Modeling Long-Distance Cue Integration in Spoken Word Recognition

Wednesday Bushong
Brain & Cognitive Sciences Department
University of Rochester
wbushong@ur.rochester.edu

T. Florian Jaeger
Brain & Cognitive Sciences Department
University of Rochester
fjaeger@ur.rochester.edu

Abstract

Cues to linguistic categories are distributed across the speech signal. Optimal categorization thus requires that listeners maintain gradient representations of incoming input in order to integrate that information with later cues. There is now evidence that listeners can and do integrate cues that occur far apart in time. Computational models of this integration have however been lacking. We take a first step at addressing this gap by mathematically formalizing four models of how listeners may maintain and use cue information during spoken language understanding and test them on two perception experiments. In one experiment, we find support for rational integration of cues at long distances. In a second, more memory and attention-taxing experiment, we find evidence in favor of a switching model that avoids maintaining detailed representations of cues in memory. These results are a first step in understanding what kinds of mechanisms listeners use for cue integration under different memory and attentional constraints.

1 Introduction

Language is a fast, temporally unfolding signal. Humans must quickly compress large amounts of information into abstract linguistic representations and meanings that contain more manageable amounts of information. However, cues to linguistic categories often do not temporallly co-occur, but are distributed quite broadly across the signal. Rational information integration thus requires maintenance of gradient subcategorical information so as to integrate cues that occur at different points in time. For example, one of the primary cues to the voicing of a syllable-final stop consonant in English is the duration of the *preceding* vowel (Klatt, 1976). Thus, in order to obtain a good estimate of the voicing of a syllable-final stop, listeners must retain some subcategorical information about the preceding vowel in memory. This is typical across languages and occurs at multiple timescales: cues to sound categories can come not only from proximate acoustic properties, but also from, e.g., later semantic context that could potentially occur an unlimited distance away from the target. This poses a memory challenge for language comprehenders: how can one possibly maintain subcategorical information for later use when such maintenance should overload working memory?

This challenge has motivated theories of language processing that contend that listeners compress input into abstract representations as quickly as possible and discard all gradient information after a categorical perceptual decision has been reached (Just and Carpenter, 1980; Christiansen and Chater, 2016). According to these accounts, listeners cannot maintain gradient sub-categorical information for cue integration at any significant timescale, at certainly not beyond word boundaries. However, a growing body of literature has suggested that listeners are in principle capable of maintaining subcategorical representations (McMurray et al., 2009), including at timescales beyond the word boundary (Connine et al., 1991; Brown-Schmidt and Toscano, 2017; Gwilliams et al., 2018). For example, Connine et al. (1991) exposed participants to sentences that contained two cues about a target word, "tent" or "dent" in the sentence. The first cue was the voice-onset time (VOT) of the first sound in the word, which was varied to form a continuum from more /t/-like to more /d/-like. The second cue was a subsequent word that contextually biased toward either the "tent" interpretation (e.g., "campground") or the "dent" interpretation (e.g., "teapot"). Participants heard sentences like "When the ?ent Sue had found in the [campground/teapot]...", and were asked to categorize whether they heard the word

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 62–70
Minneapolis, USA, June 7, 2019.

"tent" or "dent" in the sentence. They found that participants' categorizations were influenced *both* by the VOT of the sound *and* by subsequent context, suggesting that listeners maintained a gradient representation of the initial sound for later use in cue integration and categorization. Subsequent studies have confirmed that listeners can maintain subcategorical representations well beyond word boundaries (Szostak and Pitt, 2013; Bushong and Jaeger, 2017).

Despite recent interest in this phenomenon, to date there has been no comprehensive effort to spell out and quantitatively compare different models of long-distance cue integration under different memory/information constraints. This paper is a first attempt to explore this space, driven largely by previous conceptual proposals. We consider four different models that vary in the extent to which they maintain sub-categorical information and utilize multiple time-distant cues. Two of the models maintain subcategorical information about cues over time, and two do not.

These four models make distinct quantitative and qualitative predictions about how human categorization judgments should be affected by two cues. We first present the mathematical models along with their predictions. We then evaluate the models against human data from two behavioral experiments. In these experiments, participants hear sentences like those in Connine et al. paradigm. We manipulated the same two types of cues as in the Connine et al. paradigm (i.e., VOT and subsequent semantic context).

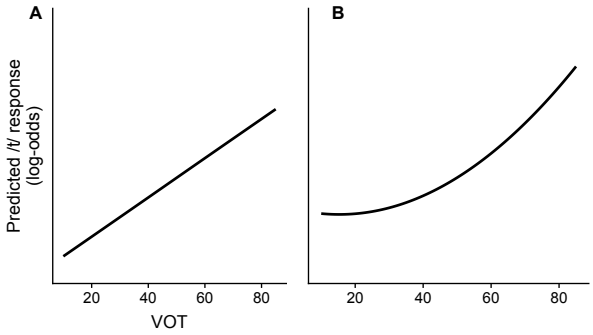

Figure 1: Linear (**A**) vs. quadratic (**B**) effect of VOT on log-odds of "t"-responses.

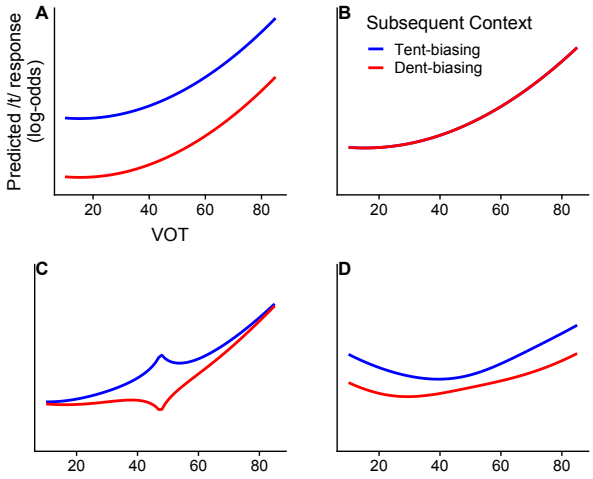

Figure 2: Qualitative predictions of each model in log-odds of "t"-responses (for a context bias of 0.95). **(A):** ideal integration model, **(B):** categorize-&-discard model, **(C):** ambiguity-only model, **(D):** categorize-discard-&-switch model. Shown predictions assume a quadratic effect of VOT (but predicted context effects are identical even if VOT has a linear effect).

2 Models

We first describe how an ideal observer would categorize stimuli based on the first cue alone (VOT). Then we describe the four potential models of cue integration, along with their predictions. Figures 2 and 3 illustrate these predictions. Predictions are shown with regard to log-odds (of a "t"-response), since the predictions of all four models look (misleadingly) similar in proportion space. The prediction plots are meant as qualitative demonstrations. For example, the predicted slope of the VOT effect depends on listeners' beliefs about the means and variances of the /t/ and /d/ categories along the VOT continuum. Similarly, the specific magnitude of the context effect depends on the bias (or information) provided by context and the perceptual uncertainty about the VOT cue. Regardless of these details, however, some qualitative differences in the context effect emerge across the four different models (see Figure 3). It is these predicted shapes of the context effect that we later compare against human responses from perception experiments.

For all predictions, we assume Luce's choice rule for the link between models' posterior probability of /t/ and the predicted decision to respond "t" or "d"—i.e., $p_{model}(\text{respond "t"}) = p_{model}(/t/|context, VOT)$

2.1 Ideal Observers: Predicting VOT Effects

Before we introduce our models of *cue integration*, we first spell out an ideal observer's predictions for the effect of VOT in the absence of any

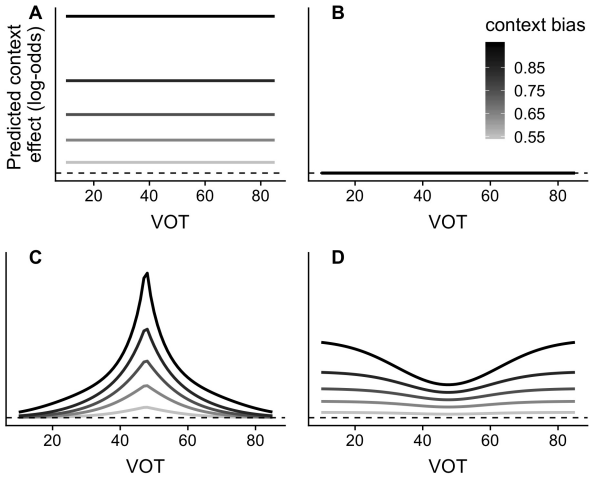

Figure 3: Predicted context effect (difference between blue and red line in Figure 3) for different possible context biases. **(A):** ideal integration model, **(B):** categorize-&-discard model, **(C):** ambiguity-only model, **(D):** categorize-discard-&-switch model. Dashed line represents 0.

second cue. If two Gaussian categories (/t/ and /d/) along VOT have equal variance, an ideal observer will exhibit linear effects of VOT on the log-odds of "t"-responses (Figure 1). However, it is well established that voicing contrasts (including /t/ vs. /d/) exhibit unequal variances along the VOT continuum (Lisker and Abramson, 1967). A standard ideal-observer model thus predicts quadratic effects of VOT on the log-odds of "t"-responses. We thus will visualize all of our model predictions with an assumed quadratic effect of VOT. Next, we turn to the four models of cue integration. We emphasize, however, that the predicted effect of context—the effect we test below—does not depend on this assumption.

2.2 Ideal Integration

The *ideal integration model* holds that listeners maintain subcategorical information about the temporally first cue (here, VOT) in memory for subsequent integration with a later cue (here, context). Note that we use the term 'ideal' in the sense of rational cue integration frameworks proposed across the literature (Ernst and Banks, 2002). These normative models, like the ideal integration model, provide an optimal baseline against which to compare human behavior. The ideal integration model always maintains subcategorical (gradient) information about VOT because optimal categorization requires access to at least $P(category|VOT)$ (or richer information about VOT) during integration with context. Specifically, ideal integration predicts *additive* effects of the two cues on the log-odds of categorization (Bicknell et al., under review).

If humans have no memory constraints and perfectly integrate all cues available to them, their behavior should resemble predictions of the ideal observer; that is, "t"-responses should be conditioned on both VOT and context:

$$p_{ideal}(\text{respond "t"}) = p(/t/|VOT, context) \quad (1)$$

We can apply Bayes' Theorem to arrive at the following:

$$p(/t/|VOT, context) = \frac{p(VOT|context, /t/)p(context, /t/)}{p(VOT, context)} = \frac{p(VOT|context, /t/)p(/t/|context)}{p(VOT|context)} \quad (2)$$

Under the plausible assumption that VOT and context are conditionally independent (as in Bicknell et al., under review)[1]:

$$p_{ideal}(\text{respond "t"}) \propto p(VOT|/t/)p(/t/|context) \quad (3)$$

As shown in Figure 2A and 3A, the ideal integration model predicts additive effects of VOT and subsequent context in log-odds space.

2.3 Ambiguity-Only

In contrast to the ideal integration model, the *ambiguity-only* model stores information about VOT to the extent to which VOT is perceptually ambiguous: the more ambiguous VOT is, the more likely listeners should be to maintain information about VOT for subsequent integration with context. The ambiguity-only hypothesis—first proposed by Connine et al. (1991)—thus sees maintenance of subcategorical information as a special case: if the signal is relatively clear then listeners immediately categorize and discard low-level information; only when the perceptual input is ambiguous is information about it maintained in memory so as to facilitate robust integration with

[1] In our descriptions of the remaining models, we will use $p(/t/|VOT, context)$ and $p(/t/|VOT)$ as shorthand rather than fully expanding them using Bayes' Theorem as in this initial example.

subsequent cues. This can be seen as serving memory economy (for related proposals, see also Dahan, 2010).

There are several ways of operationalizing the idea that information about VOT is only maintained if VOT is perceptually ambiguous. Here, we evaluate a gradient version of this hypothesis: with increasingly unambiguous VOT evidence—i.e., for $p(/t/|VOT)$ closer to 0 or 1—, listeners are assumed to be less likely to maintain gradient representations of VOT to integrate with later context, instead categorizing on the basis of VOT alone. As VOT becomes more ambiguous—$p(/t/|VOT)$ closer to 0.5—, listeners are assumed to be more likely to maintain gradient representations for later integration. We can quantify the degree of perceptual ambiguity as:

$$\lambda = 2|p(/t/|VOT) - 0.5| \quad (4)$$

We note that λ is determined by the perceptual ambiguity of the stimulus and does not constitute a free parameter for this model. We can then use λ as a weight in a mixture model that describes the relative probability of using VOT only or integrating VOT and context:

$$p_{ambiguity}(\text{respond “t”}) = \lambda p(/t/|VOT) + (1-\lambda)p(/t/|VOT, context) \quad (5)$$

Intuitively, we can think of this as listeners *not* maintaining gradient representations of VOT on λ proportion of trials, and maintaining gradient representations on the remaining proportion.

2.4 Categorize-&-Discard

The other two models we consider do *not* maintain information about VOT in memory, but rather immediately categorize based on the first cue and then discard all subcategorical information about that cue. These *categorize-&-discard* models maximize memory economy at the cost of integration accuracy. Categorize-discard models thus capture the influential view that prolonged maintenance of subcategorical information about the speech signal is not feasible given the bounds of the relevant memory systems (see, e.g., Christiansen and Chater, 2016). The most simple *categorize-&-discard model* categorizes based on VOT, discards all subcategorical information about VOT, and then never revisits the categorization decision. As this model never considers the second source of information (VOT), its categorization accuracy will necessarily be suboptimal. We formalize this model as simply making decisions on the basis of VOT alone:

$$p_{cat\_discard}(\text{respond “t”}) = p(/t/|VOT) \quad (6)$$

2.5 Categorize-Discard-&-Switch

The final model we consider also discards all subcategorical information about VOT immediately after having used it to categorize. However, unlike the category-discard model, the *categorize-discard-&-switch* model has a mechanism to take into account context: if context conflicts with the initial categorization decision, the model will change its categorization response in proportion to the evidence from context. Concretely, if the model initially categorizes a segment as /d/, but later evidence from context is more consistent with /t/, the model will switch to /t/ in proportion (over trials) to how strongly context points toward the alternative categorization. While the categorization accuracy achieved by the categorize-discard-&-switch model is better than that of the simpler categorize-&-discard model, it is still suboptimal (i.e., underperforms compared to the ideal integration model).

$$p_{cat\_switch}(\text{respond “t”}) \propto p(/t/|VOT) + (1-p(/t/|VOT))p(/t/|context) \quad (7)$$

Like the ambiguity-only model, w can think of this as a cross-trial description of the outcomes of categorization. On some proportion of trials $p(/t/|VOT)$, listeners would have categorized a stimulus as /t/ based on VOT alone. On the remaining trials where listeners would have made a /d/ categorization based on VOT alone, they sometimes switch, proportional to the evidence from context.

The categorize-discard-&-switch model is of particular relevance in light of the recent findings of Bicknell et al. (under review). In their comparison of the ideal integration model with the ambiguity-only model, Bicknell et al. (under review) found no evidence that perceptually less ambiguous VOTs were associated with smaller effects of subsequent context. Rather, the human data seemed to support a constant effect of subsequent context across the entire VOT spectrum. If

anything, some of the behavioral data considered by Bicknell et al. (under review) contained numerical trends towards *larger* effects of subsequent context for perceptually less ambiguous VOTs. As can be seen in Figures 2D and 3D, such a pattern would be predicted by the categorize-discard-&-switch model. In order to put the hypothesis of ideal integration to a stronger test, it is thus necessary to compare the ideal integration model also against the new plausible competitor we have identified, the categorize-discard-&-switch model. Next, we describe the two perception experiments that we use to model human responses.

3 Behavioral Experiments

The human data we analyze here stem from two experiments originally reported in Bushong and Jaeger (under review). In both experiments, participants are exposed to sentences and have to make categorization judgments about a target word in the sentence. We varied a critical word in the sentence to vary acoustically between "tent" and "dent", and a subsequent word in the sentence provides a contextual bias relevant to the critical target word (e.g., "campgrounds" biases towards a "tent" interpretation over a "dent" interpretation). The critical difference between the two experiments is which words participants needed to categorize. In Experiment 1, participants always were asked to make categorization decisions about our critical target words, "tent" and "dent". In Experiment 2, this was only their task on half of the trials; on the other half, they were asked to categorize a different word in the sentence that was neither our critical target word nor the subsequent contextually biasing word (see Figure 4). The basic conceptual difference here is that in Experiment 1, it is relatively easy for participants to ideally integrate cues: they always know which cue they need to maintain a gradient representation of (i.e., the initial sound of the target word). Experiment 2 increases the memory and attentional burden of maintaining gradient representations, however: now participants have several possible words they could be asked about and thus cannot perfectly predict which parts of the signal will be relevant for the task. We hypothesized that structure of Experiment 2 might bias participants towards discarding subcategorical information about the speech input (like the categorize-&-discard and categorize-discard-&-switch models).

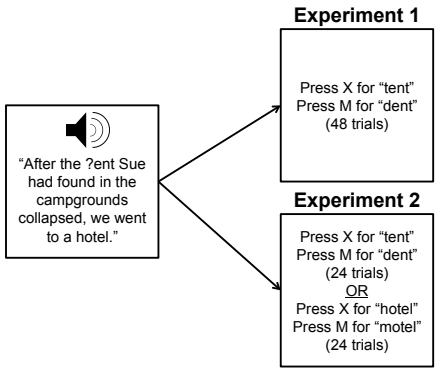

Figure 4: Visualization of an example trial.

3.1 Participants

We recruited 128 native English-speaking participants from Amazon Mechanical Turk for each experiment who were rewarded $3.00 for their participation in the experiment. No participants completed both Experiment 1 and Experiment 2.

3.2 Materials

We take the paradigm from Bushong and Jaeger (2017) as a starting point for our experiments. We constructed 12 sentence pairs like the following:

(1) After the ?ent Sue had found in the campgrounds collapsed, we went to a hotel. (tent-biasing context)

(2) After the ?ent Sue had found in the teapot was noticed, we threw it away. (dent-biasing context)

We manipulated two aspects of the sentence stimuli. First, we acoustically manipulated the "?" to range between /d/ and /t/ by changing the value of its voice-onset time (VOT), the primary cue distinguishing voiced from voiceless consonants. Based on norming and previous experiments, we chose to test VOT values of 10, 40, 50, 60, 70, and 85ms to cover a perceptual range from unambiguous /d/ to unambiguous /t/ with ambiguous points in between. Second, we manipulated whether later context biased toward a /t/-interpretation, /d/-interpretation, or neither. The onset of informative context words were between 6-9 syllables after target word offset.

3.3 Procedure

Both experiments were split into two phases: Exposure (72 trials) and Test (48 trials). The original purpose of these experiments was to test a

Experiment 1	Likelihood Ratio Test		Bayesian Analysis	
Comparison	χ^2	p	Bayes Factor	Posterior Probability
Analysis 2 vs. Analysis 1	38.78	< 0.001	3.5×10^6	> 0.999
Analysis 3 vs. **Analysis 2**	3.76	0.15	0.001	0.001
Experiment 2	Likelihood Ratio Test		Bayesian Analysis	
Comparison	χ^2	p	Bayes Factor	Posterior Probability
Analysis 2 vs. Analysis 1	71.23	< 0.001	5.66×10^{13}	> 0.999
Analysis 3 vs. Analysis 2	40.07	< 0.001	1.9×10^5	> 0.999
Analysis 3 vs. Analysis 3 control	39.27	< 0.001	6.5×10^6	> 0.999

Table 1: Model comparisons for Experiments 1 and 2, both in terms of likelihood ratio tests and Bayes Factor. Best-fitting model is bolded for each experiment.

particular relationship between exposure and test in a between-subjects manipulation (see Bushong and Jaeger, under review). The difference between the experimental groups is that one group of subjects heard sentences with no subsequent biasing context in the exposure phase, while the other group always heard sentences with subsequent context. Because of this imbalance between groups, we only analyze data from the test phase which was identical across participants[2]. What is important here is that in the test phase, all participants heard sentences that contained the full range of our VOT manipulation (evenly split between all values) and informative later context (split evenly between /t/-biasing and /d/-biasing contexts). Test sentences crossed all 6 steps of our VOT continuum with the two context conditions (/t/-biasing and /d/-biasing). All 12 combinations of VOT and context occurred equally often, so as to allow us to reliably estimate the effect of context across the VOT continuum.

Participants' task was simply to categorize whether they heard the word "tent" or "dent" after they heard the full sentence. In Experiment 1, this task was constant across all trials. In Experiment 2, on half of all trials, participants instead had to categorize another word in the sentence (e.g., for sentence (2) above they were asked whether they heard "hotel" or "motel"). Figure 4 shows the structure of the two experiments.

4 Analysis

Following previous work (Bushong and Jaeger, 2017), we excluded participants whose categorization responses were not modulated by VOT, suggesting that they did not understand the task. This resulted in the exclusion of 11 participants from Experiment 1 (8.6%) and 16 participants from Experiment 2 (12.5%).

We fit mixed-effects logistic regression analyses predicting the log-odds of /t/ responses in the test phase from predictors of interest. Regressions were fit using the `lme4` package in `R` (Bates et al., 2014). Each analysis contained the maximal random effects structure that resulted in successful model convergence. We fit four different types of analyses to each of the two experiments in order to assess each of the models outlined above:

Analysis 1: /t/ response $\sim$ VOT + VOT^2. This analysis represents the categorize-&-discard model, where participants only categorize based on VOT then immediately discard information (and thus do not integrate the subsequent context cue).

Analysis 2: /t/ response $\sim$ VOT + context + VOT^2. This analysis represents the ideal integration model, where participants optimally integrate VOT and context (i.e., use both with no interaction).

Analysis 3: /t/ response $\sim$ VOT*context + VOT^2*context. This analysis represents both the ambiguity-only and categorize-discard-&-switch models. Both models predict that there is a quadratic interaction between VOT and context. A negative quadratic coefficient supports the ambiguity-only model, and a positive coefficient supports the categorize-discard-&-switch model.

Analysis 3 control: /t/ response $\sim$ VOT*context + VOT^2. Since both a linear and squared interaction between VOT and context are necessary to support the ambiguity-only and categorize-discard-&-switch models, we fit an additional control model with only a linear

[2] Additionally, not all combinations of VOT and context were tested in the exposure phase for the group that did hear subsequent context.

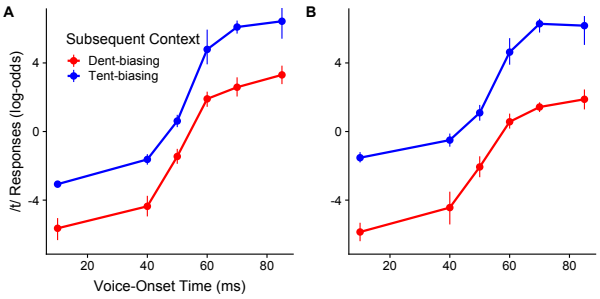

Figure 5: Log-odds of /t/-categorizations in Experiments 1 (**A**) and 2 (**B**) by VOT and subsequent context. Error bars are 95% confidence intervals over item means.

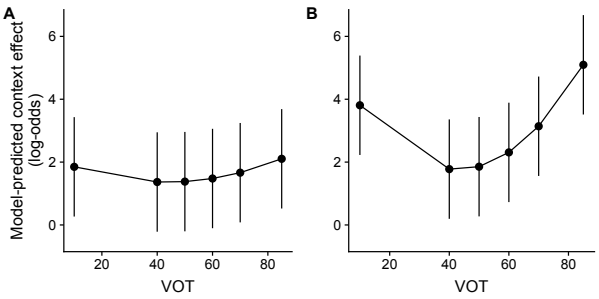

Figure 6: Model-predicted context effect (in log-odds) from Analysis 3 for Experiment 1 (**A**) and Experiment 2 (**B**). Error bars are 95% confidence intervals.

interaction between VOT and context. Thus, for us to conclude that the ambiguity-only or ategorize-discard-&-switch models have support, Analysis 3 must be a better fit compared to Analysis 2 *and* Analysis 3 must be a better fit compared to Analysis 3 control.

Note that both the ambiguity-only and categorize-discard-&-switch models also predict an overall smaller context effect, compared to the ideal integration model (see Figure 3). Additionally, the categorize-discard-&-switch model also predicts a more shallow slope for the VOT effect, compared to all other models (see Figure 2). However, the test of these more specific predictions would require precise knowledge of listeners' beliefs about both a) the distribution of VOT for /t/ and /d/, and b) the exact strength of the context cue. Since we do not have access to this information, we instead take advantage of the qualitative differences in predictions captured by Analyses 1-3.

To determine which models were the best fit for each experiment, we conducted two kinds of model comparisons between the analyses. First, we conducted standard likelihood ratio tests between each pair of models. We additionally derived Bayes Factor (BF) and posterior probability estimates by comparing the BICs of pairs of models (see Wagenmakers, 2007).

Table 1 shows the results for Experiments 1 and 2. The results of the likelihood ratio tests and the Bayesian analysis support the same conclusions.

5 Results

5.1 Experiment 1

Analysis 2 (corresponding to the ideal-integration model) was the best fit both by standard likelihood ratio tests and Bayes Factor. Within Analysis 2, we found significant effects of z-scored VOT ($\hat{\beta} = 1.43, z = 5.72, p < 0.001$), z-scored squared VOT ($\hat{\beta} = 2.43, z = 6.62, p < 0.001$), and subsequent context ($\hat{\beta} = 0.8, z = 6.67, p < 0.001$).

5.2 Experiment 2

Analysis 3 was the best fit to the Experiment 2 data both by standard likelihood ratio tests and Bayes Factor. Within Analysis 3, we found effects of z-scored VOT ($\hat{\beta} = 0.73, z = 2.5, p = 0.01$), z-scored squared VOT ($\hat{\beta} = 1.67, z = 4.873, p < 0.001$), subsequent context ($\hat{\beta} = 1.28, z = 9.75, p < 0.001$), and an interaction between z-scored squared VOT and susbequent context ($\hat{\beta} = 0.23, z = 2.494, p = 0.01$).

5.3 Discussion

Both experiments return clearly significant effects of squared VOT. This is predicted by an ideal observer model, since the /t/ and /d/ categories have unequal variance along the VOT continuum (Lisker and Abramson, 1967). With regard to the question of ideal *integration*, the results differ between the two experiments.

In Experiment 1, we found strong evidence for the ideal integration model: participants displayed effects of VOT and context, with no interaction between these factors. This suggests that participants were able to maintain gradient representations of VOT to later integrate with our contextual cue.

In Experiment 2, we found strong evidence for the categorize-discard-&-switch model: participants showed effects of VOT and context, but also showed a positive interaction between squared VOT and context such that the context effect was largest at the endpoints and smallest at the most

ambiguous points. These results suggest that participants in Experiment 2 took a more memory-efficient strategy where they did not maintain gradient information about VOT but were still able to use both relevant cues in categorization.

6 General Discussion

Language is a signal that carries thousands of bits of acoustic information per second that listeners need to somehow compress into categorical abstract representations. However, maintaining some sub-categorical detail about the original signal in memory in order to integrate it with later potentially relevant cues is beneficial for achieving optimal categorization. Several lines of work have suggested either that this kind of integration is severely limited by time (Christiansen and Chater, 2016), the ambiguity of the initial signal (Connine et al., 1991), or is actually optimal and not very constrained by time or ambiguity (Bicknell et al., under review). However, these proposals have not been formalized and tested in a rigorous way (but see Bicknell et al., under review, for a discussion of ideal observers and one formalization of the ambiguity hypothesis). Here, we took a first step toward understanding and testing these three proposals.

We enumerated four possible models for the integration of cues that occur at different points in the speech signal. Two of these models involve maintaining gradient representations of the initial speech cue in memory for later integration with the subsequent cue, either being fully optimal (the ideal integration model), or partially restricted by ambiguity of the first cue (the ambiguity-only model). The other two models reduce the burden on memory by not maintaining gradient information about the initial speech cue, either by immediately categorizing and ignoring later cues (the categorize-&-discard model), or potentially changing categorization if later information conflicts with the initial binary categorization (the categorize-discard-&-switch model).

In Experiment 1, we found strong evidence for the ideal integration model, in line with previous work (Bicknell et al., under review; Szostak and Pitt, 2013). Experiment 2 added a manipulation that made it more difficult for participants to predict which words they would need to attend to in our sentences. When we introduced this manipulation, we interestingly found strong support for the categorize-discard-&-switch model, suggesting that listeners were not maintaining subcategorical information about initial speech cues in memory. This finding is particularly noteworthy since the categorize-discard-&-switch model has not been previously considered in the literature as a possibility for cue integration during language processing. Significantly, in neither experiment did we find any evidence for the ambiguity-only model, which has been the primary proposal for how subcategorical information is maintained (Connine et al., 1991; Dahan, 2010).

Our results here suggest that listeners behave like ideal integrators under the task demands of typical right-context studies in the literature (Connine et al., 1991; Szostak and Pitt, 2013; Bushong and Jaeger, 2017; Bicknell et al., under review). However, those task demands are quite far from those of everyday language processing where listeners need to attend to many different parts of the signal and topics change rapidly. To the extent that Experiment 2 more closely reflects the task demands of natural language understanding—which strikes us as likely—our results suggest that listeners may not ideally integrate long-distance cues. Future work should continue to investigate the limits of subcategorical maintenance: what do listeners do when confronted with the typical demands of natural language use?

7 Future Work

One question not addressed in the current work is the extent to which different participants engage in different integration strategies or may change strategy over time. Our data are likely a mix of participants who show ideal integrator-like behavior and categorize-discard-&-switch behavior—what drives these differences? One possibility could be differences in working memory and attention. In addition, it is plausible that strategies could change over time as a sort of adaptation to task demands. It is possible that listeners under naturalistic demands tend to take a memory-saving suboptimal strategy for the memory benefits (like in our Experiment 2), but with a more constrained, easier-to-predict task become more inclined to switch to a more optimal strategy. Future work should investigate whether and why these changes may occur.

By making models of cue integration explicit, we inform future theoretical and experimental

work. For example, we can analyze these models to understand how well each model performs word recognition: we can directly quantify how much word identification accuracy is expected to decline for the non-optimal models compared to ideal integration. Paired with experiments that emphasize different task demands of typical language use, we can then begin to investigate (i) under what circumstances listeners are (sub)optimal and (ii) whether listeners maximize accuracy given task demands. It may be the case, for example, that in some contexts non-optimal integration is preferred to ideal integration if the expected gain in accuracy does not justify the expected memory demand of maintaining subcategorical information for ideal integration. Equipped with these formal models, we can begin to address such questions.

References

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23.

Klinton Bicknell, Wednesday Bushong, Michael K Tanenhaus, and T Florian Jaeger. under review. Listeners can maintain and rationally update uncertainty about prior words.

Sarah Brown-Schmidt and Joseph C Toscano. 2017. Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, 32(10):1211–1228.

Wednesday Bushong and T Florian Jaeger. 2017. Maintenance of perceptual information in speech perception. Thirty-Ninth Annual Conference of the Cognitive Science Society.

Wednesday Bushong and T Florian Jaeger. under review. Memory maintenance of gradient speech representations is mediated by their expected utility.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39.

Cynthia M Connine, Dawn G Blasko, and Michael Hall. 1991. Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, 30(1):234.

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

Marc O Ernst and Martin S Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429.

Laura Gwilliams, Tal Linzen, David Poeppel, and Alec Marantz. 2018. In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, 38(35):7585–7599.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.

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

Leigh Lisker and Arthur S Abramson. 1967. Some effects of context on voice onset time in english stops. *Language and speech*, 10(1):1–28.

Bob McMurray, Michael K Tanenhaus, and Richard N Aslin. 2009. Within-category vot affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, 60(1):65–91.

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

Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804.