Effects of distributional information on categorization of prosodic contours: Replication of Kurumada, Brown, and Tanenhaus (2018)

Will Clapp
Prof. Judith Degen
LINGUIST 245B

June 12, 2020

1 Introduction

In an experiment investigating the relationship between pragmatic inference and prosodic adaptation, Kurumada et al. (2018) found that speakers were able to tailor inferences based on experience with a biased training set. This line of research draws heavily from the perceptual learning literature introduced by Norris et al. (2003), which demonstrated that listeners retune category boundaries after exposure to ambiguous exemplars of fricatives experienced in lexically biased contexts. Kurumada et al. (2018) were the first to extend this finding to prosodic adaptation. The study focuses on discrimination between two possible interpretations of the phrase "It looks like an X" where "X" stands in for some target referent, e.g. "a zebra." In a noun-focus production of this sentence the referent "zebra" contains two tonal targets: H* in the first syllable and L-L% in the second. In the verb-focus production, the verb "looks" contains the tonal target L+H* while the noun "zebra" contains a target L- in the first syllable and H% in the second. These contours are shown in Figure 1. In conversational context, a typical interpretation of the noun-focus production would be that the speaker is in fact referring to the object X, but a more typical interpretation of the verb-focus intonation may be that the speaker is introducing a contrast between the object X and the actual referent, thereby inviting the listener to draw a pragmatic inference. This distinction is most probably made possible by listeners' abilities to map pragmatic meanings to prosodic features such as pitch accents and boundary tones (Pierrehumbert and Hirschberg, 1990; Beckman and Pierrehumbert, 1986; Liberman and Pierrehumbert, 1984).

Because prosodic cues are an important component of pragmatic inference but differ greatly from speaker to speaker, it follows that listeners would be able to draw statistical inferences about the way that prosodic cues are applied in a given discourse in order to facilitate mutual comprehension. To test this, Kurumada et al. (2018) conducted a perceptual learning experiment in which participants heard sentences of the type "It looks like an X" on a

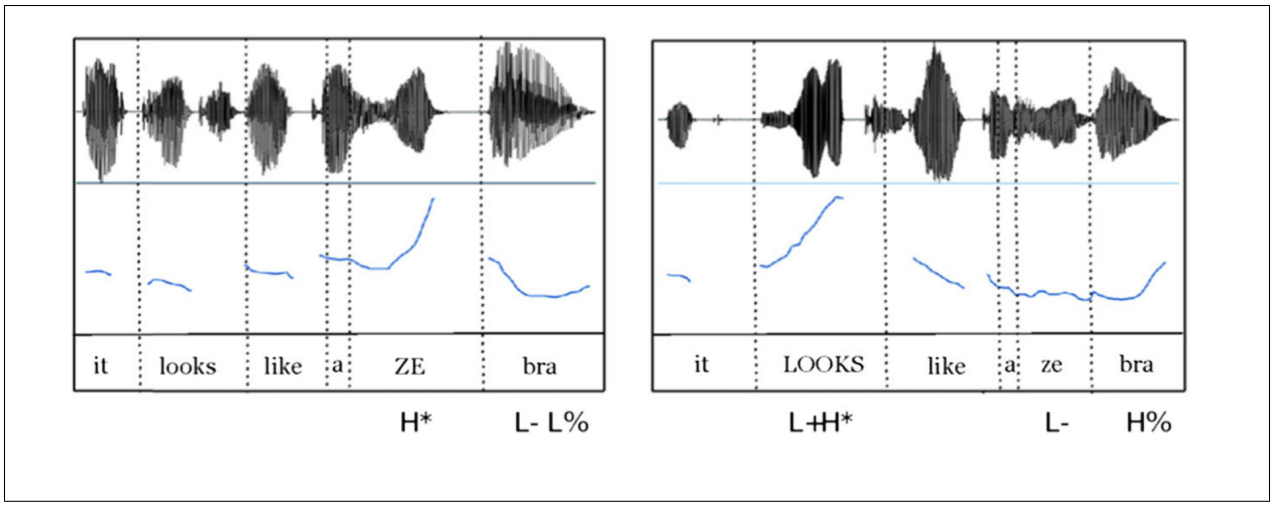

Figure 1: Waveforms and f_0 contours for a noun-focus (left) and a verb-focus (right) production of the sentence "It looks like a zebra."

digitally manipulated 12-step continuum from noun-focus to verb-focus (see Figure 2 for a visual representation). In the experiment's exposure phase, these phrases were disambiguated by a continuation phrase either in the affirmative (It is an X) or in the negative (It is not an X; it just looks like one). There were two exposure conditions. In the no-shift condition, participants heard utterances from the most unambiguously verb-focus end of the continuum disambiguated with negative continuations, and items from the middle of the continuum disambiguated with positive continuations. In the negative-shift group, only unambiguous noun-focus items were disambiguated with positive continuations, while items from the middle of the continuum were disambiguated with negative continuations. The hypothesis, which was supported by the results, was that members of the no-shift group would be more likely to answer in the negative in a subsequent test phase, demonstrating that the categorization function had been recalibrated to interpret ambiguous utterances as verb-focus. The present study sought to replicate this finding using largely the same procedure and materials as Kurumada et al. (2018).

2 Methods

2.1 Participants

One-hundred eight individuals were recruited using Amazon Mechanical Turk. To ensure that there were no repeat participants, each was required to have a unique IP address from within the United States. The expected experiment duration was 6 minutes and participants were compensated \$1.50 (\$15/hour). The average time to completion of the experiment was 6.52 minutes. All speakers reported English as their native language. Participants were also requested to only complete the experiment if they were listening on headphones and performing the task in a quiet environment. Results from 16 participants were excluded due to 'competitor' responses on the two most unambiguously noun-focused trials. These responses were taken to indicate either a fundamental misunderstanding of the task or divided

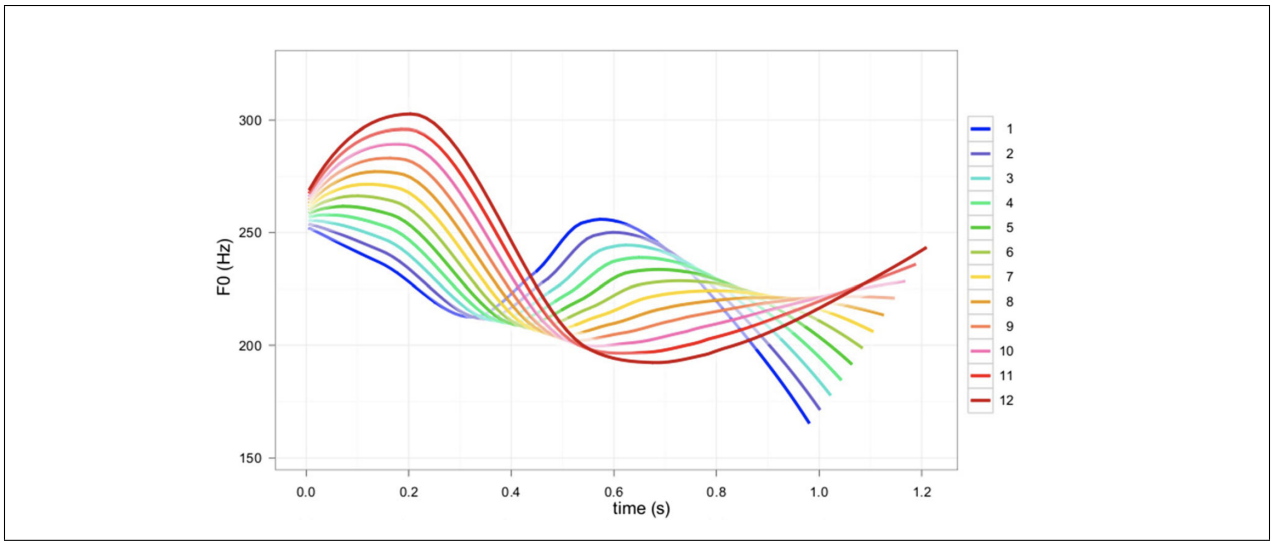

Figure 2: Intonation contours along the continuum from noun-focus (1) to verb-focus (2).

attention. Note that Kurumada et al. (2018) included 324 participants, but due to financial constraints and the robust results of the original study, it was unlikely that this reduction of sample size would endanger the results of the replication.

2.2 Stimuli

Visual and auditory stimuli were procured directly from the authors of Kurumada et al. (2018), although there were several crucial differences between the stimuli used in this study and those that were reported having been used in the original. Most centrally, Kurumada et al. (2018) reported having used prosodic continua containing 12 steps from noun-focus to verb-focus, but the continua provided contained 16 steps. Thus, some reorganization was required, particularly in the exposure stimuli. In the 12-step continuum, the 10th step was found to be most ambiguous in a norming study (i.e., induced a target response closest to 50% of the time. The 13th step of the 16-step continuum is most acoustically similar to the 10th step of the 12-step continuum, and thus stimuli were organized with the 13th step as the fulcrum. Figure 3 shows on the left the distribution of exposure stimuli used by Kurumada et al. (2018) and on the right shows the distribution of exposure stimuli used in the present study. While the relative distribution remains the same, the specific steps used differ slightly.

Auditory stimuli were recorded by a female native speaker of American English. Three recordings were produced for each of the items and with each intonational contour (i.e., noun-focus and verb-focus). Each audio file was subdivided into regions associated with each word or tonal target in cases where a word consists of multiple tonal targets. Each of these regions was then annotated for f_0 at 20 evenly spaced points. A pitch-synchronous overlap-and-add algorithm was used to blend the signals from each of the focus-types at each of these f_0 annotations in Praat (Boersma and Weenink, 2008; Moulines and Charpentier, 1990), thus creating a continuum of 16 evenly spaced prosodic contours. (For more on this process, see Kurumada et al. (2018).) In addition to the critical continua, continuation phrases were recorded by the same native English speaker. These continuation phrases

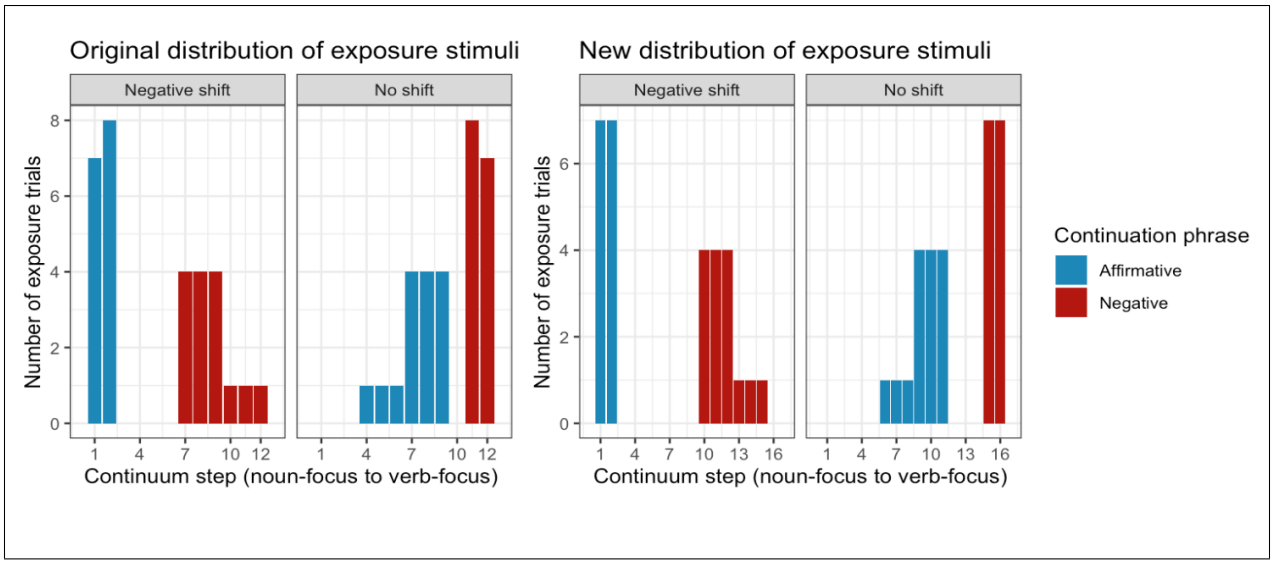

Figure 3: Distributions of exposure stimuli in Kurumada et al. (2018) (left) and the present study (right).

served to disambiguate the appropriate pragmatic inference associated with each intonation contour. For example, a phrase closer to the noun-focus end of the continuum may be followed by an affirmative continuation, such as, "It looks like a zebra . . . *because it has black and white stripes all over its body.*" A phrase closer to the verb-focus end of the continuum would be followed by a negative continuation, such as, "It looks like a zebra . . . *but it's not. It has stripes only on its legs.*" This audio was not processed. To add pragmatic context, each trial was also preceded by a context phrase (either "It's an X!" or "What's that?"), which was produced by an adult male native American English speaker and modified using Praat's 'change gender' feature to approximate the voice of a child.

Visual stimuli were also pulled directly from Kurumada et al. (2018), before which the images were normed. The only exception to this was the images for the target item 'butterfly' and the competitor image 'moth', which were absent from the provided materials. Rather than abandoning these stimuli altogether, suitable replacements were pulled from a public domain stock image database.

2.3 Procedure

The experiment consisted of an exposure phase followed by a test phase. In the exposure phase, participants heard each of 14 trial items twice—once followed by an affirmative continuation and once followed by a negative continuation—resulting in 28 exposure trials. Each exposure trial was preceded by a phrase intended to contextualize the utterance. Specifically, the context phrase was either "What's that?" or "It's an X!" (where 'X' stands in for the target referent). Because "What's that?" biases listeners towards a target interpretation and "It's an X!" biases listeners towards a competitor interpretation, context was mixed for each continuation type in order to prevent listeners from relying more heavily on the context audio than on the prosodic cues. Specifically, 3 of the 14 negative continuation trials were

preceded by "What's that?" and 3 of the 14 affirmative continuation trials were preceded by "It's an X!" After hearing the context audio and the critical audio, participants were asked to respond in a 2AFC task whether they thought the speaker was referring to the target referent or to a competitor by selecting one of two images. Once the participant submitted their answer, the continuation audio played automatically, disambiguating the intended interpretation. Participants each fell into one of two training conditions: no-shift or negative-shift. In the no-shift condition, interpretations were intended to remain approximately the same as they would if participants had not experienced an exposure phase at all. Phrases followed by affirmative continuation phrases fell between steps 6 and 11 of the continuum, while phrases followed by negative continuation phrases fell only on steps 15 and 16. Participants in the negative-shift group heard phrases followed by affirmative continuation phrases only at steps 1 and 2, while phrases followed by negative continuations ranged from step 10 to step 15. See Figure 3 for a more detailed visualization of this distribution.

After completing the exposure phase, participants moved to the test phase, which was identical to the exposure phase except that there was neither context audio nor continuation audio. Rather, participants simply heard the critical phrase, selected an image, and proceeded to the next slide. There were 12 trials, and each included audio from a different step of the continuum. Steps represented in the experiment included 2, 4, 6, and 8–16. Because results reported by Kurumada et al. (2018) indicated that participants were more sensitive to small prosodic manipulations toward the verb-focus end of the continuum, all steps at this end were included. Steps 1, 3, 5, and 7—which were likely to receive relatively consistent affirmative responses from members of both training groups—were excluded due to the previously described limitations of the auditory stimuli. Test phases were identical for members of both training conditions.

3 Results

To reiterate, Kurumada et al. (2018) found that—as was hypothesized—members of the negative-shift group provided more negative responses to trial stimuli than did members of the no-shift condition, particularly as the continuum moved toward unambiguous verb-focus prosody. Their analysis used a multi-level mixed effects logistic regression model with exposure condition and mean-centered continuum step as independent variables and image selection as the dependent variable. The maximum random effects structure was applied (Barr et al., 2013), resulting in random by-item slopes and random by-item intercepts. To the extent possible, the model in the replication was kept as close to that used by Kurumada et al. (2018) as possible. There was one exception in the random effects structure caused by a design flaw in the test phase. Specifically, test items were not rotated between different continuum steps for different participants due to a coding error (e.g. the test item 'spoon' always occurred at the 10th step). This error complicates the data analysis insofar as it creates a one-to-one mapping between items and continuum steps, meaning that random slopes and intercepts for item would make an analysis including step as an independent variable meaningless. Unfortunately, this error was not detected until the experiment had been run, but an analysis including random by-item intercepts but not random by-item slopes was still possible.

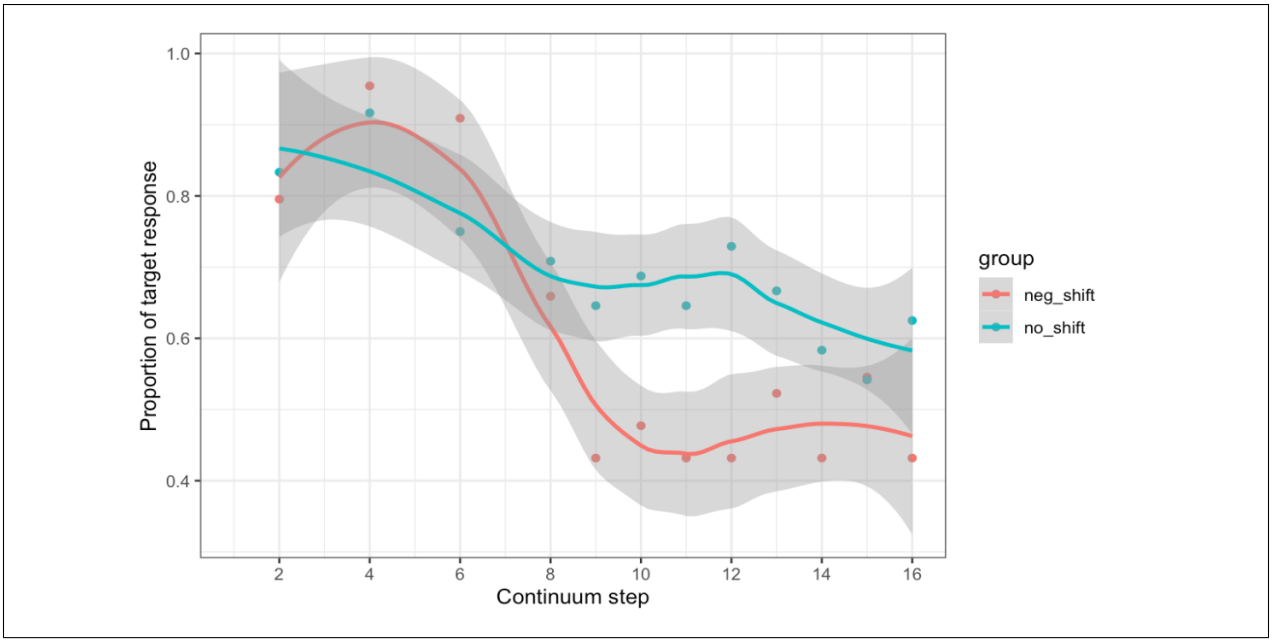

Figure 4: Mean response at each continuum step tested, where 0 corresponds to a 'competitor' response and 1 corresponds to a 'target' response.

Responses are plotted in Figure 4. The model output—run using the lme4 package in R (Bates et al., 2015)—indicated a significant main effect of both exposure condition and continuum step (exposure condition: $\beta = 0.47$, SE $= 0.13$, $p < .001$; continuum-step: $\beta = -0.17$, SE $= 0.03$, $p < .001$). Note that, following Kurumada et al. (2018), the negative-shift group was used as the reference level for the analysis of exposure condition. Thus, the observed positive coefficient is confirmatory of the hypothesis. These results are consistent with the original findings (exposure condition: $\beta = 0.32$, $p < .001$; continuum step: $\beta = -0.18$, $p < .001$; SE not reported). However, there was a small divergence regarding the interaction term. While the original report found a significant interaction between the two independent variables ($\beta = 0.05$, $p < 0.05$; SE not reported), the replication turned up only marginal significance for this interaction ($\beta = 0.06$, SE $= 0.03$, $p < 0.1$). This may be attributable to a lack of power, given that the sample size in the replication was one third of that used in the original experiment, and the p value in the replication model was only marginally greater than that reported in the original model.

4 Discussion

Taken together, these results almost fully replicate the findings of Kurumada et al. (2018). The only difference, as was described above, was the marginal significance of the interaction between exposure condition and continuum step. Due to the smaller sample size and trend-level similarity, this can be taken as an issue of power. The replication of the original results supports evidence for the hypothesis that listeners adapt interpretations of prosodic signals in order to accommodate variability in the input. In other words, listeners who heard more ambiguous intonational contours disambiguated by negative continuations successfully

adjusted subsequent predictions such that they were more likely to interpret ambiguous contours negatively (i.e. as highlighting a contrast rather than referring to the preexisting referent). Because prosodic cues exist on a continuous scale and are employed in inconsistent ways across speakers, this ability of adaptation may be part of what facilitates the ease of communication even between speakers who have never met; prosodic adaptation allows speakers to draw statistical inferences about how an interlocutor may speak in the future based on their past productions.

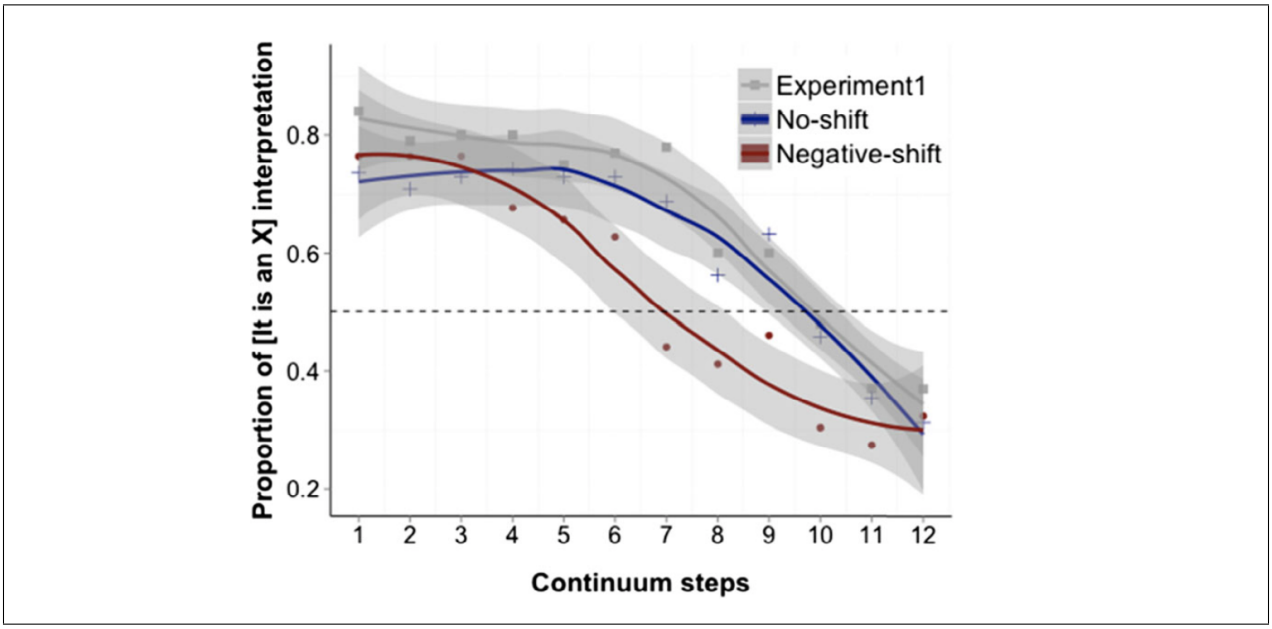

Figure 5: Original findings as visualized by Kurumada et al. (2018). X-axis indicates proportion of affirmative interpretations; y-axis shows continuum step from noun-focus (1) to verb-focus (12).

Despite the statistical similarity between the findings of Kurumada et al. (2018) and this replication, there are differences worth mentioning. First of all, one will note several differences between the contours of mean responses shown in Figure 5 and Figure 4 above. One such disparity is the abruptness of the drop in affirmative responses for members of the negative-shift group in Figure 4, which occurs between step 6 and step 9, with step 8 as the only intermediate level. This contrasts with the contour of mean responses reported by Kurumada et al. (2018), where the curve sloped more gently towards negative responses. Interestingly, in the replication, the point at which negative responses become much more likely is step 9, and the lowest step at which negative-shift participants were trained was step 10. This could be taken to suggest that participants were more likely to provide a negative response if the intonation contour was nearly identical to or more verb-focused than any item they heard in exposure. This contrasts with Kurumada et al.'s (2018) assertion that participants are not learning mappings to specific acoustic cues but rather to underlying distributional properties of phonological representations. Future research would be needed to tease apart these two possibilities.

Another discrepancy is that the results reported by Kurumada et al. (2018) show a drop to about 30% affirmative responses for members of both groups, but participants in the

replication did not display this pattern. Rather, members of the no-shift group dropped to approximately 60% affirmative responses, while members of the negative-shift group dropped to approximately 45%. Does this indicate that members of the negative-shift group were not actually learning to interpret ambiguous contours as introducing a contrast, but rather merely guessing at chance? To answer this question, an exploratory logistic regression model was run including only negative-shift participants and responses to continuum steps 9–16. This model included mean-centered continuum step as an independent variable and response as a dependent variable. Random effects were excluded due to the previously described issue with random effects structure. The results of the model turned out not to be significant ($p > 0.05$), suggesting that negative-shift participants were in fact guessing at chance on these continuum steps rather than employing the knowledge that an ambiguous or verb-focus input can be interpreted contrastively. This may be damaging to the adaptation argument, but these results are purely exploratory. To draw further conclusions, a study with a greater number of participants and an appropriate random effects structure would need to be run.

In addition to further confirmation of the original hypothesis, there are a number of possible directions for future research. One of most crucial regards whether the observed adaptation occurs pre-perceptually or at a later decision stage. This is a question that came up in the literature on perceptual learning in segmental phonology, where Clarke-Davidson et al. (2008) demonstrated using an AXB task—which is less sensitive to decision biases than something like a categorization task—that the shift did in fact reflect a manipulation of category boundaries. A similar methodology could in principle be applied to prosodic cues. Taking further inspiration from the perceptual learning literature, one could ask whether the learning demonstrated by Kurumada et al. (2018) is a general recalibration of categories or applies only to the specific speaker heard during the exposure phase. This line of inquiry could borrow experimental structure from work analyzing the effects of exposure and test phases containing multiple voices (Eisner and McQueen, 2005; Kraljic and Samuel, 2006). Following Kraljic and Samuel (2005), one could also pursue research investigating the stability and decay of prosodic adaptation over time. Are newly learned distributions as strong after 10 minutes? What about 24 hours? Further research could also seek to isolate the role of specific acoustic cues in adaptation. Perceptual learning typically modulates a single acoustic cue as the locus of learning (e.g. center of gravity in Norris et al.'s (2003) study of Dutch fricatives), but Kurumada et al. (2018) manipulated both f_0 and duration. Would these findings be replicated with only a manipulation of one of these cues but not the other? The answer to this question could reveal something about listeners' differential prioritization of various acoustic details in perception.

More generally, investigating the nature of prosodic-phonological representations may be able to illuminate features of the structure of phonological knowledge. In exemplar-theoretic approaches to phonology (e.g. Goldinger, 1998; Johnson, 1997a,b), it is often posited that the word is the basic unit of episodic memory used in speech processing, but little research has been conducted regarding the storage of phrase-level features. That said, an exemplar-theoretic account of prosody would be compatible with the findings of Kurumada et al. (2018), if such a model made space for multiple levels of granularity of episodic representations of language. Drawing from work on prosodic adaptation, it may be possible to derive an extension to exemplar theory that can account for prosody in a more rigorous way.

References

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beckman, M. E. and Pierrehumbert, J. (1986). Intonation structure in Japanese and English. *Phonology Yearbook*.

Boersma, P. and Weenink, D. (2008). Praat: Doing phonetics by computer.

Clarke-Davidson, C. M., Luce, P. A., and Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representations or decision bias? *Perception & Psychophysics*, 70(4):604–618.

Eisner, F. and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67(2):224–238.

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

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics*, 50:101–113.

Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In Johnson, K. and Mullenix, J. W., editors, *Talker variability in speech processing*, pages 145–165. Academic Press, San Diego, CA.

Kraljic, T. and Samuel, A. G. (2005). Perceptual Learning for Speech: Is there a return to normal? *Cognitive Psychology*, 51:141–178.

Kraljic, T. and Samuel, A. G. (2006). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56:1–15.

Kurumada, C., Brown, M., and Tanenhaus, M. K. (2018). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*, 26(4):1153–1160.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In Aronoff, M. and Oehrle, R., editors, *Language Sound Structure*, pages 157–233. MIT Press, Cambridge.

Moulines, E. and 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., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47:204–238.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in communication*, pages 271–311.