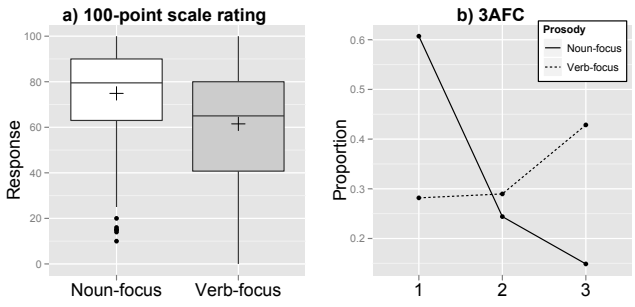

Figure 2: a) The likelihood estimation based on a 100-point scale by prosodic patterns. The crosses within the boxes indicate the mean values. b) Proportions of the responses given in the 3AFC questions [1 = MORE likely to be an X, 2 = no difference, 3 = LESS likely to be an X. The solid line and the dotted line represent the responses based on the noun-focus prosody and the verb-focus prosody, respectively.

(i.e., it is an X) based on the noun-focus prosody. A mixed effects regression analysis (Gelman & Hill, 2006) confirmed that the difference was statistically significant after controlling random effects of subjects and items ($\beta = -13.38$, $p < .001$). Notice, however, that mean values for both prosodic patterns (indicated by the crosses in Figure 2-a) were above 50%. Judgments were thus strongly biased towards the affirmative interpretation.

Figure 2b plots responses in the 3-alternative forced choice task. The difference between the prosodic patterns was significant ($\beta = .41$, $p < .001$): When participants first heard the verb-focus prosody, the noun-focus prosody was rated as more likely to refer to a denotation of the noun (61%) compared to 28% for the verb-focus when it followed the the noun-focus prosody.

Overall, participants made distinct pragmatic interpretations based on the two prosodic patterns. However, the judgments were far from categorical and strongly biased towards the affirmative reading. Taking this as our point of departure, we evaluated the adaptation hypothesis by manipulating factors which we predicted would shift listeners' judgments.

Experiment 2

The adaptation hypothesis posits that noun-focus and verb-focus prosodic contours are not directly mapped onto two distinct pragmatic meanings. Rather, these pragmatic interpretations are obtained through inference based on linguistic and non-linguistic information shared in a particular discourse context. Grice (1989) proposed that utterance meanings are derived through comparison among potential expressions that could have been used in the same situation. We hypothesized that "It looks like an X" would elicit more negative interpretations when the set of sentences produced by the speaker also included a stronger statement (e.g., *It is an X*), on the grounds that the speaker would have simply used this stronger statement to express the affirmative meaning.

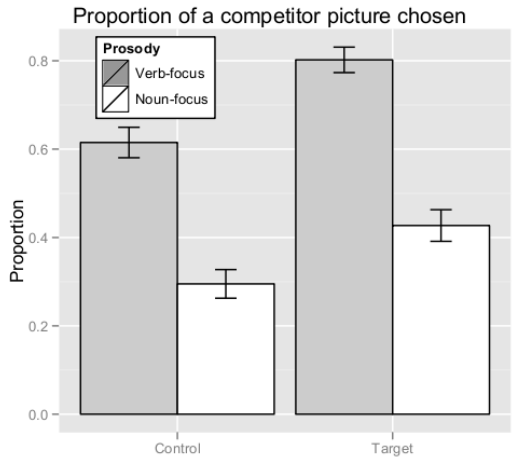

Figure 3: Proportions of competitor pictures chosen in the target and the control conditions in Experiment 2. Error bars represent standard error of the mean.

Methods

Participants We posted 50 separate HITs and received 48 HITs from distinct individuals.

Stimuli An additional 24 stimuli in the form of "It is an X" (e.g."It is a zebra") were recorded by the same speaker as in Experiment 1. Two lists were created based on the items from Experiment 1 and these newly added items. In the *control condition*, all items were in the form of "It looks like an X": 12 with verb-focus prosody and 12 with noun-focus prosody. In the *target condition*, 8 of the 12 noun-focus items were replaced by tokens of "It is an X". Each participant was randomly assigned to the control or test condition.

Procedure Participants were presented with the same cover story as that in Experiment 1 and instructed to select an intended referent of an X out of two pictures: a *target picture* (e.g. a zebra) and a *competitor* (e.g. an okapi, a four-legged animal with black and white stripes only on its legs) after listening to each sentence. Participants completed 24 consecutive trials with no feedback.

Results and Discussion

Analyses were conducted on the items that were common to the two conditions (i.e., we excluded responses to "It is an X" in the target condition and the corresponding sentences in the control condition). Figure 3 illustrates the proportion of a competitor picture chosen in each condition. A mixed logit regression (full factorial; maximum random effects (Jaeger, 2008)) confirmed that there were significant main effects of prosody ($\beta = 1.9$, $p < .0001$) and conditions (target vs. control, $\beta = 1.18$, $p = .061$) as well as a significant interaction term between them ($\beta = 0.83$, $p < .05$), with verb-focus prosody eliciting more competitor responses in both conditions.

Crucially, when the stronger statements were added, participants were more likely to select a competitor picture (indicating a negative interpretation) for both the noun-focus and verb-focus contours. This suggests that the pragmatic interpretation of prosody is not solely determined by the acoustic characteristics of utterances, consistent with the predictions of the adaptation hypothesis. Expectations based on context- and speaker-specific knowledge modulate pragmatic interpretations of contrastive prosody.

Experiment 3

The adaptation hypothesis predicts that listeners are more likely to adapt to cues that are used reliably and systematically. Experiment 3 manipulates the overall reliability with which particular prosodic contours are associated with particular pragmatic meanings.

Methods

Participants We posted 80 separate HITs and received 76 HITs from distinct individuals.

Stimuli 26 items of "It looks like an X" (16 training and 10 test items) were recorded in each of the two target prosodic patterns. For each of the training items, two continuation phrases were recorded to disambiguate the intended meaning. One continuation supported the affirmative interpretation (e.g., "It looks like a zebra because it has black and white stripes all over its body"). The other pattern supplied feedback confirming the negative interpretation (e.g., "It looks like a zebra but it's not; it has stripes only on its legs").

Procedure Participants were presented with a cover story in which a mother and a child were naming animals and objects in a picture-book. The exposure phase included 16 trials in which participants heard the mother tell the child, "It looks like an X". Their task was to choose the likely referent of N between a target and competitor picture (e.g. a zebra and an okapi). They then heard a continuation phrase disambiguating the intended referent.

Each participant was randomly assigned to one of two conditions. In the *reliable-speaker condition*, items with noun-focus and verb-focus prosody (8 items each) were invariably associated with affirmative and negative continuations, respectively. In the *unreliable-speaker condition*, half of the items with noun-focus prosody were followed by negative continuations and half of the items with verb-focus prosody were followed by affirmative continuations.

The test phase was identical across conditions. In this phase, participants made 10 additional judgments in the same format without any feedback. For each item, they were also asked to rate confidence in their judgment on a 7-point scale.

Results and Discussion

As illustrated in Figure 4, the verb-focus prosody systematically biased judgments towards competitor pictures ($\beta = 1.78$, $p < .0001$) in both conditions. Crucially, we also found a significant negative interaction term between prosody and

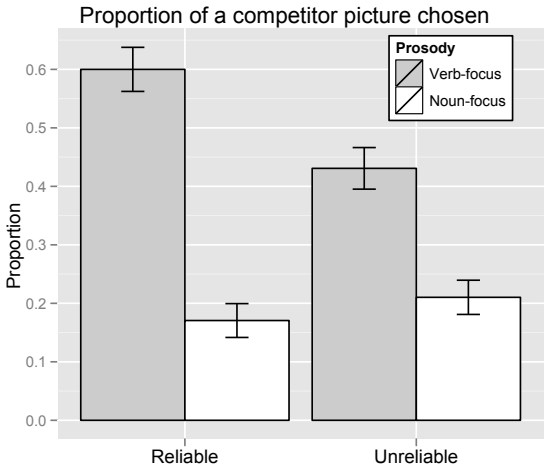

Figure 4: Proportions of competitor pictures chosen in the target and the control conditions in Experiment 3. Error bars represent standard error of the mean.

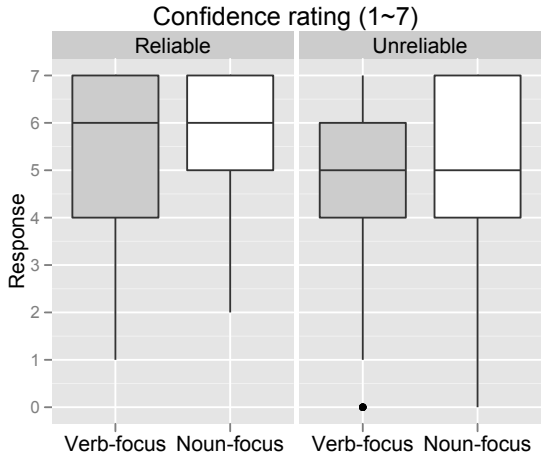

Figure 5: The responses in the confidence rating questions in Experiment 3.

speaker reliability: In the unreliable-speaker condition, the effects of prosody on judgments, and particularly on negative interpretations of contrastive prosody, were weaker overall ($\beta = -1.07$, $p < .004$). That is, participants down-weighted the contrastive accent as a cue to a competitor object after exposure to a speaker who did not use the cue systematically.

Confidence rating data exhibited the same pattern. Confidence ratings were lower overall for utterances with verb-focus prosody ($\beta = 0.3$, $p < .001$), whereas speaker reliability did not significantly affect confidence ratings ($p > .2$). We also found a significant negative interaction between these two factors ($\beta = -.2, p < .0001$). When exposed to an unreliable speaker, participants gave responses based on verb-focus prosody with diminished degree of confidence.

In sum, participants take into account prosodic features idiosyncratic to a particular speaker when deriving pragmatic meanings from prosodic contours. They down-weight prosodic information when it is an unreliable cue to intended meaning.

Experiment 4

Experiment 4 was designed to provide a stronger test of one of the central assumptions of the adaptation hypothesis. If listeners are sensitive to the probabilistic nature of prosodic cues in the input, they should adapt their pragmatic interpretations according to the distribution of tokens in the input. Using resynthesized 12-step continua between noun-focus and verb-focus prosodic contours, we examined how different distributions of prosodic patterns in an initial exposure phase affect listeners' interpretation of utterances in the test phase.

Methods

Participants We posted 360 separate HITs and received complete responses from 324 individuals.

Stimuli The stimuli created for Experiment 3 were divided into six regions corresponding to each of the four initial words (i.e. ① it ②looks ③ like ④ a) and the portions of the final word associated with each of the two tonal targets (i.e. the H* and L-L% in the noun-focus contour and the L- and H% in the verb-focus contour). The turning point in the f0 contour within the final word was used to delineate the final two regions (e.g., ⑤ zeb ⑥ ra, as illustrated in Figure 1). The f0 of each region was sampled at 20 equally spaced time points, and measures from each time point were aggregated across items to derive mean f0 contours for noun-focus and verb-focus utterances, following (Isaacs & Watson, 2010). Likewise, the durations of each region were averaged across items by contour type. Twelve-step continua for each item were derived from these mean f0 contours and durations by interpolating between values within each region and then manipulating the F0 and duration of each recording to match the interpolated values using the pitch-synchronous overlap-and-add algorithm implemented in Praat (Moulines & Charpentier, 1990; Boersma & Weenink, 2008).

These items (12 steps * 26 words) were normed by 50 people using the same 2AFC picture-selection paradigm as in Experiment 3 without feedback. The results of this norming study are summarized in Table 1. Items from most of the continuum steps were more likely to elicit affirmative responses. The items at Step 10 were judged to be the most ambiguous as to the pragmatic interpretations. Based on these norming responses, we postulated that the prosodic cue values for the Noun-focus and the Verb-focus stimuli form distributions with different means and variance, as schematically shown in Figure 6a. The distribution of cue values for the affirmative interpretation (solid line) has a mean value close to the midrange of the continuum and has relatively high variance. In contrast, the distribution for the negative interpretation (dashed line) is considered to have a mean value closer to

Table 1: Proportion of a target picture chosen at each step in the norming study.

Steps	1	2	3	4	5	6	7
Percentages (%)	84	79	80	80	75	77	78
	8	9	10	11	12		
	60	60	48	37	37		

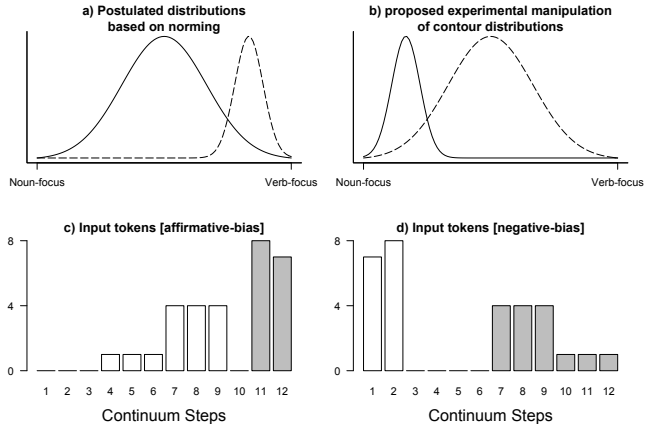

Figure 6: a) A schematic representation of distributions of prosodic cue values postulated based on the norming data. b) Proposed experimental manipulation of contour distributions. c) and d) Input frequencies of tokens sampled from each step of the continuum in the training phase of the affirmative-bias in the negative bias conditions. X-axis: continuum steps. Y-axis: Token frequencies of input utterances. Tokens indicated as white bars were disambiguated as affirmative interpretation and those indicated as shaded bars were disambiguated as negative interpretation.

the higher end of the continuum, with relatively low variance.

Based on these assumptions, we created two sets of exposure items for the current experiment. In the **affirmative-bias condition**, the distributions of the exposure items were meant to approximate the distributions observed in the norming study. Participants heard items sampled from Steps 4-9 with affirmative continuations and items from Steps 11 and 12 with negative continuations (Figure 6c). On the other hand, in the **negative-bias condition**, we tried to reverse this pattern and provided input in the distribution patterns, which are schematically illustrated in Figure 6b. In this condition, listeners heard items from Steps 1 and 2 with affirmative continuations and items from Steps 4-9 with negative continuations. Notice that in either of the conditions participants heard the same number of items from Steps 7-9 while they are identified as items from different categories (Figure 6d).

The adaptation hypothesis predicts that exposure to these affirmative-bias and negative-bias distributions should result in recalibration of the categorization function. Figure 7a plots the proportions of target pictures chosen at each step along the continuum in the norming study. We hypothesized that

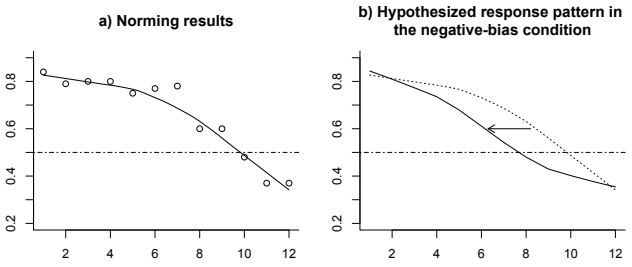

Figure 7: a) Proportions of target pictures chosen (affirmative interpretation) in the norming study. X-axis: Continuum steps (1 = prototypical noun-focus prosody, 12 = prototypical verb-focus prosody). Solid line represents lowess smoothing and dashed line indicates where the stimuli elicit most ambiguous responses (50% chance of a target picture chosen); b) A hypothesized pattern of category recalibration in the negative-bias condition in Experiment 4.

participants' categorization functions would shift towards the negative interpretation, as illustrated in Figure 7b.

Procedure The procedure of the experiment was nearly identical to Experiment 3. Participants were exposed to tokens of "It looks like an X" and selected either a target picture or a competitor picture as the more likely referent. In the exposure phase, they heard 30 items: 15 with affirmative continuations and 15 with negative continuations. The distribution of items sampled from a 12-step continuum differed across conditions, as illustrated in Figure 6c and 6d. In the test phase, which was identical across conditions, participants completed 12 trials in the same format without feedback.

Results and Discussion

Responses are plotted in Figure 8. In the affirmative-bias condition, as predicted, participants' judgments did not deviate from those in the norming study. Items from most of the steps were associated with the affirmative interpretation and the items from Step 10 was judged to be most ambiguous as to their pragmatic meanings.

In the negative-bias condition, however, a wider variety of items were assigned the negative interpretation (i.e., *it is not an X*). As shown in Figure 8, the proportion of affirmative interpretations drops to 50% around Step 7. Items that had been normed to be highly ambiguous (i.e., those from Steps 10) were more reliably assigned the negative interpretation. These results lend strong support for the adaptation hypothesis: Pragmatic interpretation of contrastive prosody does not result from an invariant mapping between sound and meanings. Depending on the patterns of input, participants can rapidly and flexibly adjust their classification criteria so that they can make optimal use of the incoming input.

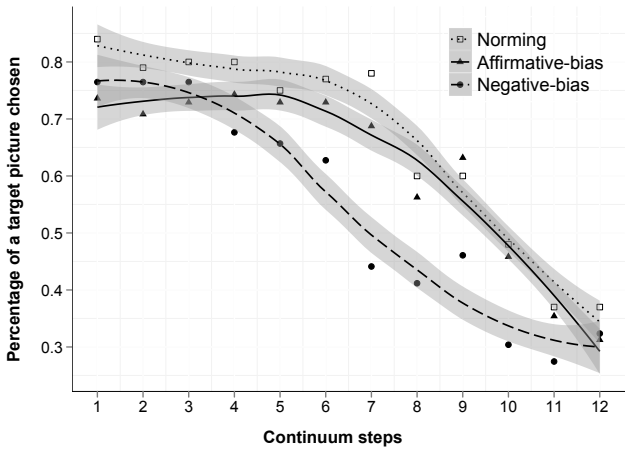

Figure 8: Percentages of target pictures chosen in the test phase [Experiment 4]. Dotted, solid, and dashed lines represent responses from the norming, the affirmative-bias, and the negative-bias conditions, respectively. X-axis plots the continuum steps. Step 1: prototypical noun-focus prosody; Step 12: prototypical verb-focus prosody.

General Discussion

Our results provide novel evidence about how listeners navigate variability in prosodic information to make inferences about an intended meaning of an utterance. We first confirmed that listeners derive affirmative and negative interpretations for "It looks like an X" based on two distinct prosodic patterns (noun-focus and verb-focus contours). The pragmatic inference associated with the contrastive interpretation, however, is affected by the range of linguistic contrasts provided in the context: Introducing a stronger statement ("It is an X") increases the proportion of contrastive implicatures for the "It looks like an X" construction. Listeners also rapidly adjust to speaker-specific use patterns of prosodic cues to pragmatic meanings. Finally, listeners optimize pragmatic interpretations based on probabilistic distributions of prosodic cue values. After hearing 30 tokens of input, listeners adjusted their classification criteria to reflect the properties of the distribution to which they were exposed.

These results provide strong support for the adaptation hypothesis: Listeners take into account the reliability of the mapping between prosodic patterns and intended meanings for a particular speaker when evaluating whether a prosodic contour provides evidence for a contrastive inference. Also, in order to effectively process noisy input, listeners integrate multiple acoustic cues as well as information idiosyncratic to a particular context. In future research we plan to extend our approach to a wider range of constructions and prosodic contours in order to further evaluate the hypothesis that adaptation to prosody can be understood within a rational inference framework.

Acknowledgments

Thanks to Eve V. Clark, Christine Gunlogson and Sarah Bibyk for valuable discussion.

References

Boersma, P., & Weenink, D. (2008). *Praat: Doing phonetics by computer (version 5.0.26) [computer program].* (Retrieved June 16, 2008, from http://www.praat.org/)

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.

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

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Grice, H. P. (1989). *Studies in the way of words* (Vol. 65) (No. 251). Harvard University Press.

Isaacs, A., & Watson, D. (2010). Accent detection is a slippery slope: Direction and rate of f0 change drives listeners comprehension. *Language Cognitive Processes*, *25*(7), 1178–1200.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Kleinschmidt, D., & Jaeger, T. F. (2011, June). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Acl workshop on cognitive modeling and computational linguistics.* Portland, OR.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453-467.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., et al. (1992). ToBI: A standard for labeling English prosody. In *International conf. on spoken language processing* (Vol. 2, pp. 867–870). Banff.

Watson, D., Tanenhaus, M., & Gunlogson, C. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H*. *Cognitive Science A Multidisciplinary Journal*, *32*(7), 1232–1244.