Maintenance of Perceptual Information in Speech Perception

Wednesday Bushong (wbushong@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester

T. Florian Jaeger (fjaeger@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester

Abstract

Acoustic and contextual cues to linguistic categories (e.g., phonemes or words) tend to be temporally distributed across the speech signal. Optimal cue integration thus requires maintenance of subcategorical information over time. At the same time, previous work suggests that finite sensory memory or processing capacity strongly limits how much subcategorical information can be maintained (or for how long). We argue that previous work might have over-interpreted the role of these limitations. In two perception experiments, we find no limit in the *ability* to maintain subcategorical information. We also find that maintenance seems to be the default, neither limited to perceptually particularly ambiguous signals, nor a learned strategy specific to our experiment. In contrast, listeners' *decision* for how long to delay categorization, we find, *is* a function of perceptual ambiguity. It is therefore crucial to distinguish between in-principle abilities (even when they reflect default processing), and decisions made within the bounds of those abilities.

Keywords: linguistics; cognitive science; speech recognition; language comprehension

Introduction

One of the most fundamental problems of auditory processing is the transient nature of the acoustic signal; the systems underlying speech perception receive large amounts of information every second. The bounds of sensory memories thus create a pressure to *incrementally* infer abstract linguistic categories (e.g., phonemes and words) from the auditory signal before that information becomes unavailable.

However, much of the information relevant to inferring a particular part of the auditory signal, for example a segment (phoneme), is not contained on the segment itself. For example, one of the main cues to coda stop voicing is duration of the previous vowel (Klatt, 1976). Thus, in order to successfully resolve the voicing of a coda stop, listeners must maintain information about the previous vowel and integrate it with the evidence they receive later. This is opposed to a scheme where the listener removes information about the previous vowel and only maintains some abstract categorical representation that does not include duration information.

Previous work suggests that listeners can indeed maintain and use subcategorical information at least at short timescales. In a classic study, Ganong (1980) found that lexical constraints can influence the perception of a word-initial sound: sounds varying on the /d/-/t/ continuum are perceived to be more /d/-like when presented before "ash" (*dash* is a word while *tash* is not). More evidence that subcategorical information is maintained within a word comes from eye-tracking studies: McMurray, Tanenhaus, and Aslin (2002) found that listeners looked to competitor items like "bear" and "pear" gradiently according to voice onset time (VOT), the cue that distinguishes /b/ and /p/, suggesting that subphonemic information is maintained and used in higher-level processes (for a review of related work, see Dahan, 2010).

However, there are other possible sources of information that follow a target segment or word that occur much later downstream. For example, persevatory co-articulation might spread information over following syllables (Magen, 1997). The identity of later segments might also contain information about earlier segements because of phonotactic dependencies within and across syllables. Even context beyond word boundaries regularly contains information that can help to resolve uncertainty about the input.

A small literature has investigated the extent to which listeners can maintain subcategorical information at longer distances across the word boundary. In a classic study, Connine, Blasko, and Hall (1991) tested whether listeners could maintain subcategorical information about a segment 3 syllables or 6-8 syllables downstream. Participants listened to sentences like "When the ?ent in the fender was well camouflaged, we sold the car." and judged whether the word they heard was *tent* or *dent*. The ? represents a sound that varied along VOT, the primary cue distinguishing between /t/ and /d/. In this example, the later word *fender* semantically biases interpretation of the target word to be *dent*. If listeners can maintain information about the identity of the ?-segment, they should integrate the biasing context into their decisions. Connine and colleagues reported two important findings, both of which have recently been revisited.

First, participants maintained subcategorical information about the ?-segment for 3 syllables: responses reflected both the specific VOT of the segment *and* the contextual bias. After 6-8 syllables responses reflected only VOT, but not biasing context. This finding is often interpreted to demonstrate the limits of subcategorical information maintenance. However, participants were allowed to respond at any point during the sentence; in fact, in the 6-8 syllables condition, participants responded before even hearing biasing context 84% of the time. This leaves open whether participants could not maintain subcategorical information for longer periods of time, or chose to respond early for other reasons.

Secondly, Connine and colleagues report that the context effect was only reliably present at ambiguous VOTs: sounds that were perceptually unambiguously /t/ or /d/ were not integrated with later context. This has been taken to mean that even information maintenance of up to 3 syllables is limited to the special case of perceptually highly ambiguous per-

cepts. This second conclusion, too, however, has to be interpreted with caution. Connine and colleagues measured the context effect in proportion of /t/ vs. /d/ responses. This is problematic (see also Jaeger, 2008): a context effect that is identical across all VOTs when measured in log-odds—i.e., equally large for perceptually clear and perceptually ambiguous VOTs—will results in smaller or insignificant context effects for perceptually clear VOTs when measured in proportions. Crucially, there are *a priori* reasons to believe that the effect should be constant in log-odds (Bicknell, Jaeger, & Tanenhaus, 2016). The analysis conducted by Connine and colleagues thus leaves open whether subcategorical information maintenance is limited to special cases.

A recent study, Bicknell et al. (2016), revisited both of these problems. Bicknell and colleagues replicated Connine et al. (1991) with one minor change to procedure. Participants were required to wait until the end of the sentence to respond, ensuring that they heard the biasing context. Unlike in the original study, Bicknell et al. (2016) analyzed the log-odds of responding /t/ vs. /d/ and found that listeners maintained subcategorical information for both the 3 and 6-8 syllable conditions (see also Szostak & Pitt, 2013 for similar results in a different phonetic contrast). This suggests that there may be an important distinction between listeners' *ability* to maintain subcategorical information and when listeners *decide* to respond.

The idea of a distinction between in-principle abilities and the decision process motivates the present experiments. Our first goal is to replicate the between-experiment comparison across Bicknell et al. (2016) and Connine et al. (1991) within the same paradigm. Anticipating our result, we indeed replicate the contrast, showing that it is important to distinguish between the ability to maintain information and the decision to provide a categorization response. Given that listeners sometimes choose to make a response before receiving additional semantic information, we ask whether subcategorical information maintenance is a default strategy employed by listeners or is specific to experience in our task. Finally, we ask what influences the decision process by investigating the role of perceptual ambiguity on when participants choose to make a response.

In order to answer these questions, we conducted a web-based experiment that closely followed the paradigm of Connine et al. (1991) and Bicknell et al. (2016). Between-participants we manipulated only one aspect of the procedure, holding everything else constant. In the *forced-response* group of participants, they were required to wait until the end of the sentence before making a response. In the *free-response* group, they could make a response whenever they wanted during the sentence. The forced-response group gives us insight into the ability to maintain subcategorical information. The free-response group allows us to ask what drives listeners' decisions to categorize.

Context	Distance	Sentence
Tent-biasing	Near (3 syllables)	When the [t/d]ent in the **forest** was ...
Dent-biasing	Near (3 syllables)	When the [t/d]ent in the **fender** was ...
Tent-biasing	Far (6-8 syllables)	When the [t/d]ent was noticed in the **forest**, ...
Dent-biasing	Far (6-8 syllables)	When the [t/d]ent was noticed in the **fender**, ...

Table 1: Example stimuli from the experiment in each biasing context and distance condition.

Experiment

Participants

We recruited a total of 96 participants from Amazon Mechanical Turk (48 for the forced-response group, 48 for the free-response group). Participants were awarded $3.00 for their participation in a 30-minute experiment.

Materials

Materials were identical across the two participant groups and were modeled on Connine et al. (1991). Table 1 shows example sentences in each context and distance condition. We manipulated context (tent-biasing vs. dent-biasing), distance (near, 3 syllables vs. far, 6-8 syllables), and VOT (10, 40, 50, 60, 70, and 85ms). We chose our range of VOTs based on simulation-based power analyses so as to maximize statistical power to assess the size of the context effect across the VOT continuum, while also ensuring that there were a range of perceptually ambiguous and unambiguous sounds (based on the VOT distributions of our recording speaker). Seven different sentence frames were constructed. Each participant heard each sentence frame in each of the context, distance, and VOT condition combinations, resulting in a total of 168 sentences in the experiment.

Procedure

Participants were instructed to listen to the sentence and report whether they heard *tent* or *dent*. In the forced-response group, participants were instructed to wait until the end of the sentence to make a response. In the free-response group, participants were instructed that they could respond whenever they wish during the sentence after hearing the critical word.

Data Exclusions

We excluded participants who showed no main effect of VOT on their responses from further analysis. That is, these were participants who did not increase their /t/ responses as VOT increased, suggesting that they had faulty audio equipment, did not understand the task, or were otherwise not paying attention. In the forced-response group, this resulted in the removal of nine participants (18.75%) from analysis. In the free-response group, eleven participants (22.92%) were removed. These exclusions hold across all analyses below.

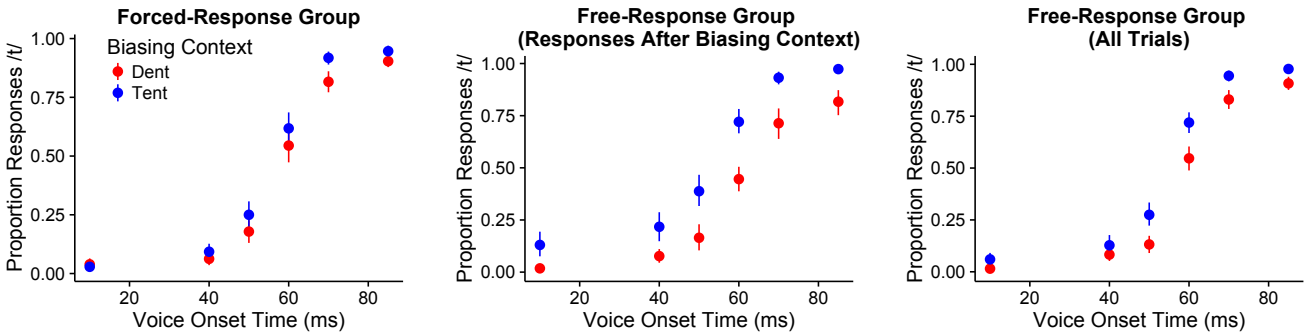

Figure 1: Proportion /t/ responses by biasing context condition for both groups of participants. Error bars are 95% confidence intervals over subject means. See text for discussion of subset vs. all trials.

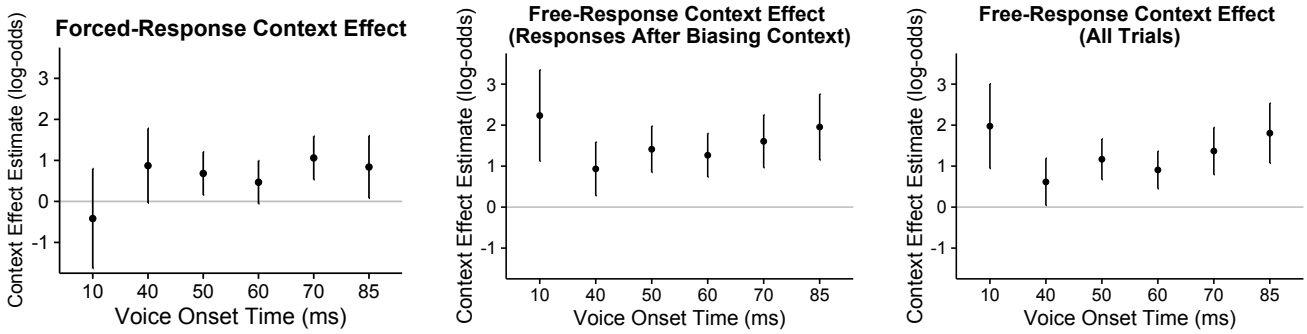

Figure 2: Size of context effect in log-odds space at each VOT for each group as estimated by our simple effects mixed models. Error bars are 95% confidence intervals.

Analysis 1: Limits of Subcategorical Information Maintenance

Analysis

Analyses 1 and 2 are based on the same mixed-effects regression, analyzing the proportion of /t/ responses as a function of VOT (a continuous variable), context, distance, trial number, and their interactions. We included random slopes for context and distance by participants and items (due to data sparsity and the consistency of the VOT effect across participants and items, we did not use the maximal random effects structure; see Bates, Kliegl, Vasishth, & Baayen, 2015). Different predictors in this model answer different questions. In Analysis 1, we focus on the overall effect of context and its interactions with VOT and distance.

In addition to the model described above, in Analysis 1 we also fit a second model which assessed the relative magnitudes of the effect of context at each VOT while removing the potentially problematic assumption that VOTs are related linearly to the log-odds of /t/ responses. This was achieved by recoding the model so as to assess the simple effects of context at each level of VOT. This analysis does thus not *a priori* assume any specific relation between VOTs and /t/ responses.

For each model of the free-response group, we present two analyses. First, we analyzed *only* the trials on which participants responded at least 200ms after offset of the biasing context. This allows a direct comparison with the forced-response group, where participants always responded after hearing the biasing context by design. Second, we also conducted the same analyses using *all* of the data in the free-response group in case these data are more comparable.

Results

Figures 1 and 2 summarize the context effect results of both groups. We found a main effect of VOT on /t/ responses (forced-response: $\hat{\beta} = 0.18, p < 0.001$, free-response subset: $\hat{\beta} = 0.13, p < 0.001$, free-response all trials: $\hat{\beta} = 0.16, p < 0.001$). We also found a context main effect (forced-response: $\hat{\beta} = 1.11, p < 0.001$, free-response subset: $\hat{\beta} = 2.08, p < 0.001$, free-response all trials: $\hat{\beta} = 1.62, p < 0.001$). In the forced-response group and the subset of trials in the free-response group where participants responded after biasing context, there was no interaction between context and distance (forced-response: $\hat{\beta} = -0.09, p = 0.57$, free-response: $\hat{\beta} = 0.18, p = 0.38$). When we analyzed all trials of the free-response group, there was a context x distance interaction such that the context effect was smaller in the far condition ($\hat{\beta} = -0.39, p = 0.02$).

A simple effects analysis revealed that the effect of context was significantly positive at 50ms, 70ms, and 85ms VOT in both groups ($\hat{\beta}s = 0.58 - 1.95$, $ps < 0.05$). In the free-response group, the context effect was also significant at all other VOTs (subset: $\hat{\beta}s = 0.93 - 2.23, ps < 0.01$, all trials: $\hat{\beta}s = 0.61 - 1.97, ps < 0.05$). In the forced-response

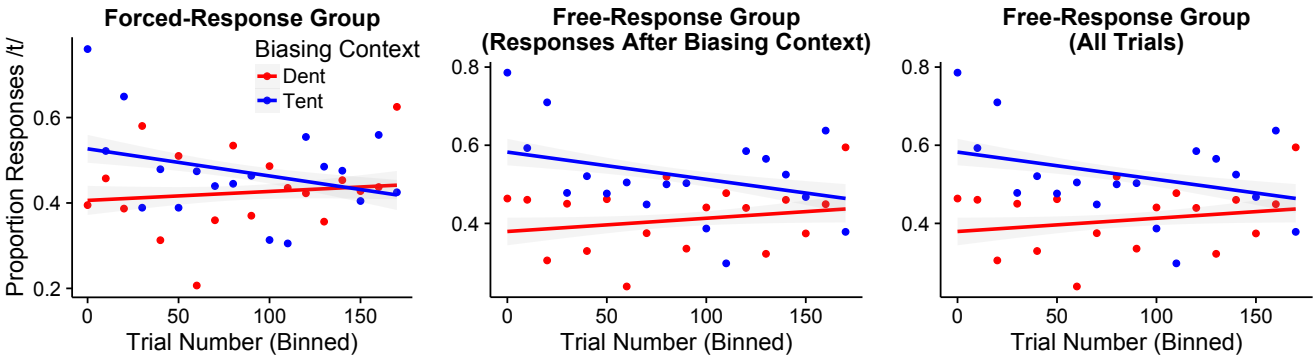

Figure 3: Interaction between context effect over trial for both groups of participants.

group, the context effect was marginal at 40ms and 60ms VOT ($\hat{\beta}s = 0.87, 0.47$, $ps = 0.06, 0.08$), and not significant at 10ms VOT ($\hat{\beta} = -0.42, p = 0.5$).

Discussion

Replicating both Connine et al. (1991) and Bicknell et al. (2016), we found that listeners have the ability to maintain subcategorical information well beyond the word boundary. When forced to wait, participants' responses reflected both the VOT and the contextual bias *even at the longest delay tested* (replicating Bicknell et al., 2016; Szostak & Pitt, 2013). Interestingly, the effect of context seemed more or less constant across the entire range of VOTs tested in both groups. This is exactly as expected by an ideal observer that integrates the perceptual signal with context (Bicknell, Bushong, Tanenhaus, & Jaeger, in preparation). It also suggests that listeners do not necessarily limit the maintenance of subcategorical information to perceptual inputs that are perfectly ambiguous. Instead, it seems listeners maintain subcategorical information even when the perceptual input is already rather unambiguous[1].

When participants were free to choose when to respond, however, we found an interaction between context and distance, such that the context effect was smaller at longer timescales. This would suggest that participants were deciding to respond before hearing biasing context: indeed, the free-response group responded before biasing context on 32% of far trials and 0.5% of near trials (a point we return to in Analysis 3).

Analysis 1 leaves open whether this tendency to maintain subcategorical information is a strategy participants adopt specifically for this experiment, rather than reflecting a more general property of speech perception. Analysis 2 begins to address this question by investigating the context effect across trials.

[1]We note that analyses in VOT space do not tell us about the context effect on the basis of individual participants' subjective perceptual ambiguity, however.

Analysis 2: Subcategorial Information Maintenance: Experimental Artifact or Default Behavior?

We analyze changes in the effect of context over the course of the experiment in both groups. If we observe an effect of context from the very beginning of the experiment, this suggests that listeners maintain subcategorical information by default. On the other hand, if we observe no context effect until later in the experiment, this suggests that listeners have learned to maintain subcategorical information.

Analysis

We used the same logistic regression model from Analysis 1 and focus on the effects of context, trial, and their interaction. Trial was coded so that the coefficient estimate for context reflects the context effect at the very first trial (by subtracting 1).

Results

Figure 3 shows the context effect over trials in both groups of participants. The context effect was significant from the very first trial of the experiment (forced-response: $\hat{\beta} = 1.11, p < 0.001$, free-response subset: $\hat{\beta} = 2.08, p < 0.001$, free-response all trials: $\hat{\beta} = 1.62, p < 0.001$). We found a significant negative interaction between context and trial for both groups of participants (forced-response: $\hat{\beta} = -0.004, p < 0.001$, free-response subset: $\hat{\beta} = -0.009, p < 0.001$, free-reponse all trials: $\hat{\beta} = -0.004, p < 0.001$).

Discussion

Particpants in both experiments exhibited clear context effects right from the beginning of the experiment. This suggests that participants have the ability to maintain subcategorical information without requiring extensive exposure to a particular task. We also found a negative context by trial interaction, such that the context effect got smaller over the course of the experiment. This could mean that participants maintain subcategorical information to a lesser extent as time goes on (e.g., because of fatigue or boredom with the task). Alternatively, participants may still be maintaining subcategorical information but may rely less on context during their

decision making process, and use VOT more (e.g., because participants become more certain of the talker-specific VOT distribution, cf. Kleinschmidt & Jaeger, 2015).

Analysis 3: Strength of Perceptual Evidence and Decision to Categorize

Although maintenance of subcategorical information seems to be a default strategy among participants, the context by distance interaction in the free-response group in Analysis 1 suggests that participants did not necessarily wait for biasing context to make their responses.

This raises questions about what determines when listeners provide a categorization response. If listeners have enough perceptual evidence to confidently make a categorization, they may tend to respond early rather than waiting for the biasing context that provides additional information about the identity of the segment (note that this leaves open whether listeners maintain subcategorical information beyond this point; we return to this below). To answer this question, we analyze *when* participants in the free-response group made responses, and whether this was dependent on the perceptual ambiguity of the stimulus.

Analysis

We used mixed-effects logistic regression to analyze the proportion of responses before biasing context as a function of perceptual ambiguity and distance. For each trial, we coded whether the participant responded before or after having heard biasing context (defined as 200ms after biasing word offset to account for motor planning). To estimate (subjective) perceptual ambiguity, we compute the distance (in probability space) of each VOT from the maximally unambiguous point based on average response probabilities[2]. If strength of perceptual evidence affects when listeners make a decision before obtaining more information (provided by the biasing context), we should see more responses before biasing context for less ambiguous stimuli.

Results

Figure 4 shows proportion of responses before biasing context by perceptual ambiguity of the stimulus. We found a significant effect of ambiguity ($\hat{\beta} = -4.06, p = 0.006$), such that participants were less likely to respond before biasing context when the perceptual stimulus was more ambiguous. We also found a main effect of distance ($\hat{\beta} = 6.93, p < 0.001$) such that participants were more likely to respond before biasing context when it occurred 6-8 syllables away from the target word than when it occurred 3 syllables away. We additionally found a main effect of VOT such that participants were less likely to respond before biasing context as VOTs became longer ($\hat{\beta} = -0.007, p < 0.001$). There were no other main effects or interactions.

[2]This perceptual ambiguity measure can also be computed on a by-subject basis and does not change the results.

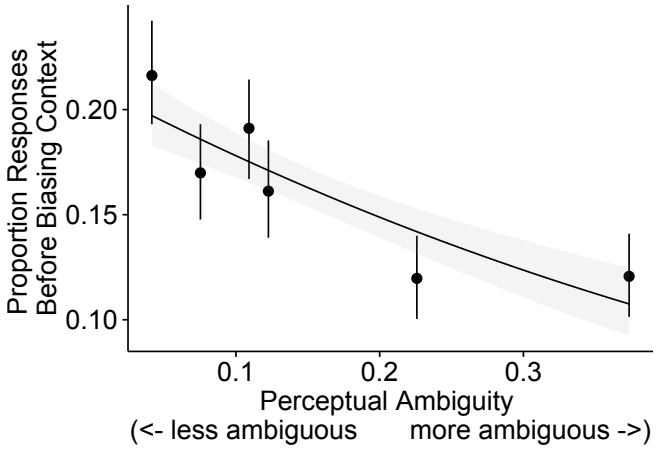

Figure 4: Proportion of responses before biasing context by perceptual ambiguity. Error bars are 95% confidence intervals.

Discussion

We found that participants were more likely to respond before hearing biasing context when the perceptual signal was less ambiguous, and when biasing context appeared farther away from the target word. We also found a main effect of distance: participants were more likely to respond before biasing context when it occurred farther away from the target word. These results suggest that while listeners have the ability to maintain subcategorical information for unambiguous stimuli over long distances, when given a choice listeners decide to respond earlier when they have stronger perceptual evidence for categorization.

General Discussion

Together, our results suggest that in principle, listeners can maintain subcategorical information well beyond word boundaries. Listeners seem to do so by default, and both for ambiguous and unambiguous percepts. This suggests that the limits of listeners ability to maintain subcategorical information are less strict than previously assumed (Connine et al., 1991; Christiansen & Chater, 2016). At the same time, listeners do not wait arbitrarily long for additional informative context. When given the opportunity, listeners responded on 16% of all trials before additional context could aid recognition. Critically, listeners' decisions to respond early were not arbitrary, but rather systematically conditioned on the ambiguity of the perceptual input: listeners were more likely to respond before biasing context when the perceptual signal was less ambiguous. This strategy seems to vary little across participants.

Three questions stand out to us as requiring further attention. First, importantly, little is known about *what* kind of information is being maintained. It is possible that listeners retain a rich representation of the original percept, some more abstract representation of their certainty in the identity of the segment, or something in between.

Second, it is unclear what becomes of these representations after listeners make a perceptual decision. It could be the case that the maintenance process and decision-making process are dependent on *or* independent of each other. The large literature on exemplar-based approach to speech perception suggests that exemplars are stored and used later in speech perception (Hay & Drager, 2010; Strand & Johnson, 1996; Goldinger, 1997). The apparent storage of this low-level information in long-term memory is puzzling if there are strict limitations on the amount of information that can be maintained during speech perception—a paradox that has, to the best of our knowledge, received surprisingly little attention.

Third, we found evidence that the maintenance of subcategorical information in the present experiments does not seem to be learned over time in a task-specific manner. It is, however, an open question whether listeners *can* flexibly adapt the degree to which (or duration for which) they maintain subcategorical information, depending on their goals or the structure of the current task. Such flexibility would suggest that listeners' decisions about at which point to categorize input might more often be constrained by the goal to quickly infer the meaning-bearing message, rather than being constrained by strong limits of perceptual memory. For example, it is possible that the limits (or lack thereof) of maintenance observed in experiments like ours (and a large body of previous work; for review, see Dahan, 2010) reflect participants' beliefs based on previous experience about the expected utility of delaying categorization. In that case, listeners might adapt these beliefs after exposure to stimuli that contain or do not contain helpful contextual information.

Acknowledgments

This work was partially funded by NSF NRT #1449828 (graduate stipend to W.B.) and NSF IIS-1150028 and NIHCD R01 HD075797 (to T.F.J.). The views expressed here do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. F. (in preparation). Listeners can maintain and rationally update uncertainty about prior words.

Bicknell, K., Jaeger, T. F., & Tanenhaus, M. K. (2016). Now or ... later: Perceptual data is not immediately forgotten during language processing. *Behavioral and Brain Sciences*, *39*, 23–24.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, *30*(2), 234–250.

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121–126.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. *Talker variability in speech processing*, 33–66.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892.

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.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in english. *Journal of Phonetics*, *25*(2), 187–205.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In *Konvens* (pp. 14–26).

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533–1546.