Prosody and intention recognition[1]

Michael K. Tanenhaus

Chigusa Kurumada

Meredith Brown

Department of Brain and Cognitive Sciences

University of Rochester

[1] Thanks to members of MTan Lab, Anne Pier Salverda, Delphine Dahan, 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).

Abstract

Listeners face multiple challenges in mapping prosody onto intentions: The relevant intentions vary with the general context of an utterance (e.g., the speaker's goals) and how prosodic contours are realized varies across speakers, accents, and speech conditions. We propose that listeners map acoustic information onto prosodic representations using (rational) probabilistic inference, in the form of generative models, which are updated on the fly based on the match between predictions and the input. We review some ongoing work, motivated by this framework, focusing on the "It looks like an X" construction, which, depending on the pitch contour and context, can be interpreted as "It looks like an X and it is" or "It looks like an X and it isn't". We use this construction to investigate the hypothesis that pragmatic processing shows the pattern of adaptation effects that are expected if the mapping of speech onto intentions involves rational inference.

In a note to the speakers before the workshop, Lyn Frazier encouraged us to flag particular aspects of our proposals that we thought were novel, promising or suspicious. Lyn also encouraged us to flag unidentified problematic assumptions in the field. In response to Lyn's suggestions, we begin with an example to illustrate some of the challenges involved in understanding the mapping of prosody onto intentions.

In the introduction to his book, *Arenas of Language Use*, Herb Clark (1992) gives a lovely example that illustrates the richness and subtlety of the context-specific based inferences that are required for the listeners to map an utterance onto the speaker's intended meaning. Clark describes a situation in which he addressed the utterance, "I'm hot", to his (then) school-age son, Damon. Clark notes that none of the plausible pre-compiled interpretations of "I'm hot" (e.g., I'm lucky; I'm on a roll, I'm uncomfortably warm; I'm saying the one thing that no child wants to hear a parent say, etc.) captures the intended (and immediately understood) meaning of his utterance. Herb and Damon were playing poker and Damon was about to make a large bet. Herb, who had uncharacteristically been winning most of the hands, was warning Damon that he should think twice about making that bet.

As this anecdote illustrates, context plays a central role in determining speaker meaning. Therefore understanding pragmatic inference requires us to tackle some of the most difficult questions in real-time language processing. For example, how do listeners (and speakers) determine what aspects of a context are relevant? And, since pragmatic inference involves considerations of not only what the speaker said, but also what she chose not to say instead, how do listeners determine those likely alternatives (Grice, 1975)?

Examining how listeners map prosody onto likely speaker intentions adds an additional set of challenges. Herb is unlikely to have uttered "I'm hot" with the "canonical" prosodic contour that he would use in an utterance that was intended to be a simple assertion, as in Figure 1a. Instead some aspects of the prosodic contour, including choice of pitch accents and boundary tones, probably signaled that his utterance should be interpreted in a non-canonical way (e.g., Figure 1b or Figure 1c). There is no way to know exactly *how* Herb said what he said, but it would likely interact with other factors, for example whether he raised his eyebrows, how he might have gestured, and various Herb-centric aspects of his speech. It seems unlikely, for example, that Janet (Fodor) would use exactly the same prosodic contour, let alone the same linguistic expression, were she in a similar situation and intending to convey the same information.

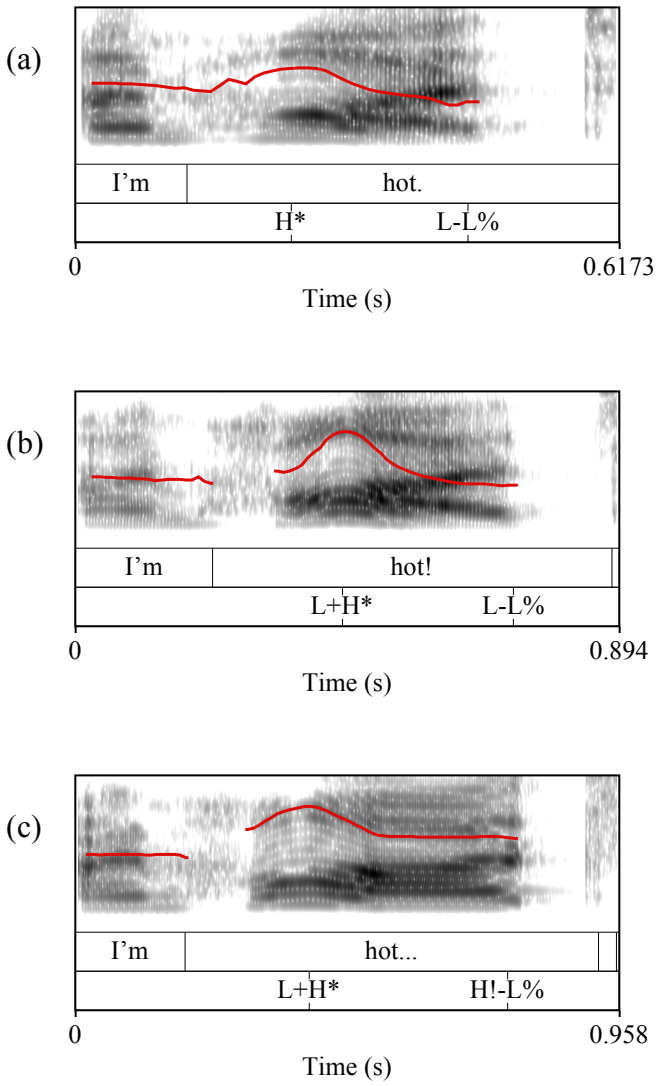


Figure 1. Spectrograms, pitch contours, and ToBI labels for the phrase "I'm hot" uttered with three different prosodic contours. Figure 1a depicts "canonical" statement prosody, whereas the less canonical prosodic contours in Figures 1b-c may be more likely to trigger pragmatic inferences (e.g., that the speaker is surprised or is advising caution).

A fundamental difficulty in studying prosody-based intention recognition is that there are few, or perhaps no, discrete units in the prosodic signal itself. The generally

accepted phonological categories for prosodic analysis, such as pitch accents and boundary tones, consist of bundles of features (e.g., pitch, duration, intensity), all of which shift in a gradient manner. Consequently many of the representations exhibit overlapping acoustic features and/or overlapping interpretations (e.g., Watson, Tanenhaus, & Gunlogson, 2008). Furthermore, actual realization of prosody varies along multiple dimensions such as the gender, age, and social background of speakers. The same speaker also shifts prosodic uses according to contexts and speech conditions (e., child-directed vs. adult-directed speech). To our knowledge, how listeners can extract subtle but pragmatically meaningful acoustic variations in the presence of this substantial variability in the acoustic realization of speech remains an unresolved question.

In an ongoing line of research we have been addressing a set of related questions on how listeners use prosodic information as they process an unfolding utterance which might trigger a pragmatic inference. What is the nature of prosodic representations that support pragmatic inferences? How does the prosody (in relation to the lexical content of the sentence) serve as a cue to unstated speaker meaning? And, how do we arrive at a particular pragmatic interpretation when we hear the particular combination of prosodic features? Underlying these questions is the core inquiry about the mechanism of a language comprehension system: How can we achieve a robust mapping between the realization of speech sounds and phonetic or phonological representations as well as the mapping between these representations and possible interpretation given the constraints provided by the relevant context? We begin by sketching out some of the assumptions that have been guiding our work.

Our framework

As we mentioned above, one of the biggest challenges to any classification model of prosodic categories is the ambiguity arising from the continuous and variable nature of the information. In developing our approach to prosody we have built upon recent work on phonetic categorization (e.g., Clayards et al. 2008; McMurray & Jongman, 2011; and especially Kleinschmidt & Jaeger, 2011; 2012; under review). We draw three analogies between prosodic interpretations and speech perception. First, prosodic representations, such as different pitch accent types, boundary tones and contours, are best characterized as distributions of relevant acoustic cue values. Therefore, just as distributions of acoustic cue values for phoneme representations (e.g., voice onset time for /p/ and /b/) show some overlap, and vary with surrounding acoustic cues, e.g., the duration of the preceding and following vowel, prosodic categories form overlapping distributions that can vary with the phonetic context (Liberman & Pierrehumbert, 1984). The distributional hypothesis explains how acoustically highly variable, and sometimes ambiguous, input can be grouped into two or more functionally contrasted abstract representations.

Second, categorical perception of prosodic representations is an outcome of inferences rather than a property of the acoustic signal itself. It is widely established that perception and recognition of phonemes are dependent not only on the acoustic information, but also on a wide range of contextually derived expectations. For instance, listeners integrate information such as lexical status of the carrier word (Ganong, 1980; Connine & Clifton, 1987; Miller, Green, & Schermer, 1984); lexical effects on compensation for coarticulation (Elman & McClelland, 1988); gender of the speaker (Kraljic & Samuel, 2007; Strand & Johnson, 1996) and information structure of the

sentence (Brown, Salverda et al., in press). Most generally, even the most robust cue-integration mechanisms cannot account for how a contrast such as voicing is perceived without taking into account expectations (McMurray & Jongman, 2011). We hypothesize that, in prosodic processing too, listeners integrate the bottom-up prosodic cues and top-down contextual expectations to inferentially arrive at a particular representation. More specifically, contextual information is expected to serve two important roles. First, it enables the listener to predict what interpretations are likely to be conveyed and how they will be encoded via prosody. For instance, an utterance following a question (e.g., What about beans? Who ate them? (Jackendoff, 1972)) is likely to contain a pitch movement signaling an informational focus (e.g., JOHN ate the beans).. Second, contextual information is used to resolve perceptual ambiguity resulting from prosodic variability and noise. Even when prosodic information is ambiguous or partially lost in noise, listeners can recover an intended interpretation relying on their contextual knowledge.

Third, prosodic processing is highly plastic: Listeners flexibly adapt their prosodic expectations according to recent experiences. As stated above, we hypothesize that listeners predict upcoming prosodic input based on contextual information. Upon receiving the input, listeners match it up with their prediction and compute how much their expectations deviated from the input. The amount of deviation is then used to update expectations for future input. Indeed, it has been demonstrated that listeners adapt their percepts of phoneme categories through rapid perceptual learning (e.g., Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2006; Vroomen et al., 2007; Clayards et al., 2008). We hypothesize that a similar mechanism in prosodic processing, allows the

comprehension system to maintain mappings between the variable perceptual input and more abstract prosodic categories that are more or less constant across speakers and contexts.

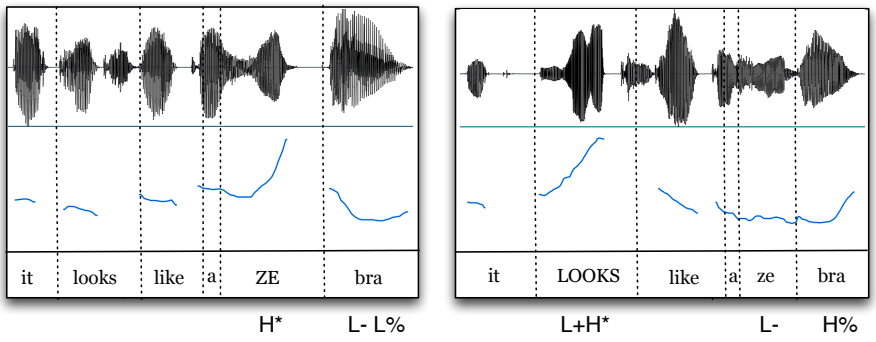
Based on these three assumptions, we propose that the pragmatic interpretation of prosodic contours can best be understood as rational inference over noisy input. Most generally, we are assuming a "data explanation framework" in which perceptual systems seek to provide an explanation for sensory data using "generative models" (Kleinschmidt & Jaeger, under review; Fine et al., 2013; Fine & Jaeger, 2013; Farmer, Brown & Tanenhaus, 2013). These models evaluate hypotheses about the state of the world according to how well they could have given rise to ("generated") the observed perceptual properties. In the case of prosodic processing, a model integrates acoustic cues such as pitch, duration, and intensity to infer which pragmatic interpretation has generated the data at hand.

To better predict future input, hypotheses are continually adjusted based on the observed differences between the predicted state of the world and the observed state of the world, with the goal of minimizing prediction error. The prediction error provides a signal that is used during learning to continuously update the generative model (for similar proposals within a connectionist framework, see Chang et al., 2006; Dell & Chang, 2013). This approach has been successful in explaining how listeners converge on coherent percepts in phoneme identification and speech perception – a domain in which a lack of categorical mappings between acoustic signals and linguistic categories has been investigated in great depth. It has also been applied to work in syntactic processing (Fine et al., 2013; Fine & Jaeger, 2013) and the role of expectations in segmentation (e.g.,

Brown, Dilley & Tanenhaus, 2012). In experiments presented below, we ask whether listeners 1) integrate a multitude of prosodic and contextual cues to constrain their inferences, and 2) adjust weights of these cues according to recent exposure.

Does it *look like* speech adaptation?

We now present an overview of four experiments from an ongoing project that uses the construction "It looks like an X" as a case study to test and refine our hypotheses about how listeners map acoustic cues onto prosodic categories. This construction has a number of desirable properties. First and foremost, it can evoke different pragmatic meanings depending on its prosodic realization (Kurumada, 2013). 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), hereafter the "It is" interpretation. In the context that we investigate, 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 2, right, hereafter verb-focus prosody), it can trigger a negative interpretation (e.g. It LOOKS like a zebra but it's actually not; see also Dennison & Schafer, 2010). The fact that these interpretations map onto different referents, e.g., a zebra or something that only resembles a zebra, makes it possible to determine which interpretation a participant has chosen.



a) Noun-focus prosody b) Verb-focus prosody

Figure 2. Examples of waveforms (top) and pitch contours (bottom) for the utterance *It looks like a zebra*. In the context that we investigate, the affirmative interpretation "It is a zebra" is typically conveyed by the pattern on the left (a), while the negative interpretation "It is not a zebra" is conveyed by the pattern on the right (b).

Second, with verb-focus prosody, we are investigating a 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 $_{\text{L+H* L"H\%}}$ [but no cake]; Ward & Hirschberg, 1985) or predicates (e.g., Lisa HAD $_{\text{L+H*}}$ the bell $_{\text{L-H\%}}$ [but she no longer has one]; Dennison & Schafer, 2010).

The fact that both the pitch accent and the boundary tone contribute to the contrastive meaning means that we can vary the reliability of two asynchronous cues. Moreover, 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; Weber et al., 2006). For example, as soon as listeners encounter the L+H* on a color adjective (e.g., "Pick up a blue ball. Now, pick up a YELLOW $_{\text{L+H*}}$..."), they begin to fixate color contrasted items that belong to the same object category as the previous referent.

In all of the work we will be describing we present participants with variations on a scenario, illustrated in Figure 3, in which the “It looks like an X” utterance occurs in a context in which an adult (e.g., a teacher or a parent) is looking at a picture book with a young child. We created pairs of pictures in which one picture had a common name (e.g., a picture of zebra) and the other picture was similar-looking but did not have a common name (e.g., a picture of an okapi). The participants were instructed that the adult is referring to a picture in the book as he or she addresses an utterance to the child. The participant then made a response to indicate which picture is being referred to.

Figure 3. An example picture used in the preamble of the current comprehension studies. Participants were told that they would be listening to a conversation between an adult and a child in a book-reading context as depicted in this picture.

This cover story was crucial in constraining the range of interpretations that listeners can draw. Without contextual constraints, the "it looks like an X" construction allows the speaker to express a range of nuanced meanings such as uncertainty, or a contrast between visual similarity and similarities in other domains (e.g., If it looks like a duck, walks like a duck, and quacks like a duck…). However, in this current scenario, we are implicitly imposing the assumption that the speaker (the mother) knows what the

identity of a referent is, and is giving a hint to the child so that he can make an appropriate inference, i.e., it is an X or it is not an X. It is an independent question of great importance how interlocutors negotiate underlying assumptions about speaker knowledge and domains (and subdomains) that determine the saliency and likelihoods of possible interpretations within an actual conversational context.

Another topic of interest, which we do not address here, is the extent to which the contrastive inference "it looks like an X but it is not" is conventionalized. The construction "it looks like an X" occurs more frequently than expressions like "smells like" or "sounds like", and it often conveys the fact that an appearance of an object conflicts with its identity. However, a corpus analysis of child-directed speech by Hansen & Markman (2005) found that adult speakers use "looks like" to talk about both appearance and identity, and the interpretation is most often largely context-dependent. For instance, if a child were to say "It's a zebra", and an adult were to answer "It LOOKS like a zebra", then the preferred interpretation is, "but it isn't a zebra. However, if the child instead had said, "It's not a zebra", then the preferred interpretation of "It LOOKS like a zebra" would be "It is a zebra". This observation has been confirmed experimentally (Bibyk, Kurumada & Tanenhaus, in preparation). This suggests that the interpretation of the "looks like" construction is not completely conventionalized and hence the uncertainty in the interpretation needs to be resolved through contextual inference.

We first summarize the manipulations and results of a series of off-line studies which used an on-line crowd-sourcing platform (Amazon's Mechanical Turk) to test: (a)

our assumption that verb-focus and noun-focus prosody with "It looks like an X" probabilistically maps onto different interpretations and (b) the hypothesis that listener adapt their mapping of the these contours onto interpretations based on the statistics of the input (Study 1). We first established that listeners preferentially map noun and verb-focus prosody onto the interpretations "it is" and "but it's not", respectively. We then asked whether listeners would adapt by (a) shifting the strength of their preferences when they were presented with evidence that a speaker often used a stronger alternative to signal the "it is" interpretation (Study 2) and (b) down-weighting prosodic cues when exposed to a speaker who used prosody unreliably (Study 3)[2].

Study 1: Prosodic representations as distributions of acoustic cues

Our most basic claim is that prosodic representations involve distributions of relevant acoustic cues such as pitch, duration and intensity. Acoustic cues should therefore map probabilistically onto different interpretations and listeners should be sensitive to properties of the distribution. Crucially, listeners should therefore adapt the mapping of contours onto interpretations according to the distribution of tokens in the input.

In order to test these hypotheses we selected a clear exemplar of a noun-focus utterance and a clear exemplar of a verb-focus utterance for each item (e.g., zebra) and resynthesized a 12-step continuum of prosodic contours. The stimuli were divided into six regions corresponding to each of the four initial words (i.e. it | looks | like | a) and the

[2] A preliminary report of these three studies, including the methodological details and results are presented in Kurumada, Brown & Tanenhaus, 2012). A longer manuscript reporting these results is under review (Kurumada et al., submitted).

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. 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).

A schematic of the F0 contours for a sample item are presented in Figure 4.

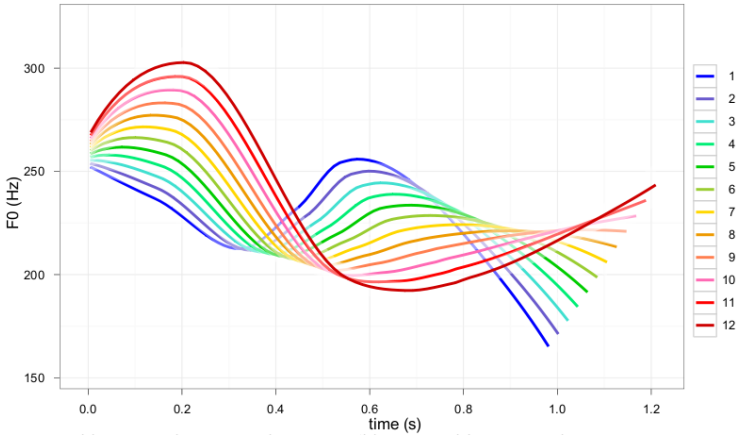


Figure 4: Schematic illustration of the manipulation of F0 and duration in resynthesizing a 12-step continuum of prosodic contours. The numbers represent the 12 steps. Step 1 and Step 12 represent mean prosodic cue values associated with typical Noun-focus and Verb-focus prosodic contours, respectively.

We first established a categorization function for the continua illustrated in Figure 4. We used these results to postulate distributions for how the phonetic contours map onto the "it is" and "but it's not" interpretations. These are illustrated in Figure 5, Panel (a). We then selected two distributions of tokens for presentation to separate groups of participants. The shaded areas in panels (a) and (b) of Figure 5 correspond to the values along the X-axis used in exposing a new set of participants to either the distribution of contours presented in Figure 5-(c) (the *affirmative-bias* condition) or Figure 5-(d) (the *negative-bias* condition). During exposure, participants heard an "It looks like an X" utterance and chose which picture they believed the teacher was intending to refer to. After making a picture selection, participants heard a second clause that disambiguated the intended referent (e.g., "because it has black and white stripes" or "but it isn't because it only has stripes on its legs"). Participants were then presented with 12 new tokens and made picture selections without feedback. The distribution of exposure tokens in the affirmative-bias condition (Figure 5-(c)) was chosen to mirror the distribution that we postulated based on the norming results. In the negative-biased condition, illustrated in Figure 5-(d), we presented ambiguous tokens from Steps 7, 8, and 9 with feedback indicating that the speaker intended to refer to the atypical referent (e.g., the okapi). The predicted effect of exposure on the distributions is illustrated in Figure 6. Figure 6-(a) shows the categorization function from the norming study and the predicted shift in the categorization function after exposure to the negative-biased tokens.

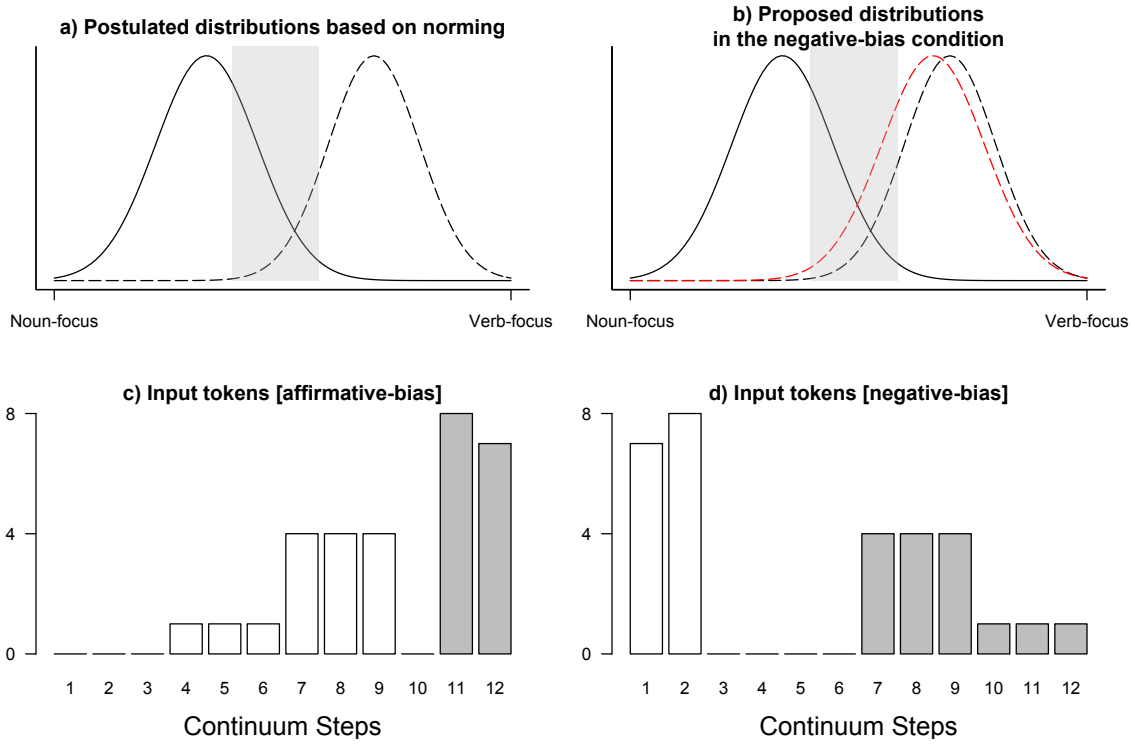


Figure 5: a) A schematic representation of distributions of prosodic cue values postulated based on the results of native-speakers' judgments. The solid line and the dashed line represent the distributions of prosodic cue values for the "it is" and "but it's not" interpretations, respectively. b) Proposed 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 and 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.

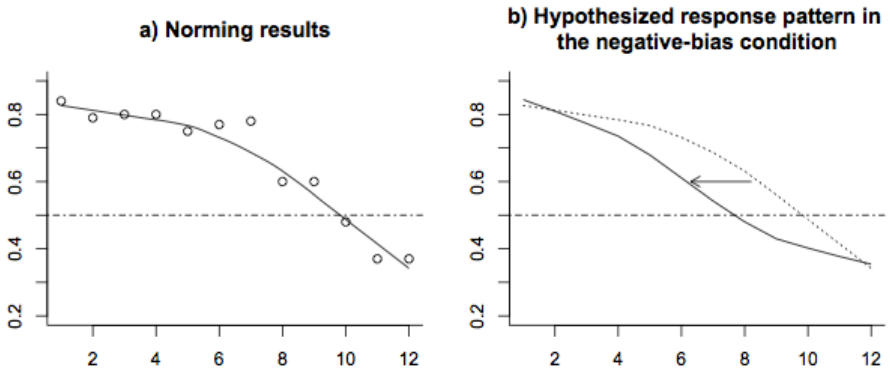


Figure 6. a) Proportion of a target picture chosen (affirmative interpretation) in the norming study. X-axis: Continuum steps (1 = prototypical noun-focus prosody, 12 = prototypical verb-focus prosody). Solid line represents lowess smoothing and dashed line indicates where the stimuli elicit most ambiguous responses (50% chance of a target picture chosen); b) A hypothesized pattern of category recalibration in the negative-bias condition.

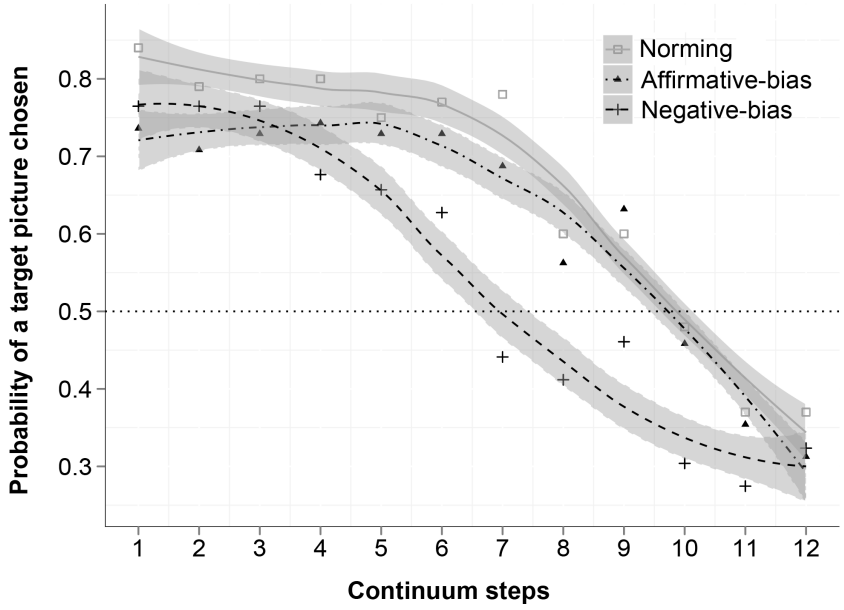


Figure 7. Proportions of target pictures (e.g. zebra) chosen in the test phase. Dotted, solid, and dashed lines represent responses from the norming, the affirmative-bias and the negative-bias conditions, respectively. Continuum steps are plotted on the X-axis (Step 1: prototypical noun-focus prosody; Step 12: prototypical verb-focus prosody).

Figure 7 presents the categorization functions for the norming study, the affirmative-bias exposure condition,and the negative-bias exposure condition. The affirmative-bias exposure condition was chosen to mirror the assumed pre-exposure distribution, whereas the negative-bias exposure condition was predicted to shift participants' categorization functions. The results closely mirror our predictions.

In sum, the results of this study establish that listener's mapping of prosodic tokens onto interpretations is probabilistic and malleable according to recent exposure. Most crucially, listeners are sensitive to the distribution of new tokens, showing the predicted adaptation effects.

Effect of alternatives

Another source of variability is the fact that pragmatic interpretation of speaker meanings relies on listener's estimates of what kind of lexical, syntactic, and prosodic elements the speaker could have produced. For example, the contrastive interpretation of "it looks like an X" depends upon an implied contrast between potential alternative predicates. Specifically, the contrastive accent on "LOOKS" signals a contrast between "(it) looks like (an X)" and its semantically stronger alternative (e.g., "it is (an X)"). This contrast supports the reasoning that the speaker could have said, "it is a zebra" but didn't, which implicates that the speaker meant it was not a zebra. The availability of the contrastive interpretation of "it looks like a zebra" thus hinges on the listener's belief about how likely the speaker would say "it is an X" if that is what she meant. If it is likely, the form "it looks like an X" is more strongly associated with the contrastive interpretation. On the

other hand, if the listener does not believe that the speaker would use "it is an X", the formal contrast does not support the contrastive inference.

To test this hypothesis, in Study 2, we directly tested the effect of semantic alternatives. In this experiment, we used only prototypical instances of the Noun-focus and the Verb-focus prosody. Participants were presented with a cover story in which a male 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 the typical or the atypical referent. Each participant heard twenty-four utterances, twelve each with Noun-focus and Verb-focus prosody. The items were rotated through presentations lists so that different participants heard the version of the utterance with Noun- and Verb-focus prosody, respectively.

For utterances with Noun-focus prosody, participants preferred the typical referent, choosing it on about 70% of the trials, whereas with Verb-focus prosody participants preferred the atypical referent, choosing the typical referent on only 40% of the trials. Thus the mapping of the contours onto interpretations was again probabilistic, with Noun-focus prosody preferentially mapping onto the "it is" interpretation and Verb-focus prosody preferentially mapping onto the "but it's not" interpretation. These results set the stage for testing our prediction that listeners would adapt by shifting their representations if they were presented with evidence that the speaker will use a less ambiguous, i.e., stronger alternative when he is expressing the "It is" interpretation.

We tested this hypothesis by replacing eight of the twelve noun-focus utterances with the stronger statement, "It is an X" (e.g., "It's a zebra!"). We then compared the judgments for the subset of items that could be directly compared to the prior study, by excluding the results from the "It is an X" trials. For details of the design, see Kurumada, Brown & Tanenhaus (2012; submitted). As predicted, the stronger alternative shifted the preferred referent of "It looks like an X" towards the atypical referent for both Noun-focus and Verb-focus prosody. The typical referent was now chosen on only 40% of the trials for noun focus prosody and 20% of the trials for verb focus prosody. These results are consistent with our proposals that direct evidence for use of "it is an X" increases the likelihood that listeners will derive the contrastive inference based on the construction "it looks like an X". In sum, the same prosodic contour produced in the same context can be interpreted differently depending on the listener's expectation about what kind of lexical, syntactic, and prosodic means the speaker could use to express a particular intention. These results are not predicted by approaches that derive an intonational interpretation based solely on the mappings between a prosodic contour, and/or combination of pitch accent and boundary tone (e.g., L+H* L-H%), and a pragmatic meaning or category, e.g., contrast.

Study 3: Speaker reliability

As we discussed in the introduction, acoustic realizations of prosodic information varies across speakers and contexts. When talking to a young baby, adult speakers tend to use a

wider pitch range with more peaks and troughs in a pitch contour (Fernald & Kuhl, 1987). In such a context, a small excursion of pitch may not be pragmatically meaningful and, hence the listener needs to suspend their pragmatic interpretations, which would be warranted in adult-adult conversation. Thus the pragmatic interpretation of prosody requires an effective "down-weighting" of prosodic information that is not a reliable indication of the speaker's pragmatic intentions.

In order to evaluate the effect of speaker reliability we used a design in which an exposure phrase (16 items) was followed by a test phase (10 items). Participants were randomly assigned to the reliable-speaker or the unreliable-speaker condition. In either condition, during the 16 utterance exposure phase, participants made a judgment based on a "It looks/LOOKS like an X" utterance, and then heard a disambiguation continuation that either indicated the "it is" interpretation or the "but it isn't" interpretation as in Study 1. In the *reliable-speaker* condition, the continuation disambiguated all eight Noun-focus utterances in favor of the "it is" interpretation and all eight Verb-focus utterances in favor of the "but it isn't", interpretation. In the *unreliable-speaker* condition, half of the Noun-focus and half of the Verb-focus utterances were followed by phrases that disambiguated the utterance in favor of each interpretation. Thus the participants were receiving feedback that the speaker's use of prosodic patterns did not provide reliable information about her intended referent.

The exposure phrase was followed by a 10 utterance test phase in which no feedback was provided after the participant's judgment. We predicted that participants in the unreliable condition would place less weight on the prosodic information in their judgments. This prediction was confirmed. With the reliable speaker, the typical referent

was chosen on 82% of the utterances with noun-focus prosody compared to only 18% of the utterances with verb-focus prosody. In contrast, with the unreliable speaker, the typical referent was chosen for 78% of the utterances with Noun-focus prosody and 55% of the utterances with Verb-focus prosody.

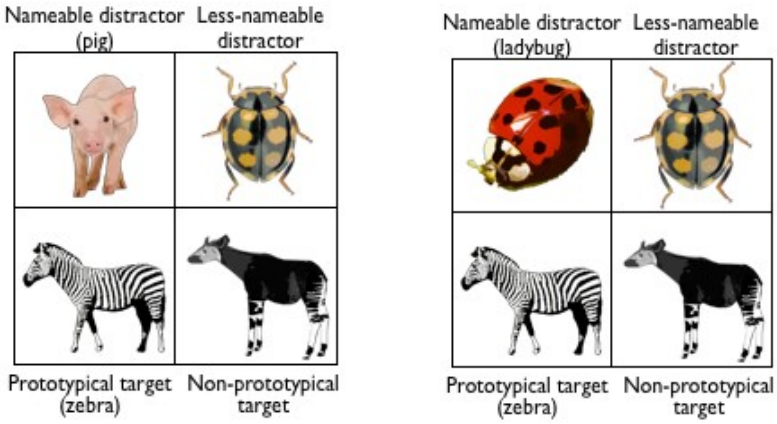
Taken together, the results of the three studies provide strong support for the adaptive nature of the mechanism employed for intonation interpretation. We have shown that listeners are sensitive to the distribution of tokens, modulating their mapping of prosodic contours onto categories just as listeners adapt phonetic categories based on new distributions. Moreover, Studies 2 and 3 suggest that listeners seem to be constantly adjusting their interpretations based on their estimates of how likely it is for a particular linguistic signal, defined with lexical and prosodic information, to convey either of the possible speaker meanings (i.e., it is an X vs. it is not an X).

The off-line judgment paradigms demonstrate that listeners adapt over the course of multiple utterances. However, these studies cannot tell us how the reliability information accumulated over time affects real-time online language comprehension. We hypothesize that online language comprehension actively employs a generative model available each point of time, and any discrepancy between a predicted pattern and an actual input signal would generate an error signal. It is this error signal that allows the listener to adapt to the statistics of the input. Thus we predict that listeners will (a) generate expectations based on cues to prosodic contours and (b) adapt these expectations based on the statistics of the input. We now briefly describe some ongoing work that examines the hypothesis listeners generate expectations based on prosodic information as an utterance unfolds, and modulate their expectations based on the reliability of the

speaker.

Study 4: Expectations in real-time processing

Our studies take advantage of the fact that the intonation contour that most naturally maps onto the contrastive "but it isn't" interpretation consists of both an L+H* pitch accent on the verb "looks" and an L-H% boundary tone. We first established that listeners process an information contour predictively when the L+H* is reliably paired with the L-H % boundary tone. In other words, when there is a unique contrast pair in a visual scene, listeners would launch an eye-movements to a less nameable referent as soon as they heard "(it) *LOOKS*…", indicating that they had generated a prediction about the likely referent. In order to do so we designed a visual world study (Cooper, 1974; Tanenhaus et al., 1995) using displays that contained either one or two contrast sets, as illustrated in Figure 8.



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

Figure 8: Sample visual displays for the 1-contrast trials (a) and the 2-contrast trials (b).

The logic of the study was based on previous work using contrast and contrastive prosody. A line of visual world studies initiated by Sedivy and colleagues (e.g., Sedivy, Tanenhaus et al., 1999), established that when listeners hear a prenominal scalar adjective, e.g., "Pick up the tall glass", they immediately look at the taller member of a contrast set (e.g., the taller of two glasses) upon hearing "tall" when a display contains only a single contrast pair. However, if there are multiple contrast sets (two glasses, and two boxes), listeners do not begin to look at potential referents (e.g., the taller of the objects in the contrast sets) until they hear the noun (Hanna et al., 2003; Heller et al., 2008). As we discussed above, several visual world studies have established that listeners are sensitive to contrastive prosody (Ito & Speer, 2007; Watson, et al., 2008; Weber et al., 2006).

We reasoned that if upon hearing the fall-rise contour (L+H*) on "LOOKS" listener incrementally develop a contrastive interpretation, then with a single contrast set, participants should make anticipatory eye movements to the less nameable referent (e.g., the okapi) about 200 ms or so after the onset of the contrastive pitch accent. We confirmed this prediction in a study conducted in collaboration with Sarah Bibyk and Daniel Pontillo (Kurumada, Brown, et al., 2013; in press). As can be seen in Figure 9, fixations to the non-prototypical target (e.g., an okapi) based on the Verb-focus prosody increased even before the segmental information of the final noun became fully available. This result demonstrates that listeners processed the acoustic cues in the contour incrementally, generating predictions before they encountered either the beginning of the noun (e.g., zebra) or the boundary tone.

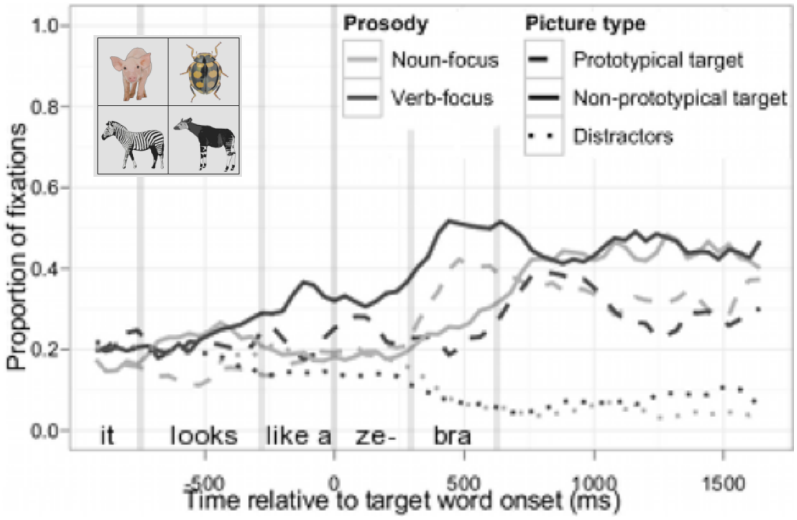


Figure 9. Proportions of fixation to pictures in response to Noun-focus (gray lines) and Verb-focus prosody (black lines) in 1-contrast displays. The x-axis indicates time with respect to the onset of the final noun.

We then tested the hypotheses that listeners would down-weight the information provided by the fall-rise contour, if prior exposure established that a speaker used contrastive focus unreliably (Kurumada, Brown, et al., 2014). The experiment consisted of an exposure and a test phase. The test phase was identical to the experiment described above. In the exposure phase the same speaker either used contrastive prosody with pre-nominal adjectives reliably or unreliably. We used pre-nominal adjectives because this allowed us to expose listeners to information about whether or not the speaker reliably used the L+H* accent to signal contrast, without giving them experience with the "It looks like an X" construction. In the prosody-reliable condition, the speaker provided instructions such as "Click on the blue circle. Now, click on the YELLOW$_{\text{L+H*}}$ circle", in which the L+H* accent highlighted a contextual contrast between two objects (Ito & Speer, 2008). In the prosody-unreliable condition, the speaker used an L+H* accent on a wrong constituent, or did not use one when it would have been informative. After being

exposed to 12 of such exposure items, participants responded to the same items from the original eye-tracking experiment. In the reliable condition, participants again made anticipatory eye movements when they heard LOOKS $_{L+H*}$. However, in the unreliable condition, listeners did not make anticipatory eye movements as they heard LOOKS $_{L+H*}$ but rather waited until the disambiguating noun with the boundary tone. These results demonstrate that the adaptation effect present in off-line judgments does indeed affect the time-course of prosodic interpretation as an utterance unfolds. Listeners generate expectations incrementally based on reliable cues about a prosodic contour, but modify their expectations when a cue becomes unreliable. It is important to note that unlike the earlier off-line experiments, the effects of reliability transferred from one construction in which contrast was signaled by an L+H* on an adjective to a construction in which the interpretation is conveyed by a prosodic contour with both contrastive focus on a verb and a subsequent boundary tone.

Summary, conclusions and implications

Let us return to our original question. Hearing Herb utter, "I'm hot", with a particular prosodic contour, how would Damon have arrived at the particular interpretation Herb intended to convey? In the four studies described in this chapter, we argued that interpretation of intonation makes use of various sources of information about the speaker, context, and the prosodic features experienced in past exposure. We suggested that the discourse context biases the listener's expectations for particular speaker meanings. In playing poker, one would expect the speaker to say something "relevant" to the situation at hand. With this expectation, the comprehension system compares possible

speaker meanings to identify which speaker meaning would be most likely to generate the given utterance. This is done through comparison between alternative hypotheses (speaker meanings) as well as alternative linguistic elements, including lexical, syntactic, and prosodic information, which the speaker could apply to convey those possible speaker meanings. While sentence prosody generally plays an important role in this process, it is effectively discounted when a recent experience indicates that it does not reliably predict the speaker's pragmatic intentions,

This approach is distinguished from the general view that one can posit a one-to-one mapping between given prosodic features of speech (e.g., pitch movement) and a particular pragmatic interpretation or function. Such an approach would not predict that listeners would process and interpret the same acoustic input differently depending on their expectations and recent experiences. We acknowledge that it was a deliberate decision for past researchers to begin their phonological analyses assuming categories abstracted away from their phonetic specifications. Many researchers have, in fact, noted that each phonological category contains substantial phonetic variation (e.g., Ladd, 2008). Our proposal, however, goes beyond claiming that listeners are somehow normalizing phonetic variability across different instances of speech. Rather we are arguing that prosodic interpretation is part of a principled inference process to arrive at speaker meanings based on noisy and variable data. In this process, each cue – contextual or linguistic – is weighted according to its reliability, which probabilistically shifts as an outcome of the inference. The prosodic features of speech alone, therefore, cannot reliably predict pragmatic interpretations of utterances, as has been assumed in previous research. Prosodic information can effectively support inferences only when it is

interpreted against appropriate models with highly-structured knowledge about discourse contexts and linguistic expressions.

As we mentioned earlier, variability is ubiquitous in speech, and the human language comprehension system needs to deal with it at all levels of linguistic representations. Recent studies have begun to address how listeners integrate recent experiences to adjust their inferences about intended messages. These attempts include investigations of speech perception (e.g., Norris et al., 2003), syntactic parsing (e.g., Fine & Jaeger, 2013; Jaeger & Snider, 2013), and semantic comprehension of quantifiers (e.g., Degen, 2013; Yildirim, et al., 2013). The evidence of prosodic adaptation outlined in this chapter demonstrates that listeners' sensitivity to the characteristics of the input extends to the mapping between prosodic profiles of speech and abstract pragmatic information. This points to an exciting possibility of a unified model for language comprehension, encompassing the low-level speech perception through the high-level intention recognition. At all levels, listeners infer an underlying representation that generated the observed input. Along with other recent studies in the same spirit, we argue that this inferential association between the signal and the representation is the key to robust language comprehension in the face of substantial variability in the input.

We acknowledge that the work we have presented is only a first step towards demonstrating the promise of this framework. As with any line of research examining adaptation, the first step is to determine whether adaptation does indeed occur. It then becomes important to determine the scope and generality of the effects, including when and how listeners generalize the learned knowledge across constructions, prosodic contours and contexts. For example, when listeners observe that a speaker uses an L+H*

accent unreliably to signal contrast, they can generalize the information in more than one way. The particular speaker might be unreliable with respect to all prosodic uses, or only to L+H* (e.g., a proficient non-native adult speaker might fail in marking contrast in prosody, yet can be otherwise fully competent). The speaker could also be incapable of reliably recognizing a contextual contrast, but otherwise capable of producing expected prosodic patterns. In consequence, the rate and outcome of adaptation can thus be modulated according to the listener's beliefs about the language, speaker, and context. Ultimately, it will be important to provide a principled account for how listeners generalize from their experience.

We also will need to provide more specific, testable, quantitative models of how listeners combine different types of cues, before concluding that they can be accounted for naturally within a rational inference framework. Perhaps the most difficult challenge is to integrate research on how listeners combine acoustic and phonetic cues, which can be observed and measured, with models of how interlocutors generate expectations for what types of intended meanings are relevant to a particular context or class of contexts, and what utterance types are likely to convey those intentions. That said, we are willing to make a substantial bet that (a) this line of investigation is likely to prove promising; and that (b) adaptation will play a crucial role in how we solve the class of variability and cue-integration problems that arise. We would even make a small wager that this approach might prove useful in understanding how readers generate and use implicit prosody.

References

Beckman, M. E., & Ayers, G. M. (1994). Guidelines for ToBI labeling. Retrieved September 23, 2008, from http://www.ling.ohio-state.edu/research/phonetics/E ToBI.

Beckman, M. E., & Hirschberg, J. (1994). The ToBI annotiation conventions. Retrieved September 23, 2008, from http://www.ling.ohio-state.edu/˜tobi/ame tobi/annotation conventions.html.

Bilyk, S., Kurumada, C., & Tanenhaus, M. K. (in preparation). Context constraints on intonation interpretation.

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/)

Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1374-79).

Brown, M., Salverda, A. P., Gunlogson, C., & Tanenhaus, M. K. (in press). Interpreting prosodic cues in discourse context. *Language, Cognition and Neuroscience*.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review,* 113(2), 234-72.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

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.

Connine, C. M., & Clifton, C., Jr. (1987). Interactive use of lexical information in speech perception. Journal of Experimental Psychology: Human perception and performance, 13 (2), 291.

Cooper, R. M. (1974). Control of eye fixation by meaning of spoken language: New methodology for real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6, 84-107.

Degen, J. (2013). Alternatives in pragmatic reasoning. Ph.D dissertation, University of Rochester.

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. 369.*

Dennison, H. Y., & Schafer, A. (2010). Online construction of implicature through contrastive prosody. *Proceedings of 5th Speech Prosody Conference*.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*, 143–165.

Farmer, T., Brown, M., & Tanenhaus, M. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences*, 36, 211-212.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech, *Infant Behavior and Development*, 10 (3), 279-293.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, 37 (3): 578-591.

Fine, A. B., Jaeger, T. F., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8(10), e77661. [doi:10.1371/journal.pone.0077661].

Ganong, W. F. (1990). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, 3, 41–58.

Hanna, J., Tanenhaus, M. K., Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49 (1), 43-61.

Hansen, M. B., & Markman, E. M. (2005). Appearance questions can be misleading: Adiscourse-based account of the appearance-reality problem. *Cognitive Psychology*, *50*(3), 233–263.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831 – 836.

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, 541-573.

Jackendoff, R. (1972). *Semantics in generative grammar*. MIT Press, Cambridge, MA.

Jaeger, T. F., & Snider, N. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition,* 127(1), 57–83.

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

Kleinschmidt, D., Fine, A., & Jaeger, T. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. *Proceedings of the 34th annual conference of the cognitive science society.*

Kleinschmidt, D., & Jaeger, T. F. (submitted). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel.

Kraljic, T. & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, 13(2), 262-268.

Kurumada, C. (2013). Navigating variability in the linguistic signal: Learning to interpret contrastive prosody. Ph.D dissertation, Stanford University.

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

Kurumada, C., Brown, M., & Tanenhaus, M. K. (submitted). Probabilistic inferences in pragmatic interpretation of English contrastive prosody.

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 Annual Meeting of the Cognitive Science Society.*

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition 133*, 335–342.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. K. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118 (2), 218-246.

Miller, J. L., Green, K., & Schermer, T. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics, 36,* 329-337.

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

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds), *Intentions and plans in communication and discourse* (pp. 271-311). MIT Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109 -147.

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* (Vol. 2, pp. 867–870). Banff.

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.

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

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

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, 49, 367-392.

Yildirim, I., Degen, J., Tanenhaus, M.K. & Jaeger, T.F. (2013**).** Linguistic variability and adaptation in quantifier meanings. *In Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*.