GLOSSA psycholinguistics

Bushong, W. (2025). Strong evidence for maintenance of gradient representations during language processing. *Glossa Psycholinguistics*, 4(1): 15, pp. 1–29. DOI: https://doi.org/10.5070/G601140229

eScholarship University of California

Strong evidence for maintenance of gradient representations during language processing

Wednesday Bushong, Wellesley College, US, wb104@wellesley.edu

To what degree listeners can maintain gradient subcategorical information about speech input in memory over time has been a matter of considerable debate. The literature has largely lacked formal computational models of potential mechanisms against which to compare human behavior. Here, we formalize several competing cognitive models of this process and quantitatively compare them to data from a series of behavioral experiments. We find consistently strong evidence in favor of models which allow for maintenance of subcategorical information over the course of an utterance. These results suggests that listeners are able to maintain relatively fine-grained details about prior linguistic input over long perceptual timescales. This work also highlights the importance of formalizing cognitive models of behavior to distinguish between competing theoretical mechanisms.

1. Introduction

Spoken language understanding is a complex cognitive activity. Listeners need to decode their interlocutors' intended meaning from the acoustic signal they produce, but the percept of this signal is high-dimensional and corrupted by listener-internal, speaker-internal, and environmental noise. Thus, any given cue in the signal inevitably leaves some degree of uncertainty about the underlying linguistic unit (e.g. phonemes, syllables, words). In speech perception, this is typically referred to as the *lack of invariance* problem: there is no one-to-one mapping between acoustic cues and phonetic units (Lisker & Abramson, 1967). The fundamental problem of real-time spoken language processing, then, is how listeners arrive at decisions about the identity of linguistic categories (like sounds, words, and syntactic structures) from inconclusive evidence.

One way for listeners to mitigate this problem is to make inferences based on multiple cues in the input, reducing (but not fully eliminating) uncertainty. In speech, each segment contains a multitude of cues that are relevant to inferring its category. In American English, for example, syllable-initial voicing (distinguishing, e.g., /t/ vs. /d/) is cued by voice-onset time, fundamental frequency, speech rate, burst duration, and other acoustic properties (Cooper et al., 1952; Kingston & Diehl, 1994; Liberman, 1957; Port, 1979). Listeners are able to combine these phonetic cues to infer an underlying phonemic category (Liberman, 1957; Lisker & Abramson, 1970). However, cues to segment identity do not always appear concurrently; they are also temporally distributed across the signal. For example, an important cue to syllable-final voicing is the duration of the *previous* vowel (Klatt, 1976). This temporal distribution also includes higher-level cues beyond acoustics. Later lexical context, for example, can provide cues to the identity of earlier segments – e.g., *-ask* following a segment acoustically manipulated to range between /t-d/ suggests that the earlier segment was more likely to be /t/ (*task* is a word, while *dask* is not). Indeed, listeners integrate these lexical cues with earlier acoustic information in spoken word recognition experiments (Ganong, 1980).[1] This effect is particularly striking because it implies that listeners can maintain subcategorical information about the initial segment /t-d/ in memory over the course of the word: in order to successfully integrate early acoustic and later lexical cues, listeners must have access to the early cue in memory, so it can be integrated with the later cue. But given human memory limitations, it is impossible for listeners to maintain every bit of the complex acoustic signal in memory indefinitely. Thus, it is critical that we reconcile how listeners are able to integrate long-distance cues with the fundamental limitation of memory. The goal of the present work is to investigate to what degree listeners have access to subcategorical information about prior speech input over time, and what limitations there are (if any) on this process.

[1] We use the terms *speech perception and spoken word recognition* interchangeably, as it is hard to disentangle whether word categorization effects reflect listeners' perception of a particular speech sound or a whole word.

These questions are important to ask, because they address the foundational underpinnings of our theoretical understanding of language processing. In particular, a long-held belief in the psycholinguistic literature, often referred to as the *immediacy assumption*, is that listeners categorize incoming input as fast as possible and immediately discard low-level information in favor of categorical representations (Christiansen & Chater, 2016; Just & Carpenter, 1980). This assumption is driven by the (correct) observation that there are memory limitations on language processing. However, there is a large empirical and theoretical literature potentially suggesting that listeners can maintain significant amounts of subcategorical information for long periods of time. Indeed, most influential formal models of speech perception assume that some degree of low-level information is maintained by listeners over time. In connectionist models like TRACE and its descendants, this manifests as lingering activation for competitors as a result of network dynamics (Magnuson et al., 2020; McClelland & Elman, 1986); in Bayesian models like Shortlist B, all relevant acoustic and sentential information is combined to produce a word categorization, implying that listeners keep track of these cues over time (Norris & McQueen, 2008). However, the predictions of these models have primarily been tested in isolated word recognition (Ganong, 1980; Gwilliams et al., 2018; McMurray et al., 2002, 2009; Toscano et al., 2010), and only sometimes are quantitatively compared to formal models.

An emerging area of empirical research investigates these effects at longer timescales. An influential study by Connine and colleagues presented participants with sentences like the following:

(1) When the **?ent** in the **fender** was noticed, we sold the car.

They varied the acoustic features of the **?** segment between /t-d/, and a later word in the sentence semantically biased toward a *tent* or *dent* interpretation of the target word (*fender*, as in (1) above, vs. *forest;* Connine et al., 1991). They found that participants' categorizations of the target word were influenced both by the acoustics of the initial target and the later semantic context. Other studies using similar stimuli and methods ranging from categorization to visual world eye-tracking have yielded similar results at strikingly long distances (up to 35 syllables away from a target word; Bicknell et al., 2025; Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2019, 2025; Falandays et al., 2020; Szostak & Pitt, 2013; Zellou & Dahan, 2019).

These findings have been interpreted to constitute evidence for subcategorical information maintenance. The key assumption is that if both early and late cues are used in categorization, listeners must have maintained subcategorical information about the early cue.[2] This assumption is based on a (usually implicit) comparison between two basic models of listener strategies,

[2] By *subcategorical information*, we mean that listeners maintain a representation with more detail than a simple categorical decision; this could range from maintenance of fine phonetic detail to a probability distribution over phonemic categories; we discuss this in more detail in 5.1.

which in this article we will call *ideal integration* and *categorize-&-discard*. **Figure 1** shows a basic demonstration of the ideal integration strategy, using example stimuli from the studies presented in this article (inspired by Connine et al., 1991). Here, we use the term *ideal* in the sense of ideal observer models, which estimate the statistically optimal solution to perceptual cue integration problems (see, e.g., Ernst & Banks, 2002).[3] The listener maintains some information about an initial stimulus varying between /t-d/ – at minimum, their degree of uncertainty about whether the sound was /t/ or /d/, but potentially more detailed, such as the value of a relevant acoustic cue, like voice-onset time (VOT). When they encounter the later biasing context *fender*, they are able to integrate information from both sources to come to a categorization decision. Notice that if the listener makes an initial categorical commitment to /t/ or /d/, this could not happen. **Figure 2** demonstrates this categorize-&-discard approach: by the time the listener reaches *fender*, they are already committed to a /t/ response and have no access to a gradient representation with which to integrate the contextual information. This is the basic line of reasoning that has led researchers in this field to argue for subcategorical information maintenance.[4]

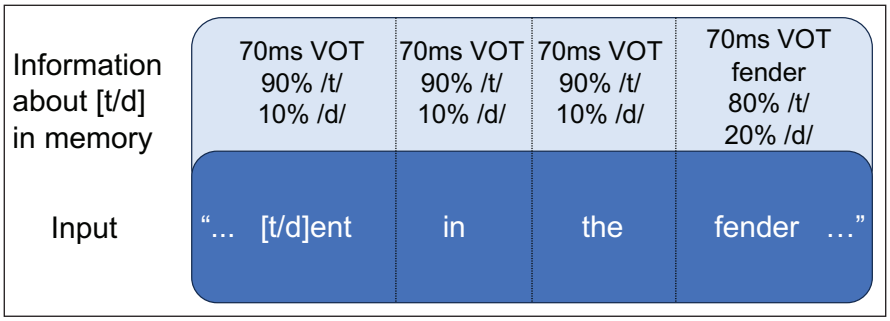

Figure 1: Schematic of the *ideal integration* model, with maintenance of subcategorical information over time.

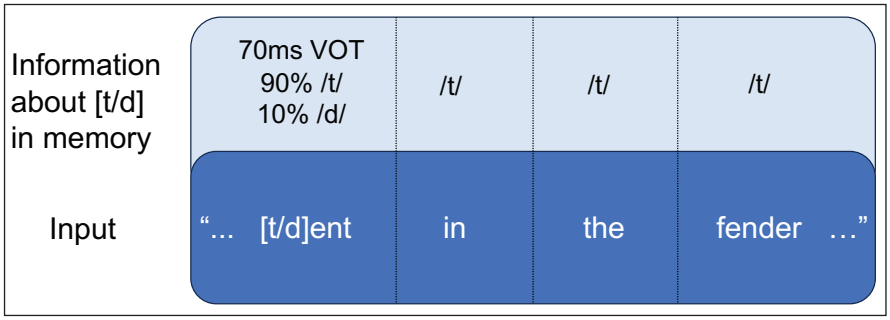

Figure 2: Schematic of the *categorize-&-discard* model, without maintenance of subcategorical information over time.

[3] Notably, these models optimize categorization accuracy; they do not optimize other reasonable goals an organism might have, like categorization speed or memory economy.

[4] This argument was first presented by McClelland and Elman (1986) as a key motivation for TRACE, on the basis of the Ganong effect (Ganong, 1980).

Surprisingly, this interpretation of the literature has gone largely unchallenged, even though there are many other possible word recognition strategies that listeners could engage in beyond the ideal integration and categorize-&-discard options presented above. Indeed, it is quite possible to derive the central qualitative finding of these studies – that listeners' behavioral responses depend on both the acoustic properties of the target sound (e.g., the /t-d/ stimulus) and subsequent sentential context (*fender* vs. *forest*) – without requiring maintenance of subcategorical information. Imagine, for example, that listeners initially categorize the /t-d/ sound as /t/ or /d/, discarding all gradient subcategorical information about it. After they encounter the subsequent context, they can choose to switch their response if the context conflicts with their original categorization. As we describe in more detail below, this strategy would yield categorization responses that exhibit dependence on both the acoustic and subsequent contextual cues: exactly the qualitative pattern observed in previous research. In short, there are plausible scenarios under which the available empirical evidence is *qualitatively* compatible with models of spoken word recognition that do not allow subcategorical information maintenance. Thus, it is clear that relying on qualitative outcomes like the presence or absence of acoustic and contextual effects in experiments is not sufficient for distinguishing between different theories of subcategorical information maintenance. Instead, we need to mathematically formalize these theories and fit them quantitatively to behavioral data.

The goal of the present study is to develop formal models of subcategorical information maintenance and test them in behavioral experiments. We formalize and evaluate a range of plausible listener strategies, from all-or-nothing approaches (like the ideal integration and categorize-&-discard models), to more nuanced strategies. Critically, by mathematically formalizing our models and fitting them quantitatively to behavioral data, we make our theoretical assumptions explicit, in contrast to previous work, which evaluates qualitative data patterns that could (in principle) be compatible with different theories.

First, we describe the formalization of each model, then present four behavioral experiments against which we fit our computational models. **Figure 3** shows the general structure of the modeled task, closely following Connine et al. (1991). Listeners hear sentences that contain a target word whose onset varies acoustically between /t-d/ (by manipulating voice-onset time, VOT); additionally, a subsequent word in the sentence biases toward a particular interpretation of the target word. The listeners' task is to categorize the target word.

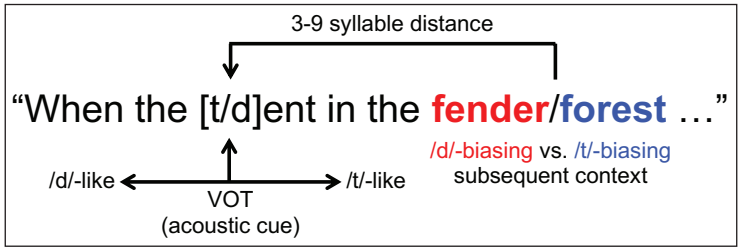

Figure 3: General design of stimuli for all experiments in this article.

2. Models

We formalize five models of how listeners may maintain subcategorical information maintenance (or not): *ideal integration, ambiguity-dependent, categorize-&-discard*, *categorize-discard-&-switch*, and *context-only*. 2.2 describes the models, and **Figures 4** and **5** show the qualitative predictions of each model. For additional details about model fitting procedures, see the Supplementary Information (SI §1 at the GitHub repository for this study).[5]

2.1 Modeling preliminaries

Before we discuss each model in more detail, we first address two basic aspects of working with speech categorization data: first, how acoustic cues alone are expected to influence categorization responses; and second, which space categorization data is most usefully analyzed in.

2.1.1 Predicting categorization responses from acoustic cues

It is important to address the major factor which impacts speech perception and can potentially change the qualitative and quantitative predictions of the cognitive models at hand: how listeners categorize voicing based on the acoustic cue of VOT alone. There are two issues to address here: (i) the link between listeners' underlying representations of acoustic evidence and decisions; and (ii) how exactly VOT affects listeners' perception of voicing.

Following previous work on other questions in speech perception, all the models we present assume that listeners' categorization responses are proportional to listeners' subjective posterior probabilities of categories (Luce's choice rule; R. D. Luce, 1963). The Luce choice rule has been found to provide a good fit to human categorization responses (Clayards et al., 2008; Feldman et al., 2009; Kleinschmidt & Jaeger, 2015; Kronrod et al., 2016; P. A. Luce & Pisoni, 1998). There have been other proposals for linking functions from underlying representations to decisions (see, e.g., Massaro & Friedman, 1990). These choices would likely affect the quantitative predictions of our models presented here, but given that Luce's choice rule is a standard assumption in the literature, we consider it outside the scope of the present work to consider other alternatives.

Although not always described in these terms, most theories of speech perception agree that the predicted slope of the VOT effect on categorization depends on listeners' beliefs about both the means and variances of the /t/ and /d/ categories along the VOT continuum. This follows from the decision rules for categorization in common models of speech perception (e.g., P. A. Luce & Pisoni, 1998; Norris, 1994; Norris & McQueen, 2008; Oden & Massaro, 1978). If two Gaussian categories (/t/ and /d/) have equal variance along VOT, an ideal observer will exhibit linear effects of VOT on the *log-odds* of /t/-responses. However, it is well established that voicing contrasts (including

[5] The GitHub repository containing full data, analyses, and supplementary information for this project can be accessed via this persistent link: https://doi.org/10.5281/zenodo.15237589.

/t/ vs. /d/) exhibit unequal variances along the VOT continuum (Lisker & Abramson, 1967). A standard ideal-observer model, thus, predicts positive quadratic effects of VOT on the log-odds of /t/-responses. Quadratic VOT effects can subtly change the qualitative and quantitative predictions of some of our models. For simplicity of visualization, we show model predictions here and in the main text without quadratic VOT effects. However, all fitted models assume that listeners' subjective $p(t|VOT)$ depends on both a linear and quadratic VOT component.

2.1.2 Log-odds vs. proportion space for visualizing and analyzing categorization data

Throughout this work, we discuss our cognitive models' predictions in log-odds space rather than proportion space. This is because this is the space where the models are clearly distinguishable. Proportions are bounded, squashing model predictions into "S"-shaped curves: a linear effect of VOT in log-odds space surfaces as a non-linear effect in proportion space, and the non-linear effect of squared VOT in log-odds space surfaces as a quite similar non-linear function in proportion space. It is, thus, difficult to distinguish the predictions of our models in proportion space. For that reason, we visualize model predictions below in log-odds space, where the qualitative differences are most obvious.

Notably, comparing model predictions in proportion space also exacerbates the problems involved in making qualitative comparisons between model predictions and experimental data. Connine and colleagues, for example, infer that the smaller effects of context at more extreme VOTs observed in their data in proportion space constitute evidence for the ambiguity model (Connine et al., 1991). However, the ideal integration model also makes the prediction that context effects should be smaller at more extreme VOTs in proportion space. The model predictions only become qualitatively distinct when compared to each other in log-odds space. Furthermore, analyzing binary categorization data using linear methods like linear regression or ANOVA further underestimates effects at proportions close to 0 or 1 (Jaeger, 2008).

2.2 Model descriptions

2.2.1 Ideal integration

The *ideal integration* model holds that listeners maintain subcategorical information about the temporally first cue (here, the acoustic cue VOT) in memory for subsequent integration with a later cue (here, context). We use the term *ideal* in the sense of rational cue integration frameworks (Bicknell et al., 2025; Ernst & Banks, 2002). These normative models provide an ideal baseline against which to compare human behavior. Under the ideal integration model, the listener always maintains subcategorical information about VOT, because ideal categorization requires access to at least $p(category|VOT)$ during integration with context. This model has been conceptually proposed in the past, but only qualitatively tested against behavioral data (Bicknell et al., 2025).

If humans have no memory constraints and ideally integrate all cues available to them, their behavior should resemble the predictions of the ideal integration model. That is, /t/-responses should be conditioned on both VOT and context:

$$p_{ideal}(/t/\text{-response}) = p(/t/|VOT, context) \quad \text{(E1)}$$

After applying Bayes' Theorem, this yields:

$$p(/t/|VOT, context) = \frac{p(VOT|context, /t/)\ p(context, /t/)}{p(VOT, context)} = \frac{p(VOT|context, /t/)\ p(/t/|context)}{p(VOT|context)} \quad \text{(E2)}$$

Under the plausible assumption that VOT and context are conditionally independent (following Bicknell et al., 2025):

$$p_{ideal}(/t/\text{-response}) \propto p(VOT|/t/)\ p(/t/|context) \quad \text{(E3)}$$

Translated to log-odds space, this results in a simple addition of the evidence from both cues (see **Figure 4a**).

2.2.2 Ambiguity-dependent integration

In contrast to the ideal integration model, under the *ambiguity-dependent* model, listeners store information about VOT to the extent to which it is perceptually ambiguous: the more ambiguous the VOT is (i.e., closer to a categorization probability of 50%), the more likely listeners should be to maintain information about VOT for subsequent integration with context. The ambiguity-dependent hypothesis – first conceptually proposed by Connine and colleagues (Connine et al., 1991), and a generally accepted theory (Dahan, 2010; Szostak & Pitt, 2013) – 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 subsequent cues. This can be seen as serving memory economy (for related proposals, see also Dahan, 2010).

Previous tests of this hypothesis have been limited to qualitative comparisons (Bicknell et al., 2025; Connine et al., 1991). Those studies have ruled out a categorical ambiguity-dependent model, in which subcategorical information is maintained only for the absolutely most ambiguous input (for a critical review and qualitative comparison to the ideal integration model, see Bicknell et al., 2025). Here we derive a quantitative model. There are several ways of instantiating the idea that information about VOT is only maintained if it is perceptually ambiguous. Here, we evaluate a gradient version of this hypothesis: with increasingly ambiguous VOT evidence, listeners are assumed to be more likely to maintain gradient representations of

VOT to integrate with later context, instead categorizing on the basis of VOT alone. We quantify the degree of perceptual ambiguity as:

$$\alpha = 2\left|p(/t/|VOT) - 0.5\right| \tag{E4}$$

Here, α is determined[6] by the perceptual ambiguity of the stimulus: α is minimized when $p(/t/|VOT)$ is .5, (the maximally ambiguous stimulus); and α is maximized when $p(/t/|VOT)$ is 0 or 1, (the least ambiguous stimuli). We can then use α as a weight in a mixture model that describes the relative probability of using VOT only or integrating VOT and context:

$$p_{ambiguity}(/t/\text{-response}) \propto \alpha p(/t/|VOT) + (1-\alpha)\ p(/t/|VOT, context) \tag{E5}$$

Intuitively, we can think of this as listeners *not* maintaining gradient representations of VOT over time on α proportion of trials. On the remaining 1–α trials, listeners do maintain a gradient representation – notice that this portion of the equation is identical to the ideal integration model. This model predicts effects of both VOT and context on behavioral categorization responses, with context effects particularly pronounced in the center of the acoustic-perceptual continuum (see **Figure 4b**).

2.2.3 Categorize-&-discard

The next three models we consider do *not* maintain information about VOT in memory over time, but rather immediately categorize, based on the first cue, and then discard all subcategorical information about that cue. These models maximize memory economy at the cost of categorization accuracy. Under the most simple *categorize-&-discard* model, listeners categorize the target word based on VOT, discard all subcategorical information about VOT, and then never revisit the categorization decision. As this model never considers later sources of information, its categorization accuracy will necessarily be suboptimal. We formalize this model as simply making decisions on the basis of VOT alone:

$$p_{cat\_discard}(/t/\text{-response}) = p(/t/|VOT) \tag{E6}$$

The hallmark predictions of this model are an effect of VOT, but a null effect of context, on behavioral responses (see **Figure 4c**).

2.2.4 Categorize-discard-&-switch

The second model of this class we consider also discards all subcategorical information about VOT immediately after having used it to categorize. However, under the *categorize-discard-*

[6] I.e., α is not a free parameter in this model.

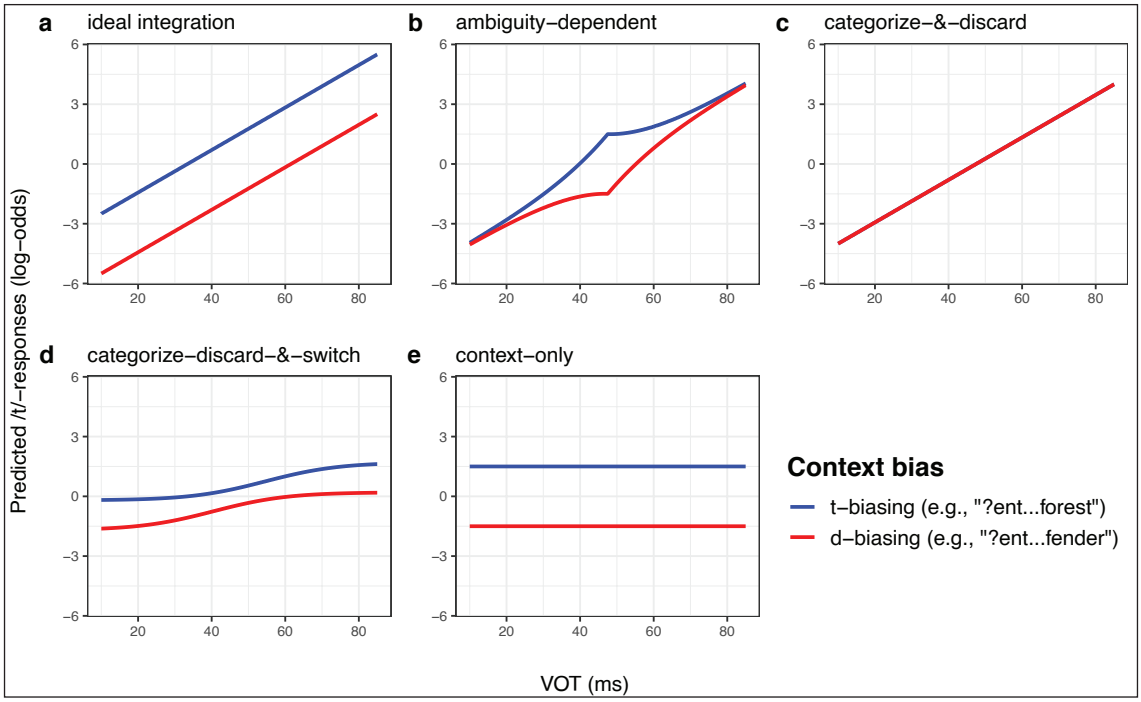

Figure 4: Qualitative predictions of each model by VOT (the acoustic cue distinguishing /t/ and /d/) and context. We set the point of maximal ambiguity to the center of the displayed VOT range, and assume that the contextual evidence for either response (*tent* vs. *dent*) is symmetric around a neutral categorization function that would result in a neutral context (not shown). These choices make it easiest to see the influence of VOT and context on the predictions of the models. For the quantitative evaluation of the models, we do not make these assumptions.

&-switch model, listeners have a mechanism to take context into account: if context conflicts with the initial categorization decision, the listener will change their categorization response in proportion to the strength of the evidence from context. To give a specific example, suppose the listener initially categorizes a segment as /d/, but later evidence from context is more consistent with a /t/ interpretation; the listener will switch their categorization decision to /t/ with probability $p(/t/|context)$.

$$p_{cat\_switch}(/t/\text{-response}) \propto p(/t/|VOT) + (1 - p(/t/|VOT))\, p(/t/|context) \tag{E7}$$

Like the ambiguity-dependent model, we can think of the categorize-discard-&-switch model as describing behavior across trials. Consider trials in the experiment containing /t/-biasing subsequent context. On some proportion of those 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 (i.e., $1 - p(/t/|VOT)$ trials), they switch their response to /t/, proportional to $p(/t/|context)$. The reverse occurs on trials with /d/-biasing subsequent context.

The categorize-discard-&-switch model, like the ideal integration and ambiguity-dependent model, predicts effects of both VOT and context on categorization responses; however, context should affect perception more at perceptual endpoints (i.e., the reverse of the ambiguity-dependent model's predictions; compare the orange and purple lines in **Figure 5b**). This prediction is of particular relevance in light of recent studies that find evidence of numerically larger context effects at acoustic-perceptual endpoints (Bicknell et al., 2025).

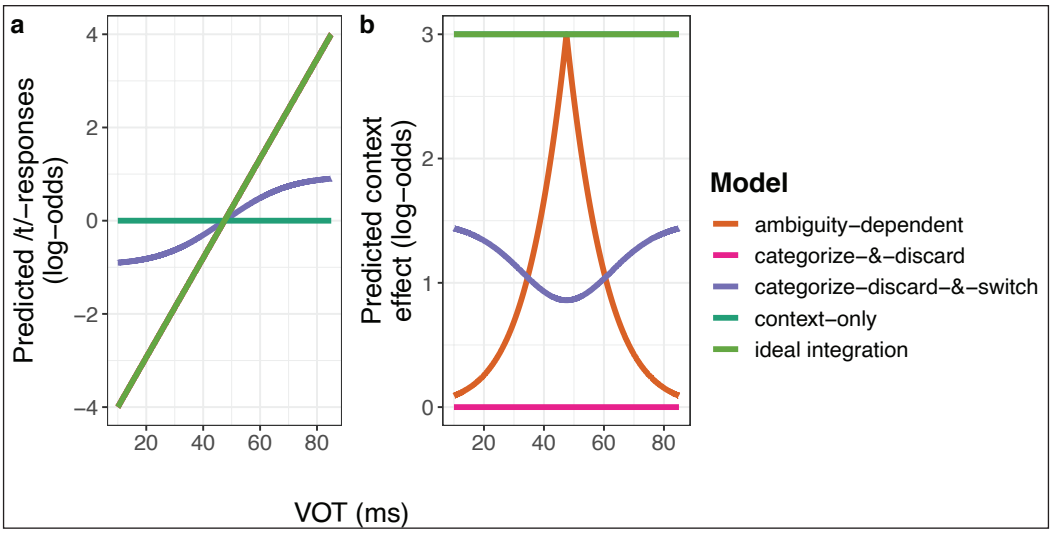

Figure 5: Qualitative predictions of the independent effects of (a) VOT and (b) context for each model. The ideal integration, ambiguity-dependent, and categorize-&-discard models make identical predictions for the effect of VOT on categorizations; similarly, the ideal integration and context-only models make identical predictions for the effect of context on categorizations.

One more point of interest here is the difference in how the ambiguity-dependent and categorize-discard-&-switch models predict differences in the context effect across the VOT continuum. Under the ambiguity-dependent model, whether context enters the listener's categorization process at all is dependent only on perceptual ambiguity; whereas in the categorize-discard-&-switch model, context always affects the listener's categorization process, but whether listeners act on contextual evidence is dependent (indirectly) on the perceptual evidence. In general, the models we present here vary not only in *how*, but also *when* acoustic and contextual information enter listeners' decision-making processes. Teasing apart these distinctions further will likely require paradigms that allow for tracking the timecourse of listener interpretation.

2.2.5 Context-only

Finally, we entertain a model that uses only context in its categorization responses:

$$p_{context\_only}(\text{/t/-response}) = p(/t/|context) \tag{E8}$$

This model captures two potential mechanisms listeners may be engaged in during spoken word recognition. Firstly, this pattern of behavioral responses would be predicted if listeners ignore VOT entirely in the task (an unlikely but possible participant strategy in the present experiments). Secondly, responding based only on context also captures a more extreme version of the categorization-switching model we described above. Our categorize-discard-&-switch model assumes that listeners only make switches when the later context conflicts with their original categorization. One possible alternative switching model is that listeners may switch their categorization choices regardless of acoustic-contextual match, proportional to the evidence from later context, regardless of their earlier categorization; this would predict only an effect of context, with no main effect of VOT. While such a model is highly unlikely to provide an adequate fit to the data, given the strength of VOT effects observed in these kinds of experiments, it serves as an informative baseline against which to compare more complex models.

2.3 Distinguishing the models

Figure 5 shows each model's predictions for the effects of VOT and context on behavior, assuming the same underlying parameters. There is sizeable overlap in the models' qualitative *and* quantitative predictions for each factor. Thus, comparing our models to the empirical data qualitatively is unlikely to be fruitful. However, each model makes unique quantitative predictions about the *joint distribution of VOT and context effects*. We thus evaluate the models quantitatively against behavioral data from four perceptual categorization experiments.

3. Experimental methods

We fit our models to four previously conducted behavioral experiments in our lab that used the same general paradigm (see **Figure 3**). Experiment 1 was previously reported in Bicknell et al. (2025) as Experiment 2; Experiment 3 was reported in Bushong and Jaeger (2019) as the "high-conflict" group. Experiments 2 and 4 have not been previously reported.

Our experimental materials, full datasets, and analysis scripts can be found in our GitHub repository at https://doi.org/10.5281/zenodo.15237589.

3.1 Participants

Participants were recruited from Amazon Mechanical Turk. Each experiment took approximately 30 minutes to complete and subjects were compensated $3.00 for their participation in the experiment.[7] 48 participants were recruited for Experiments 1—2, and 60 were recruited for Experiments 3—4. All experiments were approved by the University of Rochester Research Subjects Review Board (RSRB).

[7] These experiments were conducted between 2016–2018 and were based on a $6/hour compensation rate.

3.2 Materials

Our experiments are inspired by the paradigm first introduced by Connine and colleagues (Connine et al., 1991). **Table 1** shows an example sentence item. Following Connine and colleagues, we manipulated context (tent-biasing vs. dent-biasing), distance (near, 3 syllables vs. far, 6–9 syllables), and voice-onset time (VOT, the acoustic cue distinguishing /t/ from /d/; 6 continuum steps in each experiment). For the purposes of the present work, we do not evaluate any differences between context distance conditions, though this is an important avenue for future research to explore.

Table 1: Example sentence item in each biasing context and distance condition.

Subsequent Context	Distance	Sentence
Tent-biasing	Near (3 syllables)	When the [t/d]ent in the **forest** was well camouflaged, we began our hike.
Dent-biasing	Near (3 syllables)	When the [t/d]ent in the **fender** was well camouflaged, we sold the car.
Tent-biasing	Far (6–9 syllables)	When the [t/d]ent was noticed in the **forest**, we stopped to rest.
Dent-biasing	Far (6–9 syllables)	When the [t/d]ent was noticed in the **fender**, we sold the car.

Each participant heard seven sentence frames in each of the context, distance, and VOT condition combinations, resulting in a total of 168 sentences in each experiment.[8]

3.3 Procedure

Participants were instructed to listen to the sentence and report whether they heard the word *tent* or *dent*. Between experiments, we manipulated whether participants could make a response only after they had heard the entire sentence ("forced-response"), or were permitted to respond anytime during the sentence stimulus ("free-response").

We chose this comparatively simple paradigm because it allows a clear linking function between the input (acoustic cues and subsequent context) and listeners' categorization decisions (for more discussion, see 2.1.1). By contrast, the link between subjective probabilities and more complex measures, such as fixation latency in visual-world eye-tracking experiments (Brown-Schmidt & Toscano, 2017; McMurray et al., 2009), or MEG responses (Gwilliams et al., 2018), is less well understood. This rich temporal information has the potential to give us additional

[8] For full sentence materials and details about which materials were used in which experiments, see the stimulus files in our GitHub repository.

insight into the mechanisms listeners use when integrating incoming information, but it is not necessary to answer the basic question we seek to address in the current work.

3.4 Acoustic manipulation

We created a continuum between /t/-/d/ by following the procedure of previous studies (Bicknell et al., 2025; Connine et al., 1991). From our recordings of the full sentence stimuli, we took one recording of *dent* with a relatively short VOT (10 ms) and one recording of *tent* with a relatively long VOT (85 ms). Then, we replaced the /d/ portion of the *dent* recording by successively replacing more and more portions of the /t/ in *tent* (i.e., 15 ms VOT was created by taking the closure and burst of the /t/ recording plus 15 ms of VOT and pasting this onto the *ent* portion of the *dent* recording). The continuum created by this process then replaced the original target words produced in the full sentence recordings.

Experiments 1–2 use the same stimulus set used in Bicknell et al. (2025), and we used the same VOT steps as reported in that study. For Experiments 3–4, we developed a new stimulus set with an expanded set of sentence frames. We used the same VOT manipulation process on these stimuli; after stimulus creation, we conducted a norming study in order to choose the VOT points we would present to participants. The full details of the norming study are presented in SI §2 at the GitHub repository for this study.

3.5 Data exclusions

Following previous work using this paradigm (Bicknell et al., 2025; Bushong & Jaeger, 2019), participants were excluded from data analysis if they showed no effect of VOT, as defined by significance of a VOT coefficient in a simple logistic regression fitted to each subject. This resulted in the exclusion of 8, 11, 9, and 12 participants, respectively. For the free-response experiments, we removed trials where participants responded before hearing the biasing subsequent context (defined as 200 ms after context word offset, to account for motor planning). See **Table 2** for the number of observations remaining for each experiment after exclusions.

Table 2: Overview of each experiment after data exclusions. For the free-response experiments, we removed trials where participants responded before subsequent context, resulting in many fewer trials than their forced-response counterparts.

Experiment	Response Type	# Participants	# Observations	VOT Steps
Experiment 1	Forced-Response	40	6,720	10, 40, 50, 60, 70, 85
Experiment 2	Free-Response	37	3,470	10, 40, 50, 60, 70, 85
Experiment 3	Forced-Response	51	8,568	10, 30, 35, 40, 50, 85
Experiment 4	Free-Response	48	4,723	10, 30, 35, 40, 50, 85

3.6 Model fitting and comparison

We employed Bayesian non-linear mixed-effects regression to test each of our formal cognitive models. The advantages of this approach are (i) we can directly fit the equations derived above for each of the models, rather than relying on testing qualitative predictions (like patterns of significant results), and (ii) the Bayesian approach allows us to derive measures of evidentiary support based on posterior predictive accuracy. To implement these models, we used the nonlinear formula feature of the brms package in R (Bürkner et al., 2017; R Core Team, 2016).

We follow common practice and use weakly regularizing priors to facilitate model convergence. For fixed effect parameters, we use Student priors centered around zero, with a scale of 2.5 units (following Gelman, 2008) and 3 degrees of freedom. For random effect standard deviations, we use a Cauchy prior, with location 0 and scale 2, and for random effect correlations, we use an uninformative LKJ-Correlation prior, with its only parameter set to 1 (Lewandowski et al., 2009), describing a uniform prior over correlation matrices. Each model was fit using four chains, with 1,000 post-warmup samples per chain (after thinning to every 4th sample to reduce auto-correlations), for a total of 4,000 posterior samples for each analysis. Each chain used 2,000 warmup samples to calibrate Stan's No U-Turn Sampler. All analyses reported here converged (e.g., all $1 \leq \hat{R}s \ll 1.01$).

To compare models against each other, we used the Watanabe-Aikake Information Criterion (WAIC, also known as Widely Applicable Information Criterion; Watanabe & Opper, 2010). The WAIC is a measure for the comparison of non-nested models. It is an approximation of Bayesian leave-one-out (LOO) cross-validation, which provides a measure of a model's predictive accuracy – specifically, its estimated log predictive density (*elpd*; Gelman et al., 2014; Watanabe & Opper, 2010). LOO is very computationally intensive, requiring re-fitting of the same model many times. Considering the complexity of our models, refitting each model to thousands of observations for every experiment is computationally infeasible.

WAIC saves on this expensive computation by starting with a biased estimate of a model's *elpd* (based on its within-sample predictive accuracy), and correcting for its effective number of parameters. This is particularly important in our case, because several of our models have the same number of fitted parameters, but have a higher effective number of parameters (compare the ideal integration and ambiguity-dependent models). We chose the WAIC, as opposed to other information criteria, because it averages over the posterior density of the model, rather than relying on point estimates. This makes the WAIC useful in evaluating mixed-effects models like ours, which contain many parameters that may result in singular estimates (Gelman et al., 2014). Continuing forward, we will refer to the WAIC-estimated *elpd* as $elpd_{waic}$.

There is no general rule of thumb for what differences in $elpd_{waic}$ between models constitute evidence for a difference. One proposal by Vehtari[9] is 5 times the standard error (SE) of the difference – 2.5 SEs, to cover the 95% interval on the difference, and multiplication by 2, since this is the upper limit on the error of the 99% interval estimated by Bengio and Grandvalet (2004). For our purposes here, we will classify 2.5 SE $< elpd_{waic}$ diff $<$ 5 SE as weak evidence, and $elpd_{waic}$diff $>$ 5 SE as strong evidence.

3.7 Assessing individual differences

Recent studies of cue integration in spoken word recognition have increasingly noted that there is sizable individual variability in cue use and weighting (Crinnion et al., 2024), including in some of our recent work using this paradigm (Bushong & Jaeger, 2025). Thus, it is possible that the best-performing models fitted to an entire experiment might not accurately characterize any particular individual subject. The inclusion of random effects over subjects mitigates this issue slightly, but is inadequate for assessing whether different listeners use wholly different strategies.

To characterize possible individual differences in listener strategies, we fit each of our five models to each individual participant across the four experiments. After this process, we excluded from further analysis any subject for whom at least one model did not converge (which we define as at least one $\hat{R} \geq 1.01$ or ≤ 0.99, a slightly looser criterion than our standard for the aggregate models). We then conducted the $elpd_{waic}$ model comparisons within each individual participant. We calculated which model was the best fit for each subject and the degree of evidence for that model over the next-best-fitting model (defined, as above, as a difference in $elpd_{waic}$ of $>$2.5 SE for weak evidence $>$5 SE for strong evidence).

4. Results

At the whole-experiment level, the model comparisons yielded strikingly similar results across all experiments. Models with subcategorical information maintenance always outperformed models without subcategorical information maintenance; in fact, there was strong evidence against the categorize-&-discard, categorize-discard-&-switch, and context-only models, compared to the ideal integration and ambiguity-dependent models, in twenty-three out of twenty-four comparisons across the experiments. The ideal integration model was the best-fitting across the board, strongly outperforming the ambiguity-dependent model in Experiments 1–2, and weakly outperforming it in Experiments 3–4. For full pairwise model comparisons for each experiment, see **Table 3**.

[9] https://discourse.mc-stan.org/t/interpreting-elpd-diff-loo-package/1628/2.

Table 3: Pairwise comparison of model fits ($elpd_{waic}$) for Experiments 1–4. Each cell shows the fit difference and the standard error of the difference in parentheses. Negative values indicate that the model listed in the row is a better fit than the model in the column (i.e., the top left cell shows the ideal integration model is a better fit than the ambiguity-dependent model for Experiment 1). Italicized cells indicate weak evidence for a difference ($elpd_{waic}$ difference $>$ 2.5 SEs), with bolded cells indicating strong evidence (difference $>$ 5 SEs).

Experiment 1	ambiguity	cat.-discard	cat.-discard-switch	context-only
ideal	**–25.6 (4.6)**	**–55.6 (9.9)**	**–1009.7 (42)**	**–2415.4 (52.4)**
ambiguity		*–30 (7.9)*	**–984.2 (43.9)**	**–2389.8 (53)**
cat-discard			**–954.1 (45.5)**	**–2359.8 (53.4)**
cat-discard-switch				**–1405.6 (33)**
Experiment 2	ambiguity	cat.-discard	cat.-discard-switch	context-only
ideal	**–30.7 (5.5)**	**–187.3 (17.8)**	**–294.8 (25.3)**	**–861 (33.9)**
ambiguity		**–156.6 (16.7)**	**–264.1 (27.2)**	**–830.3 (34.3)**
cat-discard			**–107.5 (30.9)**	**–673.7 (38.9)**
cat-discard-switch				**–566.2 (22.9)**
Experiment 3	ambiguity	cat.-discard	cat.-discard-switch	context-only
ideal	*–26.3 (6.6)*	**–180.3 (17.6)**	**–1481.4 (42.7)**	**–2662.2 (57.1)**
ambiguity		**–154 (16.4)**	**–1455.2 (44.9)**	**–2635 (57.6)**
cat-discard			**–1301.1 (48)**	**–2481.9 (59.5)**
cat-discard-switch				**–1180.8 (44.3)**
Experiment 4	ambiguity	cat.-discard	cat.-discard-switch	context-only
ideal	*–17.1 (5.8)*	**–130.9 (15.6)**	**–579.5 (33.2)**	**–1187.5 (42.8)**
ambiguity		**–113.8 (14.9)**	**–562.4 (35.1)**	**–1170.4 (42.7)**
cat-discard			**–448.6 (38.5)**	**–1056.7 (45.4)**
cat-discard-switch				**–608.1 (32.4)**

To illustrate the fit of the different models to listeners' responses, we visualize the predictions of all models for Experiment 2 in **Figure 6**.[10] It is clear from these fits why the *a priori* plausible categorize-discard-&-switch model (and its more extreme counterpart, the context-only model) performed so badly: the model predicts quite a shallow effect of VOT, which does not fit well to the relatively steep average slope we observe in behavior. By contrast, the ideal integration and ambiguity-dependent models fit the VOT effect quite well, while also explaining the presence of the context effect.

[10] We show Experiment 2, because this dataset had the largest overall context effect, which makes the qualitative differences between the model fits more clear. However, the model fits to the three other experiments showed the same quantitative and qualitative patterns (see Figures S3–6 in the SI at the GitHub repository for this study).

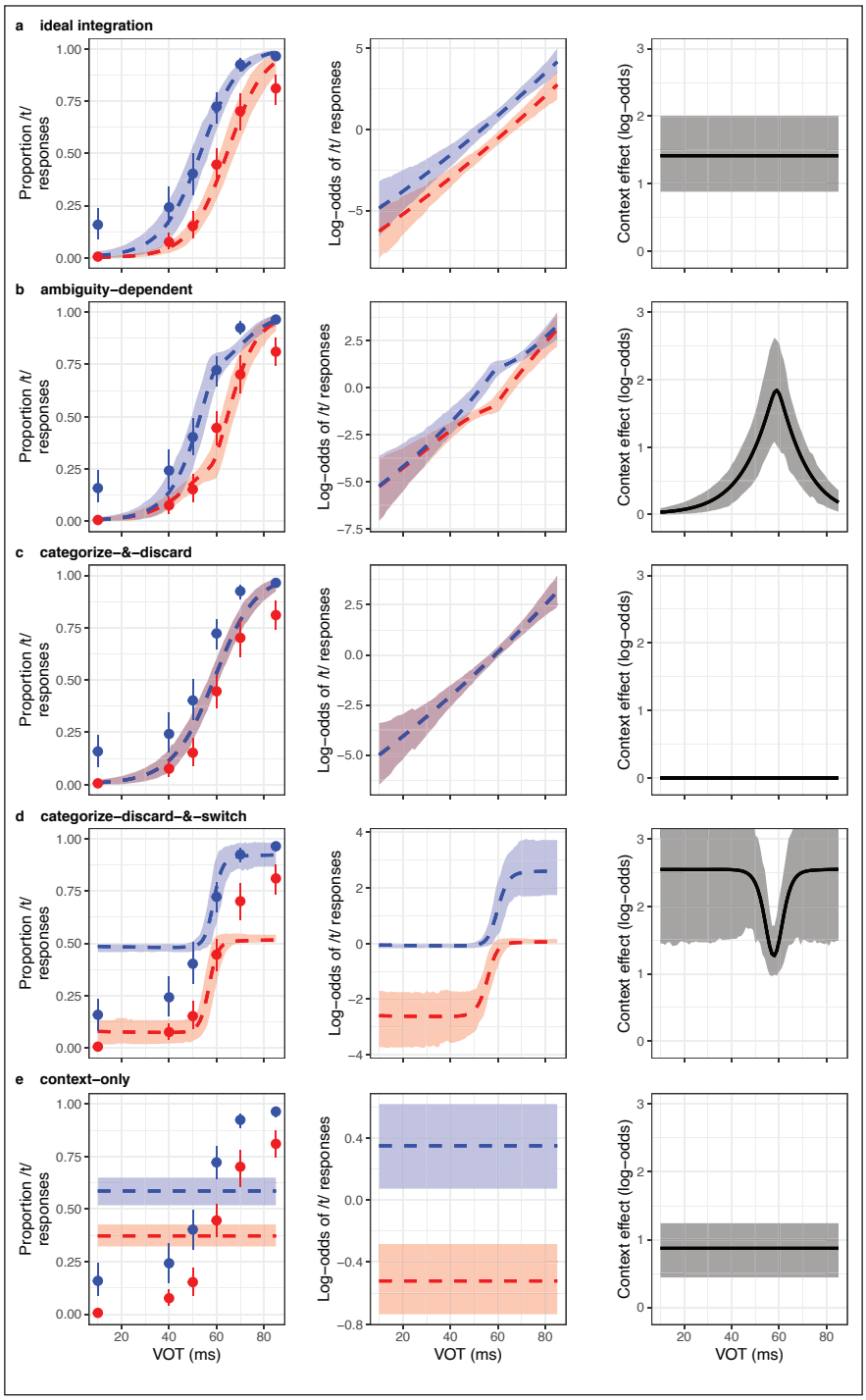

Figure 6: Predictions of the five models fit to Experiment 2 in proportion space (left panel), log-odds space (center panel), and context effect predictions (right panel). Point ranges in the left panel show means and bootstrapped 95% confidence intervals over empirical by-subject means. Dashed lines and shaded regions are mean and 95% highest-density continuous interval (HDCI) of model predictions, drawn from 1,000 random posterior samples.

4.1 Individual results

Of 176 participants, at least one model failed to converge for 25, leaving us with 151 participants who had analyzable results. The results of the model comparisons are summarized in **Figure 7**. Unlike the models fit to whole experiments, the results for individual participants were less clear. For every participant, the best-fitting model was not statistically distinguishable from the next-best-fitting model (i.e., $elpd_{waic}$ difference < 2.5 SE). Numerically, for most participants, the best-fitting model was the ideal integration model (62, 41% of participants), followed by categorize-&-discard (61, 40.4%), ambiguity-dependent (14, 9.3%), categorize-discard-&-switch (13, 8.6%), and context-only (1, .6%).

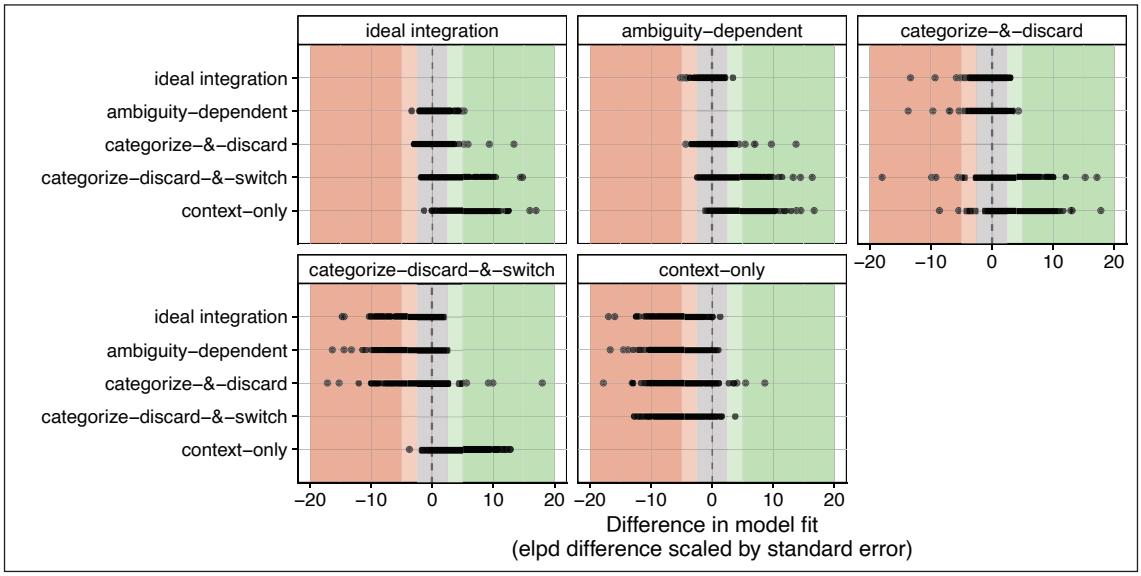

Figure 7: Summary of models fit to individual participants. Each panel represents a model, and each position on the y-axis indicates the model it is compared against. Each point represents an individual subject. The position on the x-axis is the degree of evidence for the model represented by the panel. Shaded regions indicate degree of evidence for or against the model (gray: inconclusive evidence, light green/light red: weak evidence for/against, green/red: strong evidence for/against). Note that there was no subject for whom the best-fitting model performed significantly better than the *next-best-fitting* model. So while, for example, there are many instances of the ideal integration model being a significantly better fit than either the ambiguity-dependent or categorize-&-discard model (see top-left panel), it was never the case that the model was a significantly better fit than *both* of those models within an individual participant.

5. General discussion

There is a substantial body of work that seeks to answer the question of whether listeners are able to maintain subcategorical information about previous input (Bicknell et al., 2025; Brown-Schmidt & Toscano, 2017; Connine et al., 1991; Falandays et al., 2020; Ganong, 1980; McMurray et al., 2009;

Szostak & Pitt, 2013; Zellou & Dahan, 2019, inter alia). The inferences made by these studies have rested on the assumption that observing effects of both initial acoustic input and later contextual information on behavioral responses constitutes evidence that listeners have maintained gradient subcategorical information about prior input. While some studies have proposed conceptual cognitive models that can be compared to behavior (Bicknell et al., 2025; Connine et al., 1991), there has been no concerted effort to formalize and quantitatively test these alternatives.

Here, we formalized five cognitive models that allow us to distinguish different kinds of information maintenance using results from perceptual categorization studies. Two of these models, ideal integration and ambiguity-dependent, were based on prior conceptual proposals in the literature (Bicknell et al., 2025; Connine et al., 1991). We introduced three additional models that assume listeners do not maintain any uncertainty about prior input after initial word recognition: the categorize-&-discard models and context-only model. The categorize-discard-&-switch is a novel contribution to this literature – to our knowledge, such a cognitive process has not been proposed before to explain subsequent context effects. At first blush, this new model seemed to provide an alternative explanation for behavioral patterns that reflect both early and later cues: if listeners simply switch their categorizations when later information conflicts with initial categorizations, one would expect this pattern.

The quantitative comparison of the competing models yielded strikingly consistent results across experiments: the ideal integration models always outperformed the four non-ideal models. The ambiguity-dependent model was also a strong contender, but it always patterned after the ideal integration models, and in two of our four experiments, the evidence against it in favor of the ideal integration model was strong. The three models that assume listeners discard subcategorical information were systematically worse, with our novel proposal, the categorize-discard-&-switch, patterning consistently second-worst. On the whole, these results very strongly suggest that listeners are capable of maintaining subcategorical information about input over long perceptual timescales (3–9 syllables).

Since there may be variability in strategies between participants, we also assessed model fits within individuals. These results were less conclusive, because within each participant, the best-fitting models were statistically indistinguishable from the next-best fit, making it difficult to draw firm conclusions. The qualitative pattern of results, however, showed some divergence from the models fit to whole datasets: while the ideal integration model was the best-fitting model for a plurality of participants, the categorize-discard-&-switch model was a close second, and the ambiguity-dependent model was the best fit for only a small fraction of the total participants (as in the full-experiment results, the categorize-discard-&-switch and context-only models were the worst-performing). We discuss these results further in 5.2.

The failure of the categorize-discard-&-switch model is illustrative of the importance of formalizing and quantitatively testing theories. On its face, it appears to be a plausible competitor

to models that assume maintenance of subcategorical information. It also predicts effects of both acoustic and contextual cues across time, and calls into question the assumption in previous work that finding these effects must imply maintenance of subcategorical information (Bicknell et al., 2025; Brown-Schmidt & Toscano, 2017; Connine et al., 1991). When we quantitatively evaluated this model, however, it provided a very poor fit to the data. This work, thus, highlights the importance of directly fitting quantitative predictions of cognitive models to behavioral data. With an eye to the future, we see two major avenues for advancement in this area.

5.1 What kind of subcategorical information do listeners maintain?

While the present work reveals that listeners can maintain gradient representations of previous input, it is unclear what *kind* of information is contained in these representations. Throughout this paper, we use the general term *subcategorical information* to refer to any kind of representation of past input that is below the level of a categorical decision. But how detailed these representations are has significant implications for the language processing system. For example, listeners could maintain information about specific cue values over time, which would likely be a highly resource-intensive process. By contrast, listeners may maintain something as general as a probability distribution over possible categories, which would be less resource-intensive, but still sufficient to perform ideal cue integration (under some simplifying assumptions). It is also possible that there is some mixture of representations maintained over different timescales; listeners may maintain fine phonetic detail over limited timescales, moving to uncertainty over categories as more time passes. In this work, we aimed to show that maintenance of *some kind* of subcategorical information is possible over long timescales, but our paradigm cannot adjudicate between these different types of representations.

This issue is not trivially solvable. In a series of neuroimaging studies, Gwilliams et al. (2018, 2022) use MEG to reveal across time the neural activity of brain regions known to be associated with phonetic processing. In particular, Gwilliams et al. (2022) are able to decode phonetic features from these regions (as subjects listen to natural speech) for modest perceptual distances ($\sim$ 300 ms). However, these data do not necessarily disambiguate whether listeners have access to more detailed information: indeed, uncertainty about phonetic feature identity may paradoxically lead to *worse* decoding accuracy (particularly in noisy natural speech), precisely because listeners have access to information more detailed than the binary phonetic feature category level, leading to a higher degree of uncertainty at the category level. Furthermore, listeners could, in principle, make perceptual *commitments* while continuing to maintain subcategorical detail over time – what pattern of neural responses this would predict is unclear.

There is a second line of work that tackles the problem of representational detail behaviorally, using the perceptual recalibration paradigm. Caplan et al. (2021) find that lexical labeling following exposure to acoustically manipulated words failed to induce perceptual recalibration effects (in contrast to lexical labeling preceding acoustic information). Perceptual recalibration

requires that listeners be able to track acoustic cues and re-map them to phonemic categories, so the absence of recalibration suggests listeners do not have access to representations as detailed as acoustic cue values at the time of lexical labeling. However, other work using a different accent adaptation paradigm has found effects with delayed lexical labeling (Burchill et al., 2018).

Given the results from the above lines of work, we find it likely that listeners in our studies maintain a more general uncertainty, such as a probability distribution over phonemic categories, rather than a more detailed representation of acoustic feature values. However, we cannot rule it out, and testing these questions is very tricky. Careful model-building and highly controlled experiments are likely to be key to future work in this area.

5.2 How general is subcategorical information maintenance?

How generalizable our results are to naturalistic language comprehension depends on two factors: (i) how well our models, which are fit to entire experiments, capture what any particular *individual* listener does; and (ii) whether our task is reflective of typical language use.

Psycholinguistic experiments generally assume a *modal* language user – that is, we operate under the assumption that most humans share the same fundamental language production and comprehension processes, with some limited exceptions (for recent discussion of this issue, see McMurray et al., 2023). This is in contrast to an approach which views psycholinguistic processes as fundamentally variable and under which each individual's behavioral patterns are considered. Thus, it is important in our work to at least begin to address to what degree our average results (here, the models fit to entire experiments) are sufficient descriptors of individual participants. To tackle this issue, we fit each of our models to individual participants. Unfortunately, given the small amount of data at the individual level, the results were inconclusive. The most notable result to us was that the categorize-&-discard model performed much better on an individual level than at the aggregate level; in particular, it was the best-fitting model for nearly the same number of participants as the ideal integration model. The ambiguity-dependent model, by contrast, performed much worse for individuals than in the aggregate models. To some degree, this is likely driven by the strong VOT effects present within subjects. However, we want to avoid speculating about these results too much – there were no participants who had a statistically clear best-fitting model. Ultimately, to address the question of whether there is individual variation in the mechanisms of subcategorical information maintenance, we need a different approach. Future work should collect significantly more data per participant and fit one model that implements a mixture over the base models – in this way, one could get an estimate of the degree to which each model captures variation in strategies between individuals. Such an approach could also help us understand whether there may even be changes in strategy between trials.

A second concern about the generalizability of the present work concerns task. The experiments presented here use highly predictable, repetitive stimuli: participants are always

asked about the first phoneme of the third word of the sentence, which is predictably followed by additional sentence context 3–9 syllables later. Thus, it is worth considering to what extent our results here reflect real, day-to-day language comprehension, versus a learned task strategy that develops based on exposure to our particular stimuli. To some degree, we can address this question empirically. If participants in our studies show effects of VOT and context from the very beginning of the experiment, this would constitute evidence (albeit limited) that use of these time-disjoint cues is not a task-dependent strategy that requires repeated exposure to our stimuli. To that end, we conducted additional trial analyses (presented in SI §3 at the GitHub repository for this study) on each of our experiments. We find that there is strong evidence for effects of both context and VOT from the very first trial of the experiment. Of course, this does not constitute evidence that there are no task-specific adaptations that may result in behavioral patterns that are not present in natural language comprehension. To mitigate this problem, future work using a paradigm like ours should take steps to draw listeners' attention away from critical manipulations, including introducing filler items, probing alternative words in the sentences, and developing a larger set of sentence items to reduce repetition.

Even if we take as a given that the effects we find here are not task-dependent, it is worth asking whether subcategorical information maintenance is a static, unchanging *mechanism* of language comprehension, or is a *strategy* that is malleable and under listeners' control. To this point, we have used these terms interchangeably, but they imply quite different things about the language processing system. Consider what it would mean for subcategorical information maintenance to be a general mechanism: it would imply that listeners always maintain subcategorical representations about every segment of speech input on an indefinite timescale – the memory demands this process would imply seem immense (and, to some degree, contradictory to the general principle of incrementality in language processing; Christiansen & Chater, 2016). Some work has begun to test whether subcategorical information maintenance can change across time; for example, Bushong and Jaeger (2025) propose that the expected utility of context modulates whether listeners maintain subcategorical information; they find that when sentence context is less informative, listeners subsequently down-weight its use in a spoken word recognition task. Some lexical garden-path studies have also started to investigate whether individual listeners' perceptual abilities modulate acoustic-lexical cue integration (Kapnoula et al., 2021). And, as we mentioned above, extensions to our current work to model mixtures of strategies may also begin to elucidate these processes. As of yet, however, there are no concrete theories of how perceptual, attentional, and memory processes together play a role in maintaining and updating linguistic representations of uncertainty in real time spoken language understanding.[11] We see this as a fruitful area for future work to address.

[11] Notably, some work in sentence processing has begun to address these issues (e.g., recent extensions of noisy-channel surprisal, such as Hahn et al., 2022).

6. Conclusion

The present work suggests that there is strong evidence that listeners can maintain subcategorical representations of previous linguistic input for long perceptual timescales beyond the single word. The present results point to a need for broader theories of speech perception (and language processing generally) to recognize that listeners have access to low-level information even after initial processing. Converging evidence from other domains (e.g., maintenance of uncertainty about syntactic parses over time; Hahn et al., 2022; Levy et al., 2009) suggests that maintaining intermediate representations about linguistic input may be the norm, rather than the exception, of the human language processing system.

Data accessibility statement

Our experimental materials, full datasets, and analysis scripts can be found in our GitHub repository, which can be accessed via this persistent link: https://doi.org/10.5281/zenodo.15237589.

Ethics and consent

All experiments were conducted when W.B. was affiliated with University of Rochester; the experiments were approved by the University of Rochester Research Subjects Review Board Case No. 00045955.

Acknowledgments

This research was funded by NICHD HD075797 to T. Florian Jaeger and NSF NRT 1449828 to Wednesday Bushong. The views expressed here do not necessarily reflect those of the funding agencies. I would like to thank T. Florian Jaeger for many helpful discussions of this project over the years, and for key contributions to the model fitting procedures and data visualization; and for funding support. Evan Hamaguchi and Chelsea March assisted with stimulus sentence creation and recording. Thanks to Mike Tanenhaus and Aaron White for critical discussions about early forms of this work. I am grateful to those who read early drafts of this manuscript, especially Maleka Donaldson and Jennifer McLeer. Thanks to three anonymous peer reviewers for their insightful feedback on this work.

Competing interests

The author has no competing interests to declare.

Authors contributions

W.B. conceptualized the models and experiments, collected and analyzed the data, and wrote the manuscript.

ORCiD IDs

Wednesday Bushong https://orcid.org/0000-0002–1837-0689

References

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, *5*(Sep), 1089–1105.

Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. F. (2025). Maintenance of subcategorical information during speech perception: Revisiting misunderstood limitations. *Journal of Memory and Language*, *140*, 104565. https://doi.org/10.1016/j.jml.2024.104565

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *32*(10), 1211–1228. https://doi.org/10.1080/23273798.2017.1325508

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PloS One*, *13*(8), e0199358. https://doi.org/10.1371/journal.pone.0199358

Bürkner, P.-C., et al. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bushong, W., & Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135–EL140. https://doi.org/10.1121/1.5119271

Bushong, W., & Jaeger, T. F. (2025). Changes in informativity of sentential context affects its integration with subcategorical information about preceding speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001443

Caplan, S., Hafri, A., & Trueswell, J. C. (2021). Now you hear me, later you don't: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, *32*(3), 410–423. https://doi.org/10.1177/0956797620968787

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, E62. https://doi.org/10.1017/S0140525X1500031X

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. https://doi.org/10.1016/j.cognition.2008.04.004

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*(1), 234. https://doi.org/10.1016/0749-596x(91)90005-5

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, *24*(6), 597–606. https://doi.org/10.1121/1.1906940

Crinnion, A. M., Heffner, C. C., & Myers, E. B. (2024). Individual differences in the use of top-down versus bottom-up cues to resolve phonetic ambiguity. *Attention, Perception, & Psychophysics*, 1–11. https://doi.org/10.3758/s13414-024-02889-4

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121–126. https://doi.org/10.1177/0963721410364726

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, 104088. https://doi.org/10.1016/j.jml.2020.104088

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037//0096-1523.6.1.110

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), 6606. https://doi.org/10.1038/s41467-022-34326-1

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119. https://doi.org/10.1073/pnas.2122602119

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. https://doi.org/10.1016/j.jml.2007.11.007

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037//0033-295x.87.4.329

Kapnoula, E. C., Edwards, J., & McMurray, B. (2021). Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(4), 578. https://doi.org/10.1037/xhp0000900

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419–454. https://doi.org/10.2307/416481

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. https://doi.org/10.1121/1.380986

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), 148–203. https://doi.org/10.1037/a0038695

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712. https://doi.org/10.3758/s13423-016-1049-y

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090. https://doi.org/10.1073/pnas.0907664106

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117–123. https://doi.org/10.1121/1.1908635

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28. https://doi.org/10.1177/002383096701000101

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences*, *563*, 563–567.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61. https://doi.org/10.1037/h0039723

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), e12823. https://doi.org/10.1111/cogs.12823

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*(2), 225. https://doi.org/10.1037//0033-295x.97.2.225

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McMurray, B., Baxelbaum, K. S., Colby, S., & Bruce Tomblin, J. (2023). Understanding language processing in variable populations on their own terms: Towards a functionalist psycholinguistics of individual differences, development, and disorders. *Applied Psycholinguistics*, *44*(4), 565–592. https://doi.org/10.1017/s0142716423000255

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191. https://doi.org/10.1037//0033-295x.85.3.172

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, *7*(1), 45–56. https://doi.org/10.1016/s0095-4470(19)31032-0

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533–1546. https://doi.org/10.3758/s13414-013-0492-3

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532–1540. https://doi.org/10.1177/0956797610384142

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594.

Zellou, G., & Dahan, D. (2019). Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in English. *Journal of Phonetics*, *76*, 100910. https://doi.org/10.1016/j.wocn.2019.06.001