Journal of Memory and Language 140 (2025) 104565

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Maintenance of subcategorical information during speech perception: Revisiting misunderstood limitations

Klinton Bicknell [a,b,1], Wednesday Bushong [c,d,e,1], Michael K. Tanenhaus [b,f], T. Florian Jaeger [b,g,*]

[a] Duolingo, Inc, USA
[b] Department of Brain & Cognitive Sciences, University of Rochester, USA
[c] Department of Psychology, University of Hartford, USA
[d] Cognitive & Linguistic Sciences Program, Wellesley College, USA
[e] Department of Psychology, Wellesley College, USA
[f] School of Psychology, Nanjing Normal University, China
[g] Department of Data Science, University of Rochester, USA

ARTICLE INFO

Keywords:
Ideal observer
Right context
Maintaining uncertainty
Subcategorical information
Word recognition

ABSTRACT

Accurate word recognition is facilitated by context. Some relevant context, however, occurs after the word. Rational use of such "right context" would require listeners to have maintained uncertainty or subcategorical information about the word, thus allowing for consideration of possible alternatives when they encounter relevant right context. A classic study continues to be widely cited as evidence that subcategorical information maintenance is limited to highly ambiguous percepts and short time spans (Connine et al., 1991). More recent studies, however, using other phonological contrasts, and sometimes other paradigms, have returned mixed results. We identify procedural and analytical issues that provide an explanation for existing results. We address these issues in two reanalyses of previously published results and two new experiments. In all four cases, we find consistent evidence against both limitations reported in Connine et al.'s seminal work, at least within the classic paradigms. Key to our approach is the introduction of an ideal observer framework to derive normative predictions for human word recognition expected if listeners maintain and integrate subcategorical information about preceding speech input rationally with subsequent context. We test these predictions in Bayesian mixed-effect analyses, including at the level of individual participants. While we find that the ideal observer fits participants' behavior better than models based on previously proposed limitations, we also find one previously unrecognized aspect of listeners' behavior that is unexpected under any existing model, including the ideal observer.

Introduction

Language comprehension requires inferring linguistic structure from perceptual input. Following Marslen-Wilson's seminal work (Marslen-Wilson, 1973, 1975), models of language comprehension have adopted some variant of what Just and Carpenter (1980) labeled the "immediacy assumption": input is fully processed, i.e., integrated into representations at multiple levels, immediately. The immediacy assumption is often accompanied by a second assumption: once input is integrated, uncertainty about how to categorize the input—that is, gradient information about how consistent the input is with multiple possible categories—is rapidly discarded (hereafter we will refer to this as the *categorize-and-discard* assumption).

In speech perception, the categorize-and-discard assumption, though often implicit, has a long and influential history (for review, see Christiansen & Chater, 2016 and replies in the same volume). It is reflected in standard views of categorical perception: whereas listeners are initially sensitive to within-category differences in phonetic cues such as voice onset time (VOT), sensitivity rapidly decays as inputs are parsed into categorical representations (e.g., Goldstone & Hendrickson, 2010). Some well-known models of spoken word recognition also made this assumption. In the original cohort model (Marslen-Wilson & Welsh, 1978), lexical candidates are immediately discarded once they become inconsistent with the input. Later theories of spoken word recognition

* Corresponding author at: Department of Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Rochester, NY 14620, USA.
E-mail address: fjaeger@ur.rochester.edu (T.F. Jaeger).
[1] Joint first authors.

https://doi.org/10.1016/j.jml.2024.104565
Received 4 March 2024; Received in revised form 5 September 2024; Accepted 6 September 2024
Available online 20 September 2024

weakened this assumption (e.g., TRACE, McClelland & Elman, 1986; NAM, Luce & Pisoni, 1998) and eventually abandoned it altogether (e.g., DIANA, ten Bosch, Boves, & Ernestus, 2022; EARSHOT, Magnuson et al., 2020; Shortlist B, Norris & McQueen, 2008). Yet, it is our experience that intuitions rooted in this view continue to persist in the field, and questions remain about the limits of subcategorical information maintenance during speech perception.[2] This motivates the present work.

There is now increasing evidence that listeners can maintain *some* subcategorical information about preceding speech input beyond the moment. To start with, subcategorical details seem to form part of listeners' long-term memory representations of speech (for review, Hay, 2019). Another line of work—the one that we seek to contribute to here—has asked how long subcategorical details about preceding speech input remain available in *short*-term memory. These studies have found that listeners maintain some subcategorical information about preceding speech input for integration with subsequent "right context" (for review, see Dahan, 2010; Falandays, Brown-Schmidt, & Toscano, 2020). Some effects of local right context have been known for decades. For example, the most important cue to the perception of syllable-initial stop voicing in English (e.g., "pa" vs. "ba") is the VOT of the initial sound. But the duration of a vowel immediately following a word-initial stop is also known to affect the perception of voicing: an ambiguous VOT is more likely to be perceived as a voiced consonant when followed with a long vowel and as a voiceless consonant with a following short vowel (Miller & Volaitis, 1989; McMurray et al., 2008; Summerfield, 1981). Later work demonstrated that such right context effects extend several syllables within a word (e.g., Gwilliams, Linzen, Poeppel, et al., 2018; McMurray, Tanenhaus, & Aslin, 2009) or even *beyond* the word (e.g., Brown, Tanenhaus, & Dilley, 2021; Burchill et al., 2018; Connine, Blasko, & Hall, 1991; Szostak & Pitt, 2013), with some subcategorical information persisting for dozens of syllables (Brown-Schmidt & Toscano, 2017; Falandays et al., 2020). Taken together, these works suggest that subcategorical information can sometimes remain available in listeners' short-term memory for longer than previously thought.

At the same time, previous work has identified potential *limits* of subcategorical information maintenance. A classic study by Connine, Blasko and Hall (1991), for example, is widely cited as evidence that maintenance is (1) restricted to highly ambiguous segments close to a category boundary, and (2) only possible for a few syllables beyond the word boundary. Connine et al. (Experiment 1) examined subcategorical information maintenance for VOT continua for stop consonants between 'dent' and 'tent' by manipulating the lag at which disambiguating information was presented in the right-context (3 syllables [near] or 6–8 [far] downstream). An example item is shown in (1). Six steps along a VOT continuum from 4 ms (voiced, clear *dent*) to 56 ms (voiceless, clear *tent*) were inserted into each sentence frame condition.

(1) [*dent*-biasing, near] When the__in the fender was well camouflaged, we sold the car.
[*tent*-biasing, near] When the__in the forest was well camouflaged, we began our hike.
[*dent*-biasing, far] When the__was noticed in the fender, we sold the car.
[*tent*-biasing, far] When the__was noticed in the forest, we stopped to rest.

The results of Connine et al. (1991)—replotted in Fig. 1—seem to suggest two limitations on subcategorical information maintenance: (1) maintenance seems to be restricted to highly ambiguous VOTs closest to the category boundary (middle of left panel); and (2) maintenance seems to be short-lived, so that right context only affects categorizations when it occurs 3 syllables downstream (left panel) but not 6–8 syllables downstream (right panel). As Connine and colleagues point out, their findings are consistent with a weakened version of the categorize-and-discard assumption: whereas left context effects are pervasive and ubiquitous, right context effects are both limited to special cases approaching maximal ambiguity and are very limited in time.

The two potential limitations of subcategorical information maintenance, maintenance-restricted- to-ambiguity (hereafter the ambiguity hypothesis) and short-lived maintenance have remained influential, even though later studies suggest a more nuanced view. Szostak and Pitt (2013, Experiment 2) conducted an experiment similar to Connine et al. (1991) that focused on another phonetic contrast (place of fricatives: /s/ vs. /ʃ/). Like Connine and colleagues, Szostak and Pitt reported that maintenance seems to be restricted to highly ambiguous speech inputs. Unlike Connine et al., Szostak and Pitt found that subcategorical information is *not* short-lived but maintained to 8–9 syllables beyond the word, the longest lag tested. Later studies—also on a place of fricative contrast (/h/ vs. /ʃ/), using a different paradigm, found long-lived subcategorical information maintenance even after 35 syllables, the longest distance tested so far (Brown-Schmidt & Toscano, 2017; Falandays et al., 2020). Unlike Szostak and Pitt, these later studies also found indirect evidence that, even at the longest distances tested, right context effects were *not* limited to the most ambiguous cases.

Previous work thus seems to suggest a heterogenous set of findings, potentially pointing to differences in subcategorical information maintenance that depend on the type of phonetic contrast. Szostak and Pitt (2013), for example, proposed that fricatives might evoke longer-lasting perceptual memory than stops. But do previous studies really require explanations of this type? Or are there simpler explanations to reconcile these findings? And why do some paradigms seem to reliably suggest that subcategorical information is only maintained for highly ambiguous input, whereas other paradigms have come to different conclusions?

Overview of the present study

The present study revisits the seminal work of Connine et al. (1991). Key to our approach is the introduction of an *ideal observer* framework. We derive normative predictions for human word recognition under the assumption that listeners optimally maintain and integrate subcategorical information with subsequent context. By comparing these predictions against listeners' responses, we assess the extent to which listeners maintain and integrate subcategorical information optimally. This leads us to conclude that the original results by Connine et al. (1991)—and to some extent those by Szostak and Pitt (2013)—have been misunderstood: neither of these studies provide evidence for the two limitations proposed by Connine and colleagues. We argue that all studies conducted so far are compatible with the hypothesis that some subcategorical categorical information is maintained by default—even for highly *un*ambiguous speech—and that subcategorical information maintenance can be detected even at the longest distances tested so far (up to 35 syllables in Falandays et al., 2020). While this conclusion might be surprising, given strong intuitions about the limits of short-term sensory memories, we discuss how such limitations can be reconciled with our findings.

We begin by laying out our argument. We first address the seemingly conflicting findings about the longevity of subcategorical information maintenances. We identify a simple procedural difference between Connine et al. (1991) and subsequent studies, whether participants are allowed to respond before hearing the relevant right context, that offers a straightforward explanation as to why Connine et al.'s study found maintenance to be short-lived while later studies did not. Our explanation does *not* entail that subcategorical information maintenance

[2] We use the term *maintenance of subcategorical information* as an umbrella term to refer to the maintenance of any type of information beyond the category itself. At the very least, this includes *uncertainty* about (or relative activation of) phonological categories, but it could theoretically include richer information, closer to phonetic or even perceptual representations (for relevant discussion, see Burchill, Liu, & Jaeger, 2018; Caplan, Hafri, & Trueswell, 2021). We return to this question in the general discussion.

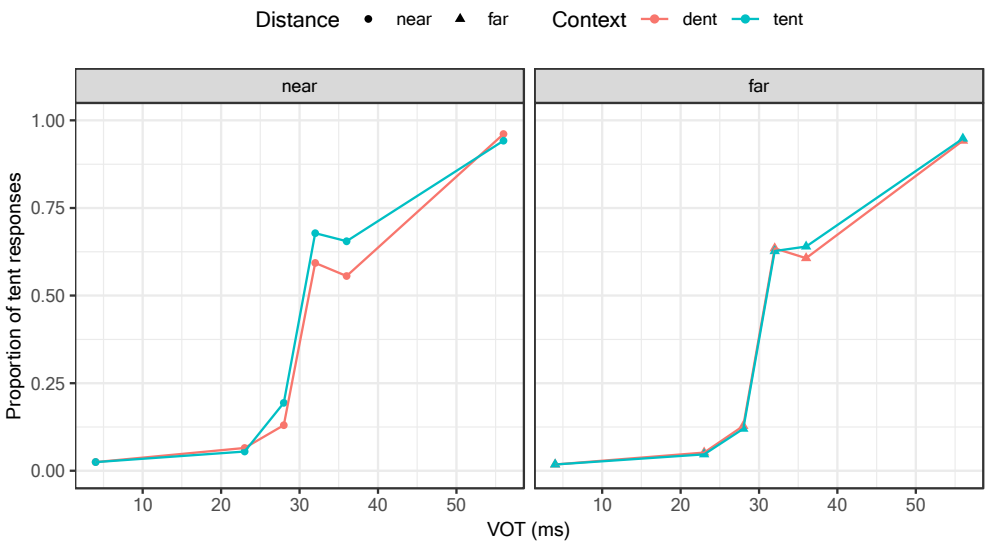

Fig. 1. Experiment 1 from Connine et al. (1991) re-plotted by us. Mean proportions of "tent" responses in each combination of VOT and sentence frame. A significant difference between dent-biasing and tent-biasing contexts was found only for the near condition, and this difference appeared to be driven by VOTs near the category boundary (28, 32, and 36 ms).

depends on the type of phonetic contrast. We then address the ambiguity hypothesis. We introduce the ideal observer framework and derive its predictions for right context effects. We demonstrate that results routinely interpreted as evidence for the ambiguity hypothesis are, in fact, equally compatible with optimal maintenance and integration of subcategorical information.

This leads us to design stronger tests that compare predictions from the ideal observer with stronger and weaker versions of the ambiguity hypothesis. In the strong version, modeled on Connine et al.'s proposal, maintenance is restricted to the most ambiguous stimuli. In the weaker version, maintenance is reduced for less ambiguous stimuli. We use Bayesian mixed-effects logistic regression to compare these hypotheses against both the /s/-ʃ/ data from Szostak and Pitt (2013, Experiment 2) and /d/-/t/ data (as in Connine et al., 1991) from one previously published and two new experiments. We find that the preponderance of the evidence is incompatible with either form of the ambiguity hypothesis, instead favoring the ideal observer hypothesis. However, we also identify a suggestive pattern in all four data sets—increasing effects of right context for perceptually less ambiguous continuum steps—that is inconsistent not only with the ambiguity hypothesis, but also with the predictions of the ideal observer. This leads us to investigate whether this unexpected pattern could be the result of attentional lapses and/or task-related strategies, either of which might cause participants to sometimes respond based on only *parts* of the speech stimulus (e.g., based on only the acoustics of the target word or only the right context). As a preliminary step in this direction, we outline modified versions of both the ambiguity hypothesis and the ideal observer model that account for (intentional or unintentional) attentional lapses and compare them against our results.

Data avail ability

All stimuli, sound recordings, raw result files, and analysis code are available as part of the OSF repository for this project (https://osf.io/6fng2/).

Reconciling seemingly conflicting results

Is subcategorical information maintenance sometimes short-lived?

We submit that there is a simple procedural explanation as to why some previous studies found subcategorical information maintenance to be limited to at most 3 syllables (Connine et al., 1991), and others did not (Brown-Schmidt & Toscano, 2017; Falandays et al., 2020; Szostak & Pitt, 2013). In Connine et al.'s study, participants were allowed to respond at any point during the sentence, whereas all subsequent studies cited above forced participants to respond after the sentence recording finished playing. Indeed, 84 % of responses in the 6–8 syllable condition in Connine et al.'s study (versus 15 % in the 3-syllable condition) occurred *before* the biasing right context was heard. Even though these responses cannot *possibly* be affected by right context, they were still included in analyses. The analysis presented in Connine et al. (1991) thus primarily assessed for how long listeners decide to delay their response given that subsequent context might contain additional information. It does not, however, provide a strong test of how long listeners *can* maintain subcategorical information.

One question we seek to address in the present study is whether this simple difference in procedure might explain the differences in the observed longevity of subcategorical information maintenance across previous studies. We do so by analyzing three experiments—two new experiments and one re-analysis—that employ the /d/-/t/ contrast (as in Connine et al., 1991) while not allowing participants to respond until the end of the sentence (following Szostak and Pitt, 2013 but deviating from Connine et al., 1991). This allows us to ask whether subcategorical information maintenance is observed at longer distances even for /d/-/t/, and thus for all phonetic contrasts tested so far.

Derivation of quantitative predictions from an ideal observer

In ideal observer analysis, one develops an explicit mathematical model of how a system that uses available information optimally would be expected to perform given a set of specified constraints (Geisler, 2003). When human behavior is consistent with an ideal observer, we need not invoke specialized mechanisms: any system making rational use of the evidence would behave as humans do. When, however, the behavior differs, there is clear evidence that there is something more to

be explained (e.g., for speech perception, see Massaro, 1989; Clayards et al., 2008; Norris & McQueen, 2008; Feldman et al., 2009; Kleinschmidt & Jaeger, 2015). For the present case, we start by developing a basic ideal observer that has unlimited perceptual memory, always pays attention, and always uses all information available in the acoustic input and context.[3] As we show, even this basic model suffices to call into question whether findings like those in Fig. 1 really imply that listeners maintain subcategorical information primarily for ambiguous speech input. Beyond this specific question, the fact that the ideal observer approach yields a far more specific linking hypothesis than previous work—a quantitative normative expectation—lays the foundations for future work to develop and test stronger predictive theories of subcategorical information maintenance (in the sense of Yarkoni & Westfall, 2017), a promise we return to in the discussion.

In Connine et al. (1991) and Szostak and Pitt (2013), the participants' task was to answer whether they heard one word (e.g., *tent*) or another (e.g., *dent*). The probability of a word w being *tent* given just a particular subsequent linguistic context c is given by Bayes' rule:

$$p(w = \text{tent}|c) = \frac{p(c|w = \text{tent})p(w = \text{tent})}{p(c)} \tag{E1}$$

Conditioning all terms on the perceptual evidence e for the word w yields the probability that an ideal observer should believe w is *tent*, given both acoustics and right context:

$$p(w = \text{tent}|c, e) = \frac{p(c|w = \text{tent}, e)p(w = \text{tent}|e)}{p(c|e)} \tag{E2}$$

Assuming the context c is conditionally independent from the evidence e given that we know word w (i.e., speakers do not change what subsequent context they produce based on the perceptual realization e of w), we can simplify:

$$p(w = \text{tent}|c, e) = \frac{p(c|w = \text{tent})p(w = \text{tent}|e)}{p(c|e)} \tag{E3}$$

As we show next, this relationship becomes especially clear and simple (as well as convenient to test) in log-odds space. To obtain the log-odds that w is *tent* (as opposed to *dent*), we first take the ratio of the two posterior probabilities to yield odds:

$$\frac{p(w = \text{tent}|c, e)}{p(w = \text{dent}|c, e)} = \frac{\frac{p(c|w = \text{tent})p(w = \text{tent}|e)}{p(c|e)}}{\frac{p(c|w = \text{dent})p(w = \text{dent}|e)}{p(c|e)}} \tag{E4}$$

$$= \frac{p(c|w = \text{tent})}{p(c|w = \text{dent})} \frac{p(w = \text{tent}|e)}{p(w = \text{dent}|e)} \tag{E5}$$

and converting odds to log-odds:

$$\log \frac{p(w = \text{tent}|c, e)}{p(w = \text{dent}|c, e)} = \log \frac{p(c|w = \text{tent})}{p(c|w = \text{dent})} + \log \frac{p(w = \text{tent}|e)}{p(w = \text{dent}|e)} \tag{E6}$$

Thus, the log-odds of *tent* given both perceptual evidence and subsequent context equals the sum of one term depending on subsequent context and one depending on perceptual evidence (and mutatis mutandis for *dent*). For an ideal observer distinguishing between two words, context and acoustics have additive effects on log-odds.[4]

This derivation shows that context and acoustics are predicted to have purely additive effects on log-odds for an ideal observer—at least, for the basic ideal observer presented here. As this prediction of additivity arises from the relationship between the log-odds scale and the ideal observer, it is specific to the log-odds scale: an additive effect in log-odds would appear as a *(non-additive) interaction* in proportion space—the space that both Connine and colleagues as well as Szostak and Pitt used to plot and interpret effects. The effects of right context *on the* log-*odds* of categorization should thus be constant across the acoustic continuum (e.g., the specific VOT value), whereas the same context effects should appear to vary depending on the acoustic input when expressed in terms of *proportions*. Indeed, as we show next, the type and direction of the interaction between context and VOT that has been observed in previous work (cf. Fig. 1) follows the predictions of the ideal observer.

What does this mean for the interpretation of previous work?

The core finding taken in previous work to support the hypothesis that subcategorical information maintenance is limited to highly ambiguous speech inputs is that the effect of right context on the *proportions* of t/d and s/ʃ responses were largest for tokens near the category boundary. However, as we show next, this finding is also consistent with the predictions of the ideal observer. Specifically, an additive, or constant-sized, effect in log-odds space is largest in proportions around 0.5, i.e., exactly for the most ambiguous cases at the category boundary and decreases as proportions go towards 0 or 1. As a concrete example, consider Fig. 2. The left panel shows a hypothetical right context effect that is additive in the log-odds: the blue vs. orange lines are the same distance from each other across the entire phonetic continuum. The right panel shows the same effect (still constant in log-odds) expressed in proportion space. In this latter space, the effect—the vertical distance between the two lines—appears largest at the point of maximal ambiguity along the phonetic continuum, and appears to continuously decrease with increasing distance from that point.

In summary, the findings of both Connine et al. (1991) and Szostak and Pitt (2013) are qualitatively consistent with the predictions of an ideal observer: both report an interaction of right context effects with VOT in proportion space, such that right context effects expressed in proportions are largest for the highly ambiguous cases near the category boundary. That is, the same data that were previously interpreted as evidence that subcategorical information is only maintained for highly ambiguous tokens, is in fact, qualitatively consistent with the predictions of an ideal observer, in which subcategorical information is maintained equally for *all* tokens.[5]

This calls for a stronger test of the ideal observer: if right context effects were estimated in log-odds space, they should be constant across the phonetic continuum. If so, this would argue that perceptual evidence and subsequent context are integrated in an optimal fashion (at least in experiments of this type). This prediction contrasts with the ambiguity hypothesis, which predicts that context effects should be larger for more ambiguous tokens close to the category boundary even when responses

[3] We revisit some of these assumptions in the general discussion. Ideal observers are often used to derive predictions that are unconstrained by mechanistic limitations, such as perceptual noise, attentional lapses, or memory limitations. This is also our starting point here. It, however, also common to extend ideal observers to integrate some mechanistic limitations—e.g., to describe optimal behavior *under those constraints* (e.g., perceptual noise, Jacobs, 2002; Feldman et al., 2009). We return to this issue in the general discussion.

[4] In the general case, for an ideal observer distinguishing n words, the speech input and context have additive effects on multivariate log-odds, expressed as an n-1 dimensional vector of log-odds ratios, where element i gives the log-odds of word i compared to word n.

[5] This would also explain why some recent findings seem to suggest that listeners can maintain subcategorical information even for highly unambiguous inputs (Brown-Schmidt & Toscano, 2017; Falandays et al., 2020): these studies employed different response measure that do not suffer from the issue we identified here for proportions. We return to this point in the discussion.

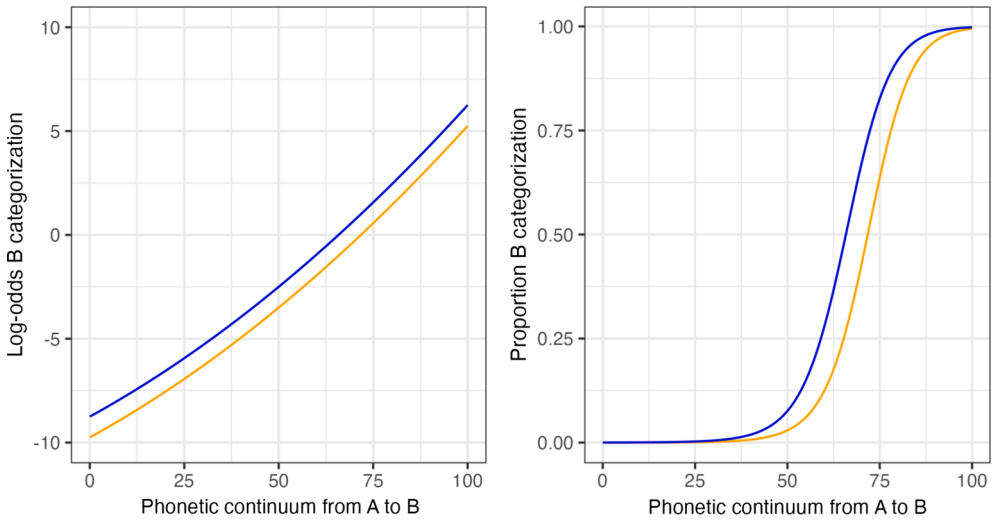

Fig. 2. An effect of right context that is additive in the log-odds of 2AFC response (here: categorizing tokens along a hypothetical phonetic continuum into one of two phonological categories A and B) will appear non-additive if expressed in proportions. **Left:** hypothetical effect of phonetic continuum on log-odds of categorization (mostly linear with some quadratic effect, as in many of the data sets analyzed below). The blue and orange lines represent the two hypothetical contexts, and have constant distance of 1 logit throughout the entire continuum—an effect comparable to what we observe in the experiments presented below. **Right:** The same two conditions but plotted in proportion space. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are interpreted in log-odd space. Next, we compare these two hypotheses against data from both /s/-/ʃ/ as well as /d/-/t/ contrasts.

Reanalysis of Szostak and Pitt (2013, expt. 2)

We begin with a re-analysis of the /s/-/ʃ/ data from Szostak and Pitt (2013, Expt. 2). Since Szostak and Pitt found subcategorical information maintenance even at the longest distance tested, the primary purpose of the re-analysis is to contrast the ideal observer against the ambiguity hypothesis. We do, however, also test whether we can replicate the evidence for longer-lived subcategorical information maintenance that Szostak and Pitt reported.

Approach

Experiment 2 in Szostak and Pitt (2013) crossed acoustic continuum step (5 steps from /s/ to /ʃ/, including two clear endpoints [steps 1 and 51] and three intermediate values [steps 14, 18, and 22]), right context bias (/s/- or /ʃ/-biased), and right context distance (near [2 syllables downstream] or far [8–9 syllables downstream]).

We compare the log-odds additivity prediction from the ideal observer model—i.e., that the effect of right context will be the same size in log-odds space for each acoustic continuum step—with the predictions of the ambiguity hypothesis, in which context has monotonically increasing effects for more ambiguous stimuli. Specifically, we consider two versions of the ambiguity hypothesis, illustrated in Fig. 3: the strong version of the ambiguity hypothesis (left panel) predicts context effects for only the most ambiguous tokens; the weaker version (right panel) merely predicts that context effects are larger for the most ambiguous tokens and decrease for less ambiguous tokens (though not necessarily to zero). In contrast to both versions of the ambiguity hypothesis, the ideal observer (purple) predicts that the effect of right context is constant for all continuum steps.

We assess the predictions in Fig. 3 through two types of analyses. These analyses complement each other, in that they address different aspects of our predictions. The *independent analysis* estimates the effect of right context in log-odds independently for each continuum step. This analysis enables visual inspection in log-odds of the extent to which the data are consistent with the predictions of each of the two hypotheses visualized in Fig. 3. Additionally, this analysis provides the statistical test of whether there are significant effects of right context on the continuum endpoints. To the extent that there are, this provides strong evidence against a strong ambiguity hypothesis. There are, however, two reasons why this analysis by itself cannot address all our questions. First, effects of right context at one or both of the endpoints are compatible with the weak ambiguity hypothesis: that right context effects are *smaller*, but not zero, at the continuum endpoints. Second, the absence of statistically reliable effects at one or both endpoints does *not* provide strong evidence against the ideal observer model: as shown in Fig. 4, the statistical power to detect an effect of context *inevitably* decreases with increasing distance from the most ambiguous point on the acoustic continuum (i.e., the point for which mean proportions of both answers are.5). In Fig. 4, even a 100-fold increase in the amount of data, which would increase power close to the category boundary from less than 25 % to close to 100 %, would barely change the statistical power to detect an effect towards the continuum endpoints (separate power simulations for each experiment are presented in the SI and confirm this point). It would thus not be particularly informative if the independent analyses found no evidence for an effect of right context at the continuum endpoints.

We thus present a second type of analysis, the *combined analysis*, which analyzes all continuum steps together to test whether there is statistical evidence that the effect of right context becomes smaller for the continuum endpoints. If this analysis reveals evidence for attenuation at the endpoints of the continuum, it would provide strong evidence for the ambiguity hypothesis and evidence against the ideal observer model. Together, these two types of analyses provide complementary evidence distinguishing the additivity prediction of the ideal observer from the strong and weak ambiguity hypothesis.

We excluded from all analyses participants who did not exhibit significant effects of the phonetic continuum in the expected direction (as assessed by participant-wise logistic regressions). This ensures that data from participants who respond randomly or who respond solely based on the right context do not confound our analyses. For the Szostak and Pitt reanalysis, this excluded 0 participants. All analyses employ Bayesian mixed-effects logistic regression. Bayesian hypothesis testing has the advantage that it provides us with a coherent measure of the evidentiary support for/against each of the two hypotheses (Bayes

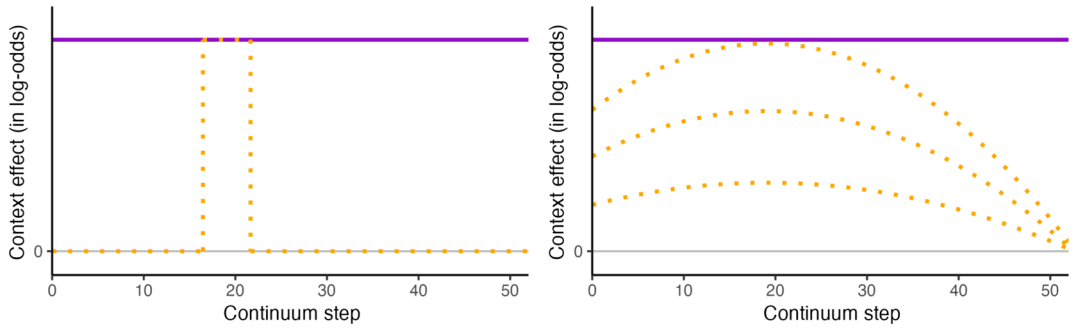

Fig. 3. Schematic predictions for right context effects (in log-odds) at each continuum step under different hypotheses. The ideal observer (purple) predicts that the effect of right context is constant for all continuum steps. The ambiguity hypothesis (orange) predicts that right context has a larger effect for highly ambiguous continuum steps (14, 18, 22) than for continuum endpoints (1, 51). **Left:** a strong version of the ambiguity hypothesis, predicting zero effects of right context at the continuum endpoints. **Right:** a weak version of the ambiguity hypothesis predicting reduced, but not necessarily zero, effects of right context at the continuum endpoints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

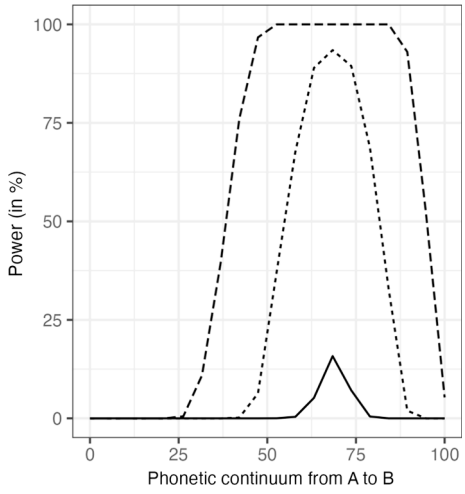

Fig. 4. Statistical power to detect the hypothetical context effect shown in Fig. 2 (constant in log-odds). Power was calculated at 20 equi-spaced points along the phonetic continuum on the x-axis. The different line types represent different amounts of data from 10 data points per continuum step and context condition (solid line), to 100 data points (short dashed) to 1000 data points per continuum step and context condition (long dashes; for details and additional power simulations, see SI). Note also the asymmetry in power at the two 'endpoints' due to the asymmetry in how categorical responses are expected to be at the left and right 'endpoint' (due to the quadratic effect of the phonetic continuum shown in the left panel of Fig. 2). All datasets in this article exhibit this type of asymmetry.

Factors; for discussion, see Jeffreys, 1961; Rafferty, 1995; Wagenmakers, 2007).

All analyses were conducted in the *brms* package (Bürkner, 2017) in *R* (R core team, 2019), using *Stan*'s No-U-Turn Hamiltonian Monte Carlo sampler (Carpenter, Gelman, Hoffman et al., 2016). We follow common practice and use weakly regularizing priors to facilitate model convergence. Specifically, both the coding of predictors and the choice of priors followed our previous work (Xie, Liu, & Jaeger, 2021). For fixed effect parameters, we use Student priors centered around zero with a scale of 2.5 units (following Gelman et al., 2008) and 3 degrees of freedom. We use a Cauchy prior with location 0 and scale 2 for the standard deviations of the random effects (in order to aid convergence), and an uninformative LKJ-prior for the correlations of the random effects. The only parameter of the LKJ was set to 1 (Lewandowski, Kurowicka, and Joe, 2009), describing a weak uniform prior over correlation matrices. As an additional benefit, these priors facilitate model convergence without introducing bias, allowing maximal random effect structures that would otherwise not converge.

All analyses were fit using 4 chains with 2500 warmup and 5000 posterior samples each. Inferences are based on the 20,000 posterior samples.[6] All analyses converged and met common diagnostics (e.g., all $1 < \widehat{R}s < 1.001$; no divergent transitions; chains mixed).

Results and discussion

Independent analysis

The response data are visualized in proportion space in Fig. 5A.

The first analysis we performed estimated the size of the effect of right context in log-odds space for each step of the acoustic continuum separately. We fit separate logistic mixed-effects regressions (for an introduction, see Jaeger, 2008) to the data for each continuum step, each time predicting the proportion of "ship" responses. Each analysis included fixed effects for right context (deviation-coded: +.5 = ship-biasing vs. −.5 sip-biasing) and distance (deviation-coded: +.5 = far vs. −.5 = near), and their interaction. The random effect structure was maximal, including the full by-participant variance–covariance matrix for the factorial design (10 DFs) and the full by-item variance–covariance matrix for random intercepts and slopes for distance (since in Szostak & Pitt's materials, right context was manipulated between items, 3 DFs).

For each independent logistic mixed-effects regression, we assessed the support for the hypothesis of a positive context effect ($H_{\beta_{context}>0}$) against the alternative ($H_{\beta_{context}<0}$). Bayesian hypothesis testing—here implemented through *brms*'s *hypothesis* function—provides a coherent measure of this support. This support is the Bayes factor, $BF_{H_{\beta_{context}>0}, H_{\beta_{context}<0}}$, the ratio between the likelihood of the data under $H_{\beta_{context}>0}$ and the likelihood of the data under $H_{\beta_{context}<0}$. Bayes factors are thus likelihood ratios, ranging from 0 to positive infinity. A Bayes factor of 1 indicates that the support is equally strong for the hypothesis of a positive context effect ($H_{\beta_{context}>0}$) and against it. Values above 1 indicate support for $H_{\beta_{context}>0}$, values below 1 indicate support against $H_{\beta_{context}>0}$. Assuming that the competing hypotheses are *a priori* considered equally probable, a BF > 20 means that the posterior probability of the hypothesis, $p_{posterior}$, is > .95 (since $p_{posterior}$ = BF / (1 + BF)). A BF > 150 means that the posterior probability of the hypothesis is > .99. Following convention, we use verbal labels to describe BFs of 1–3 as

[6] The large number of posterior samples was motivated by additional tests that we had originally planned: tests of the null predicted by the strong ambiguity hypothesis at the continuum endpoints, and tests of the null interactions predicted by the ideal observer. These tests turned out to be both unnecessarily complicated and, for reasons we lay out in the general discussion, uninformative.

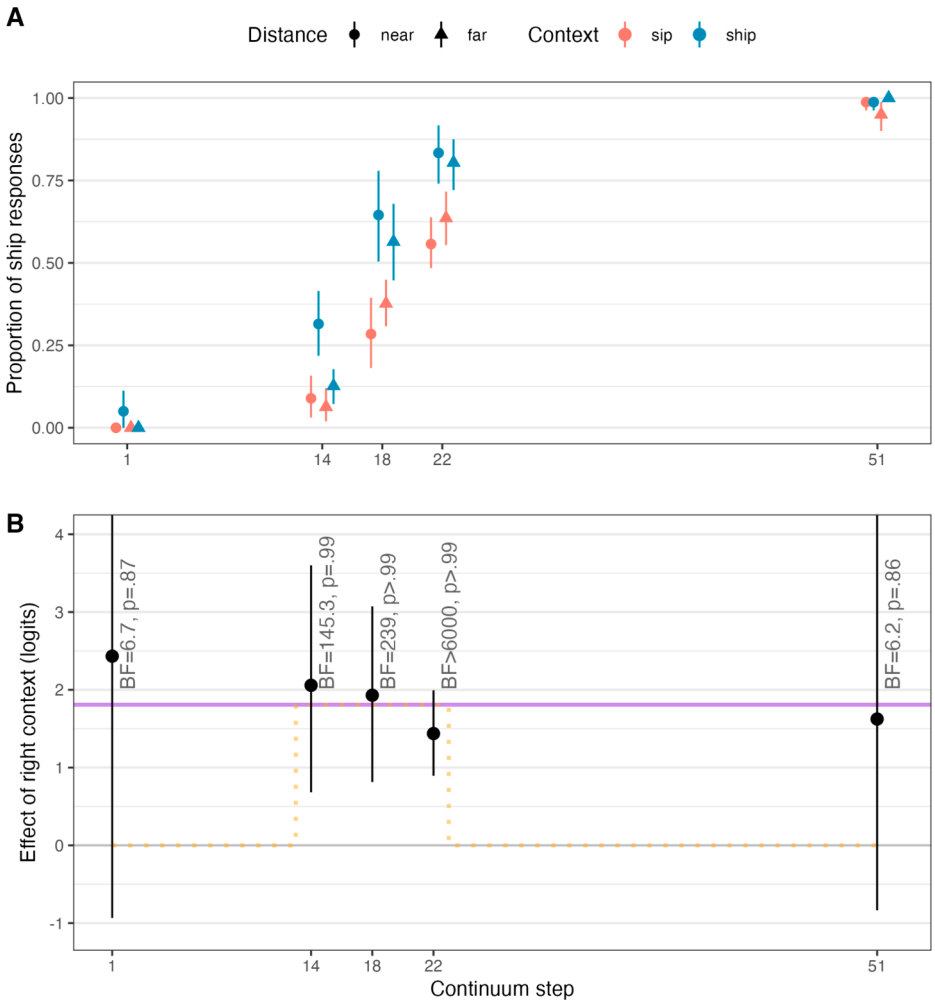

Fig. 5. Panel A: Proportion of 'ship' responses in each condition of the re-analysis of Szostak and Pitt (2013, Experiment 2). Error bars show 95% confidence intervals, bootstrapped over participant means. Panel B: Bayes factor and posterior probability of a positive context effect (in log-odds) obtained from the independent Bayesian logistic mixed-effects regressions. Point ranges show 95% credible intervals (CIs). Also plotted are predictions from the ideal observer (solid purple) and the strong ambiguity hypothesis (dotted orange). For the predictions of the ideal observer, we set the purple line to the average context effect of all continuum steps except for the endpoints.

"weak" or "anecdotal" support, BFs > 3 as "positive" or "moderate" support, BFs > 20 as "strong", and BFs > 150 as "very strong" support (Raftery, 1995; Wagenmakers, 2007). We note though that some have argued for even lower threshold for "strong" (BFs > 10) and "very strong" support (BFs > 30) on the basis of large-scale simulation studies, suggests that even "moderate" evidence can be more reliable than statistical significance under traditional null-hypothesis significance testing (Lee & Wagenmakers, 2013; Schönbrodt et al., 2017). When reporting the degree of support for Bayesian analyses we will use quotation marks to avoid confusion.

Fig. 5B shows 95 % posterior credible intervals for the effect of right context at each continuum step, along with the Bayes factor and the posterior probability of $H_{\beta_{context}>0}$. The right context effect is estimated to be a similar size at all six continuum steps (the points in Fig. 5B), consistent with the predictions of the ideal observer. For the three steps near the category boundary there is "strong" to "very strong" evidence in support of a positive context effect ($p_{posterior}$s $\geq$.99). However, for the continuum endpoint steps 1 and 51 the support is only "moderate" ($p_{posterior}$s $\geq$ .86). This lack of more decisive evidence at the endpoints of the continuum is not necessarily surprising, given the drastically reduced power at the continuum endpoints (recall Fig. 4; confirmed also specifically for the Szostak and Pitt data in Figure S1 in the SI).

While the results of the independent analysis *favor* the ideal observer hypothesis over the strong ambiguity hypothesis, they do not decisively reject the strong ambiguity hypothesis. This is also evident in Fig. 5B: the credible intervals of the context effect at the continuum endpoints include both the purple line (predictions of the ideal observer) and the zero line. The results of the independent analysis also do not speak to the credibility of the weaker version of the ambiguity hypothesis. The analysis we present next provides a way to address these points.

Combined analysis

The second analysis we performed pooled the data across all continuum steps to test statistically for possible interactions between acoustic continuum steps and the effect of right context by fitting a logistic mixed-effects regression to the full dataset. The critical prediction of the ambiguity hypothesis is that effects of right context are primarily found at intermediate acoustic step values near the category boundary and are absent, or smaller, at continuum endpoints. This would result in a negative interaction between a quadratic function of continuum step and the effect of right context (cf. predictions in Fig. 2). The ideal observer, on the other hand, predicts a null effect for this interaction (as well as for the interaction between the linear function of the acoustic continuum and the effect of context). Thus, the regression for the combined analysis included right context and distance (deviation-coded in the same way as in the independent analysis), as well as (orthogonal)

linear and quadratic terms for continuum steps,[7] and interactions between context, distance, and continuum steps (for a total of 12 DFs). Following Gelman (2008), all categorical predictors were coded to have unit-distance (specifically, −.5 vs. + .5) and the continuous predictors (the orthogonal linear and quadratic components of continuum step) were divided through twice their standard deviations. This puts all predictors on a comparable scale, making comparison of effect sizes across predictors more meaningful. The random effect structure was again maximal, including random intercepts by participant and item, random slopes by participant for all fixed effect terms, and random slopes by item for all fixed effect terms except those including right context, which Szostak and Pitt manipulated between items (i.e., 78 by-participant DFs and 21 by-item DFs, for a total of 111 DFs inferred from 2,469 observations).

The full model output is reported in the supplementary information (SI). Table 1 summarizes the Bayesian hypothesis tests for the effects of interest. With regard to our primary goal, we ask whether we find credible effects of both the acoustic continuum and right context—indicating that listeners maintain subcategorical information—and whether the effect of right context interacts with the quadratic effect of the phonetic continuum. As shown in the first two rows of Table 1, we find "very strong" evidence both for a positive linear effect of the continuum on the log-odds of responding "sip" ($\widehat{\beta} = 9.88$, $BF \geq 19999$, $p_{posterior} > .999$) and for a positive main effect of right context ($\widehat{\beta} = 1.81$, $BF = 9999$, $p_{posterior} > .999$). Both effects are also evident in Fig. 6A. To further quantify the pervasiveness of these effects, Fig. 7 summarizes the two effects by participant. This goes beyond the intent of previous work, and should thus be interpreted with caution. In the Szostak and Pitt data, all participants exhibited evidence of both acoustic and context effects. As is often observed for cue integration, the magnitude of these effects traded-off against each other: participants who exhibited larger context effects also exhibited smaller effects of acoustic continuum.

Of primary interest, we find no support for the ambiguity hypothesis compared to its inverse—i.e., the hypothesis that the context effect exhibits a negative, rather than positive, interaction with quadratic continuum step ($\widehat{\beta} = 0.79$, $BF = 0.3$; $p_{posterior} < .24$). Put differently, there is "anecdotal", but not decisive, support *against* the weak ambiguity hypothesis ($BF_{\neg H1,\, H1,} = 1/BF_{H1,\, \neg H1,} = 1/.3 = 3.3$). The reason for this is apparent in Fig. 6B, which visualizes the effect of context across the continuum, as estimated by the combined analysis: if anything, the effect of context *in*creases as one moves away from the most ambiguous point in the continuum.[8]

With regard to our second goal, we ask whether the effect of right context decreases at longer lags—i.e., in the far, compared to the near, condition. That is, we ask whether there is a negative interaction between distance (far vs. near) and the context effect. In line with the original findings of Szostak and Pitt (2013), we find only "anecdotal" evidence in support of this hypothesis ($\widehat{\beta} = -.58$, $BF = 3.0$, $p_{posterior} >$ .74). There was "strong" evidence for a positive effect of right context even at the far distance, and "very strong" evidence at the near distance. Finally, to ascertain that the effect of context at the far distance is not driven by a decreased reliance on the acoustic input (e.g., responding purely based on the most recent words in the sentence), we also tested whether there was a negative interaction between distance and continuum. As summarized in Table 1, we found little support of such an interaction ($\widehat{\beta} = .96$, $BF = 0.3$, $p_{posterior} < .25$): the acoustic continuum had a "very strong" effects both at near and at far distances.

Discussion

Together, the independent and combined (re)analyses of Szostak and Pitt's data favor the additivity prediction of the ideal observer model over the strong or weak ambiguity hypothesis. The evidence is not, however, decisive. While the independent analysis estimated the effect of context to be of similar magnitude throughout the continuum (similar coefficient estimates, i.e., the points in Fig. 5B), it yielded only "moderate" evidence that right context has an effect at the continuum endpoints. The combined analysis revealed no evidence that context effects were smaller on the endpoints. If anything, we found anecdotal support for the opposite trend: in an analysis that contained interactions between context and linear as well as quadratic effects of the phonetic continuum, we find that the data favor *in*creased effects of context for less perceptually ambiguous continuum steps. This does not mean, of course, that the dependence of context effects on the phonetic continuum has exactly the quadratic shape shown in Fig. 6B. For example, just as high uncertainty about the effects at the continuum endpoints can affect the independent analyses, it is possible that the positive quadratic interaction is primarily driven by the three continuum steps in the center of the phonetic continuum (a possibility to which we return in the general discussion).

Notably, this trend—if confirmed in our remaining experiments—would also be unexpected under the ideal observer hypothesis, which predicts that right context effects are constant across the acoustic continuum. Finally, post-hoc frequentist power simulations presented in the SI found the power to detect the effect predicted by the weak ambiguity hypothesis in the combined analysis to be very low (~5%, see Figure S2). In short, the re-analysis of Szostak and Pitt's data provides no credible support for either version of the ambiguity hypotheses, but it also does not decisively reject it.

We thus conducted three experiments, which we present next. These experiments are designed to further distinguish between the ideal observer and the ambiguity hypothesis. They also shed light on the reasons for the discrepancy between Connine et al. (1991) and Szostak and Pitt (2013) in how long subcategorical information is maintained.

Experiment 1

We designed Experiment 1 to closely follow Connine et al. (1991, Experiment 1), except that participants could respond only after the sentence (following Szostak & Pitt, 2013). Post-hoc power simulations presented in the SI found that Experiment 1 provided substantially higher power to test the predictions of both the strong (>75 %) and the weak ambiguity hypothesis (50 %), compared to the Szostak and Pitt data (~5% and 30 %, respectively).

Method

Participants

Forty-eight workers on Amazon Mechanical Turk participated in the experiment between 10/26–11/07/2013. Two participants were excluded for performing the experiment twice. Seven additional participants were removed because they did not exhibit significant effects of VOT. The results of this and all other experiments reported below

[7] We employ orthogonal polynomials to reduce collinearity. For all predictions and visualizations, we transform the model fit back into the original continuum steps.

[8] As requested by reviewers, we repeated the same hypothesis tests while excluding the two continuum endpoints. This addresses the possibility that the two endpoints exert disproportionate influence on the results due to their distance from the mean of the continuum. The analyses removed between 33% (Experiments 1–3) to 40% (re-analysis of Szostak & Pitt, 2013) of the data, and are reported in full in the SI. Despite the substantial loss of data, there was always support for a positive context effect (Szostak and Pitt reanalysis: $BF = 29.4$; Exp 1: $BF = 7.8$; Exp 2: $BF = 415.7$; Exp 3: $BF = 25.9$). Support for the weak ambiguity hypothesis was *at best* anecdotal (Szostak and Pitt reanalysis: $BF = 1.5$) with most analyses delivering anecdotal to moderate evidence *against* the ambiguity hypothesis (Exp 1–3: $BF < 0.9$). We thus do not discuss this question further.

Table 1
Summary of the Bayesian hypothesis tests conducted over the combined analysis of Szostak and Pitt (2013, Experiment 2). Columns provide the estimated effect in log-odds ($\widehat{\beta}$), its standard error (SE), the 95% credible interval of the test, the associated Bayes factor (BF), and the posterior probability of the hypothesis (assuming uniform prior probabilities of the hypothesis being true vs. false). The first part of the table summarizes the effects of the acoustic continuum and right context on participants' responses. The second part summarizes tests assessing the predictions of the ambiguity and ideal observer hypotheses. The third part summarizes the test of whether the effects of the phonetic continuum and right context decrease at longer distances. See SI for full model summary.

Hypothesis	Est.	SE	CI_L	CI_U	BF	$p_{post}(h)$	
Continuum > 0	9.88	1.075	8.21	11.75	>19000.0	1.000	*
Context > 0	1.81	0.475	1.04	2.61	9999.0	1.000	*
Ctxt:Cont^2 < 0	0.79	1.128	−1.04	2.65	0.3	0.240	
Ctxt:Dist < 0	−0.58	0.889	−2.06	0.86	3.0	0.749	
Ctxt at near Dist > 0	2.10	0.610	1.12	3.11	1817.2	0.999	*
Ctxt at far Dist > 0	1.52	0.688	0.39	2.66	82.0	0.988	*
Dist:Cont < 0	0.96	1.432	−1.37	3.33	0.3	0.242	
Cont at near Dist > 0	8.92	1.775	6.11	11.95	>19000.0	1.000	*
Cont at far Dist > 0	10.84	1.806	8.04	13.97	>19000.0	1.000	*

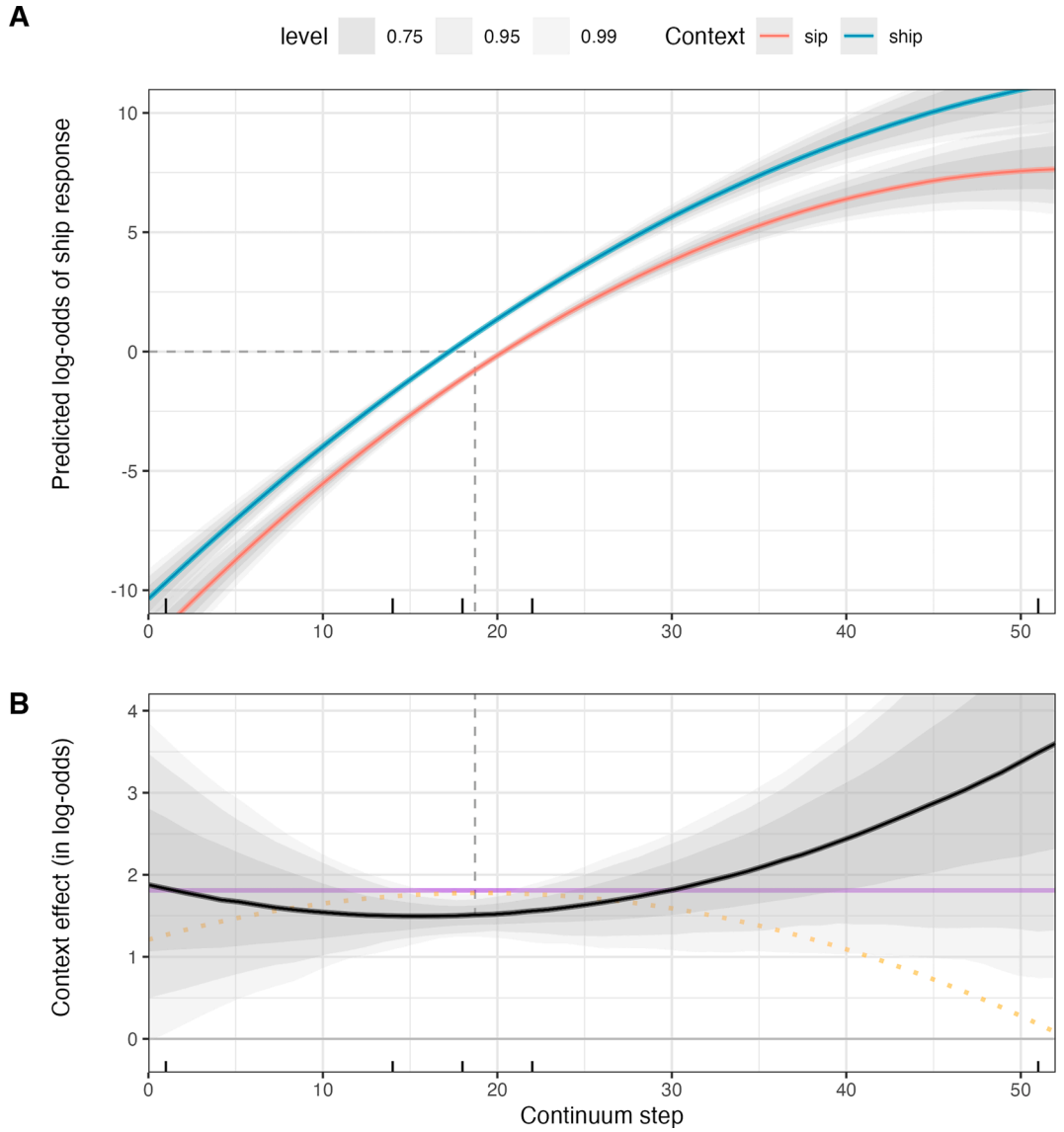

Fig. 6. Panel A: Marginal effects of continuum—including its linear and quadratic effect, and all their interactions with other predictors in the model—on participants' categorization responses in the combined analysis of Szostak and Pitt (2013, Experiment 2), shown for both context conditions. The dashed gray lines indicate the point of maximal ambiguity. Panel B: Marginal effect of context in the same combined analysis—i.e., the difference between the two lines in panel A. Also plotted are the qualitative predictions from the ideal observer (solid purple) and the weak ambiguity hypothesis (dotted orange), as implemented for the hypothesis tests over the combined model (see appendix for details). The continuum steps participants heard during the experiment are indicated by the upwards ticks along the x-axis. Shaded intervals show 75–99% CIs.

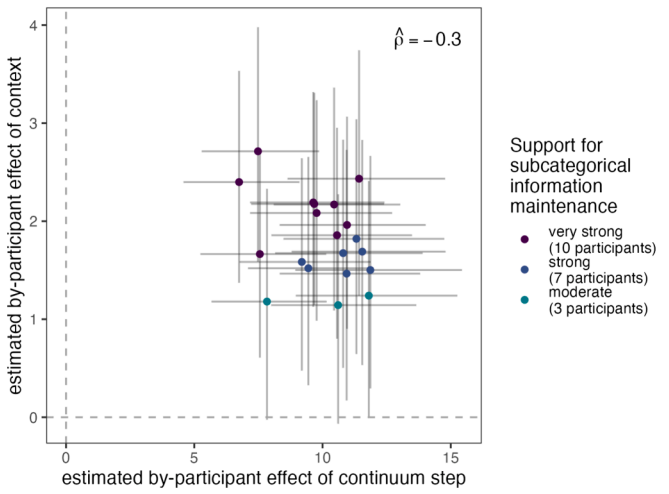

Fig. 7. Summary of participant-specific effects of the acoustic continuum and right context in the Szostak and Pitt data, as well as the correlation between these effects, derived from the Bayesian mixed-effect logistic regression for the combined analysis. Pointranges show 95% CIs. Support indicates the lower level of support between the effects of acoustics and right context. Note that participants who did not exhibit acoustic effects were removed prior to analysis.

qualitatively replicate when participants who do not use VOT are not excluded.

Materials

We used a tent–dent continuum and seven sets of sentence frames taken from Connine et al. (1991, Appendix A, sets 1–6 and 8). Each set had four conditions identical to those in Connine et al. (Fig. 1a), yielding a total of 28 sentence frames. For each set, the material preceding the target word was identical.

A female speaker in a noise-attenuated booth recorded the sentence frames with the target word biased by the right context. To create identical pre-target frames across the four versions, we concatenated one version of the first word for each set ("*when*" in example 1) with a single recording of "*the*", used for all sentences.

Following Connine et al., we created the *tent–dent* VOT continuum by cross-splicing onsets from a *tent* recording onto the rhyme of a *dent* recording (Fig. 8). For the latter, we selected a *dent* recording with a relatively long vowel in order to allow us to mimic the natural compensation between vowel duration and VOT. For example, to create a token with 45 ms VOT, we marked the location in the *tent* recording at which there had been 45 ms of VOT—i.e., 45 ms after the onset of the burst. We then cut the onset of this recording up the mark (blue shading in bottom-left of Fig. 8), and cross-spliced it before the rime of the *dent* recording minus its first 45 ms (blue shading in top-left of Fig. 8). While this cross-splicing approach has been used in many studies on speech perception (including Connine et al., 1991), we note that it implies that even the continuum endpoints are blends of two different recordings. For example, the /t/ endpoint was created by taking the onset from a "tent" recording, and the remainder of the rime from a "dent" recording.

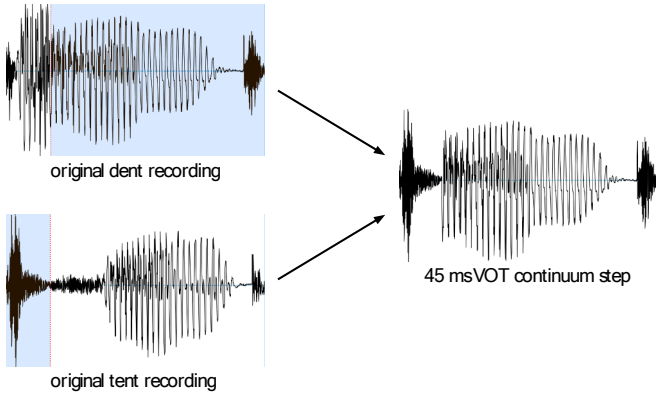

Fig. 8. Illustration of VOT continuum step creation. From two original recordings of "dent" and "tent" (**left**), each starting with the burst from the initial [t]/[d], we created a VOT continuum step waveform with a particular VOT (e.g., 45 ms, on **right**) by concatenating the first 45 ms of the "tent" waveform (in blue, lower left) with the remainder of the "dent" waveform starting at 45 ms after the burst (in blue, upper left). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Also following Connine et al., we then conducted a norming study in which participants identified members of this continuum without context. We then selected six VOT values: two unambiguous endpoints (10 and 85 ms) and four values around the category boundary (40, 45, 50, 55 ms). While the specific VOTs we selected differ from those in Connine et al.,[9] the procedure we employed to select the steps was identical to the original study. The 28 target frames were combined with the 6 VOTs to create 168 sentences. The sentences ranged in duration from 3 to 5 s total.

Procedure

After giving informed consent, participants listened to each of the sentences (following Connine et al.), in individually randomized order. Verbatim instructions to participants are provided in the SI. After each sentence, participants responded whether it contained *dent* or *tent* by pressing "x" or "m" on their keyboard (key assignment was counterbalanced across participants). There was no time-out. Before completing 168 trials of this type, participants completed 4 practice trials to ensure they understood the task, consisting of two example recordings of each continuum endpoint.

Results

Our analysis approach is identical to that employed in the re-analysis of the Szostak and Pitt (2013) data. The response data are visualized in proportion space in Fig. 9A.

Independent analysis

The independent analyses employ the same model, predictors, and random effects as in the re-analysis of the Szostak and Pitt data. The only difference was that the analyses of Experiment 1 included the full factorial variance–covariance matrix of random effects for both participant and items (10 DFs each; unlike Szostak and Pitt, 2013, our Experiments 1–3 manipulated right context within items).

Bayes factors, posterior probabilities, and 95 % credible intervals for the effect of right context in log-odds space for each continuum step are plotted in Fig. 9B. The results resemble those we found for the Szostak and Pitt (2013) data but are more decisive. We again find "strong" to "very strong" support for a context effect for the four steps near the category boundary ($p_{posterior}$s $\geq$.99), and "moderate" to "strong" support for context effects at the continuum endpoints. Compared the Szostak

[9] Our norming experiment (and the experiments reported below) found the "dent"-"tent" category boundary at somewhat larger VOT values than those reported in Connine et al. (1991). This could be due to differences in how VOT are measured, or due to differences in the speech rate of recordings (which is known to affect the perception of VOT, Miller et al., 1986). We were unable to assess these possibilities, as we did not have access to the recordings from Connine and colleagues (our own stimuli, and all materials and scripts used to construct them, are available on OSF). In other experiments with other clearly enunciated stimuli, we have found category boundaries at similarly large VOTs between 40–55 ms (e.g., Tan & Jaeger, 2024).

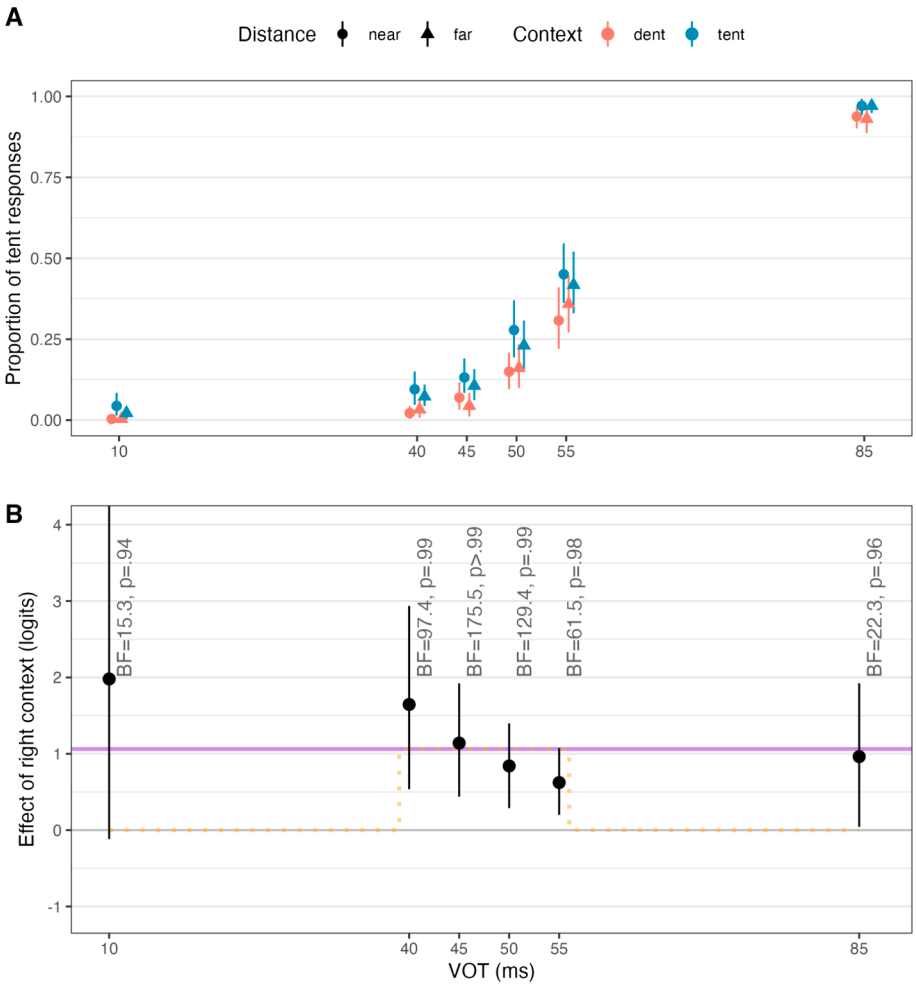

Fig. 9. Panel A: Proportion of 'tent' responses in each condition of Experiment 1. Intervals show 95% confidence intervals, bootstrapped over participant means. **Panel B:** Bayes factor and posterior probability of a positive context effect obtained from the independent Bayesian logistic mixed-effects regressions in Experiment 1. Point ranges show 95% CIs. Also plotted are schematic predictions from the ideal observer (solid purple) and the strong ambiguity hypothesis (dotted orange). For the predictions of the ideal observer, we set the purple line to the average context effect of all continuum steps except for the endpoints. This makes it informative that the credible intervals at the endpoints overlap with the purple line.

and Pitt data, the support for a context effect at the endpoints was stronger ("strong" at the /t/ endpoint, $p_{posterior} = .96$, and "moderate" at the /d/ endpoint, $p_{posterior} = .94$).

These results again favor the ideal observer over the strong ambiguity hypothesis, and do so more decisively than for the Szostak and Pitt data. Next, we present the combined analysis, which evaluates the weaker ambiguity hypothesis.

Combined analysis

The analysis was identical to the combined analysis for the Szostak and Pitt data, except that the full random effect structure now also included the full factorial random effects by item, as all manipulations were within item (for a total of 168 DFs inferred from 7,912 observations). The full model output is reported in the SI. Table 2 summarizes the Bayesian hypothesis tests for the effects of interest.

Table 2
Summary of the Bayesian hypothesis tests conducted over the combined analysis of Experiment 1. The first part of the table summarizes the effects of the phonetic continuum and right context on participants' responses. The second part summarizes tests assessing the predictions of the ambiguity and ideal observer hypotheses. The third part summarizes the test of whether the effects of the phonetic continuum and right context decrease at longer distances. See SI for full model summary.

Hypothesis	Est.	SE	CI_L	CI_U	BF	$p_{post}(h)$	
VOT>0	8.52	0.688	7.42	9.66	>19000.0	1.000	*
Context > 0	1.43	0.359	0.86	2.04	9999.0	1.000	*
Ctxt:VOT^2 < 0	1.54	0.749	0.38	2.84	0.0	0.011	
Ctxt:Dist < 0	−0.28	0.433	−0.98	0.43	3.0	0.748	
Ctxt at near Dist > 0	1.57	0.412	0.91	2.27	6665.7	1.000	*
Ctxt at far Dist > 0	1.29	0.427	0.60	2.00	1051.6	0.999	*
Dist:VOT<0	0.45	0.676	−0.63	1.55	0.3	0.244	
VOT at near Dist > 0	8.07	0.947	6.54	9.63	>19000.0	1.000	*
VOT at far Dist > 0	8.97	0.982	7.40	10.61	>19000.0	1.000	*

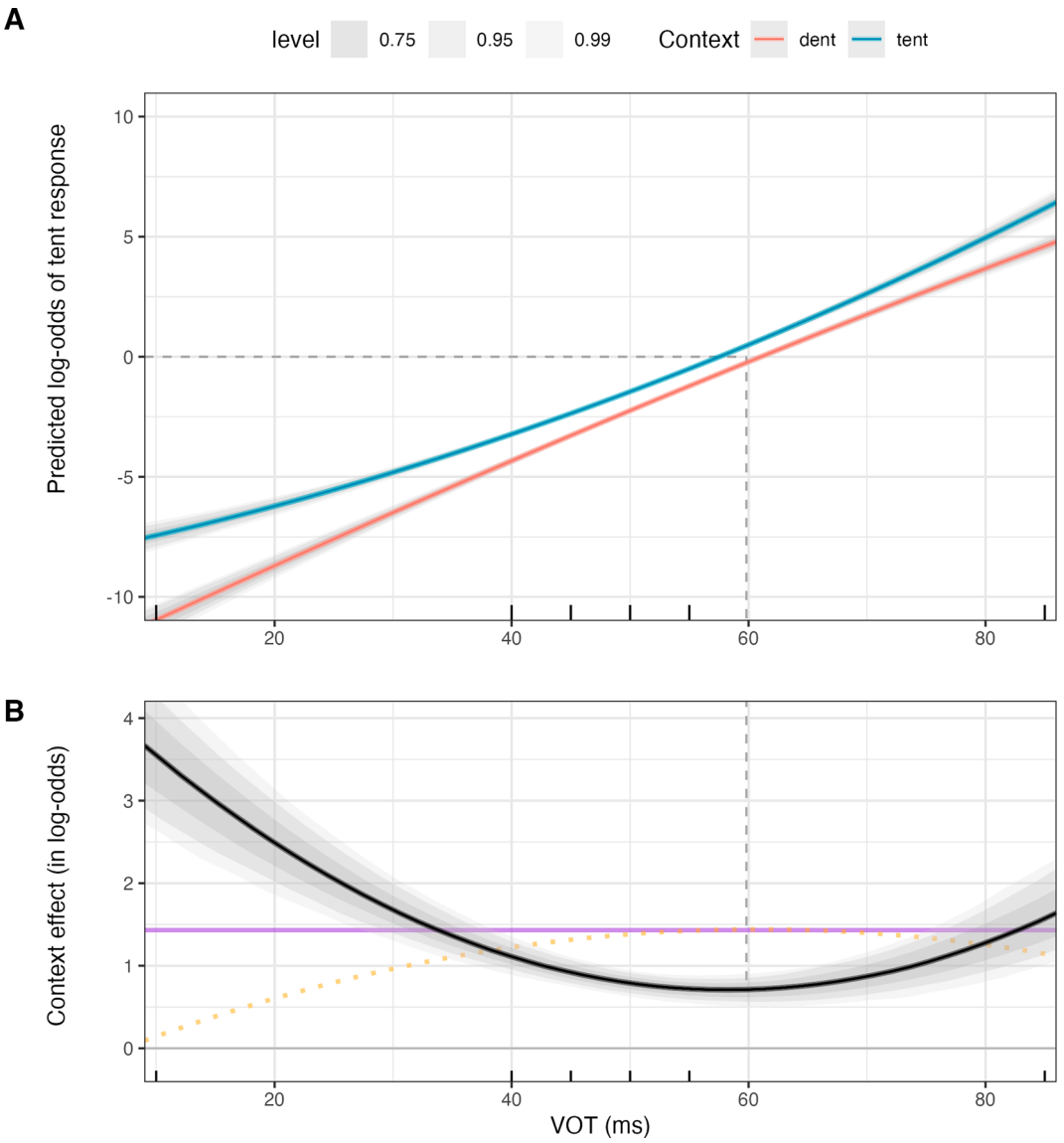

Fig. 10. Panel A: Marginal effects of continuum—including its linear and quadratic effect, and all their interactions with other predictors in the model—on participants' categorization responses in the combined analysis, shown for both context conditions in Experiment 1. The dashed gray lines indicate the point of maximal ambiguity. **Panel B:** Marginal effect of context in the same combined analysis—i.e., the difference between the two lines in panel A. Also plotted are the qualitative predictions from the ideal observer (solid purple) and the weak ambiguity hypothesis (dotted orange), as implemented for the hypothesis tests over the combined model. Shading in both panels shows credible intervals. The continuum steps participants heard during the experiment are indicated by the upwards ticks along the x-axis. Shaded intervals show 75–99% CIs.

With regard to our first goal, we find both "very strong" evidence for both a positive linear effect of VOT on the log-odds of responding "tent" ($\widehat{\beta} = 8.52$, BF ≥ 19999, $p_{posterior} > .999$) and a positive main effect of right context ($\widehat{\beta} = 1.43$, BF ≥ 9999, $p_{posterior} > .999$). Fig. 11 shows that—like in Szostak and Pitt's data—all participants exhibit some evidence for both effects, and the magnitude of these effects seems to trade-off against each other. Critically, we find no support for the weak ambiguity hypothesis compared to its inverse ($\widehat{\beta} = 1.54$, BF < 0.1, $p_{posterior} < .02$). Compared to the reanalysis of Szostak and Pitt's data, the support *against* the ambiguity hypothesis was stronger ($BF_{\neg H1,\, H1,} = 1/BF_{H1,\, \neg H1,} = 87.9$). The reason for this is apparent in the visualization of the context effect across the continuum (Fig. 10B): the combined analysis finds that the effect of context *in*creases for perceptually less ambiguous continuum steps, and this effect is clearer in Experiment 1 than in the Szostak and Pitt data.

With regard to the second goal, we find only "anecdotal" evidence for the hypothesis that effects of right context decrease with increasing distance ($\widehat{\beta} = -.28$, BF $= 3.0$, $p_{posterior} > .74$): there was "very strong" evidence for a positive effect of right context at both the near and the far distance. Finally, there was "moderate" evidence that the effect of VOT was *larger* at far distances ($\widehat{\beta} = .45$, BF $= 3.0$, $p_{posterior} > .75$), though there was "very strong" evidence for an effect of VOT at both distances.

Discussion

The independent and combined analyses of Experiment 1 support similar conclusions as in the re-analysis of Szostak and Pitt's data. They provide evidence relevant to our two goals of (1) testing the predictions of the ideal observer model quantitatively and (2) determining whether comprehenders can maintain subcategorical information for 6–8 syllables, even with stop contrast stimuli modeled after those used by Connine et al.

Relevant to our primary goal, right context is estimated to have a positive effect in log-odds space on all VOT steps, as predicted by the

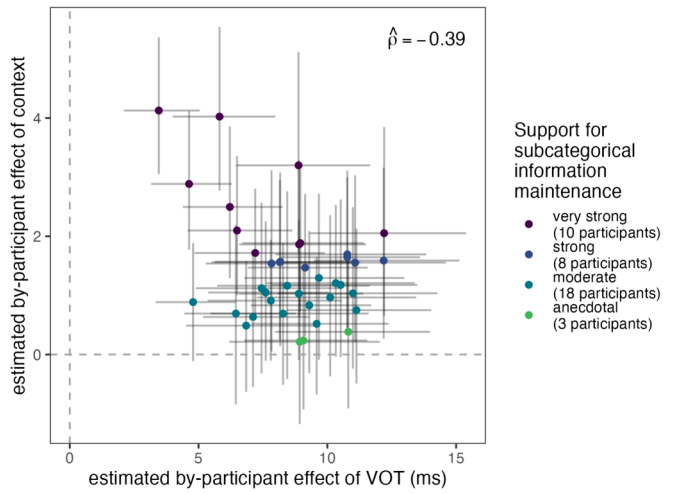

Fig. 11. Summary of participant-specific effects of VOT and right context in Experiment 1, as well as the correlation between these effects, derived from the Bayesian mixed-effect logistic regression for the combined analysis. Pointranges show 95% CIs. Support indicates the lower level of support between the effects of VOT and right context. Note that participants without a VOT effect were removed prior to analysis.

ideal observer model. Compared to the Szostak and Pitt data, the evidence for a context effect at the continuum endpoints is stronger in Experiment 1, though it is worth noting that the evidence is again weaker at one endpoint than the other. For both the fricative continuum in the re-analysis of Szostak and Pitt's data, and the VOT continuum in Experiment 1, this was the left continuum endpoint. This is the endpoint for which responses are particularly close to categorical (see log-odds at the left endpoint in Fig. 10A, compared to right endpoint), resulting in particularly low power to detect an effect of right context (recall Fig. 4; confirmed also specifically by power simulations for Experiment 1, summarized in Figure S2A in the SI).

Relevant to our second goal, our findings tease apart the two competing explanations for the difference in results between Connine et al. (1991) and Szostak and Pitt (2013). Experiment 1 used stop contrast stimuli modeled after those from Connine et al. with a procedure similar to that of Szostak and Pitt. The fact that the results of Experiment 1 parallel those of Szostak and Pitt suggests that the procedural difference between the Connine et al and Szostak and Pitt studies is sufficient to explain the difference in their results. That is, comprehenders *can* maintain subcategorical information about even stop contrasts in relatively syntactically complex sentences for at least 6–8 syllables, provided they are not allowed to respond before the end of the sentence. That said, both our reanalysis of Szostak and Pitt's data and Experiment 1 also find "anecdotal" evidence that the effect of right context decreases with increasing distance. This trend thus is something to revisit in our remaining experiments.

In summary, both experiments analyzed so far favor the ideal observer over the ambiguity hypothesis, suggesting that listeners maintain some subcategorical information for all tokens. However, the data we have analyzed so far leaves us with an important caveat to this conclusion: we observe a *positive* interaction between context and the quadratic effect of the acoustic continuum. This effect is the opposite of what the ambiguity hypothesis predicts. It is, however, also unexpected under the ideal observer hypothesis. Recall that one of the features of the ideal observer approach is that when the data are inconsistent with the ideal observer, then additional mechanisms are necessary to explain human behavior. We return to this point in the general discussion. The evidence for this unexpected effect was stronger in Experiment 1 than in the Szostak and Pitt data. This provides additional motivation for the remaining two experiments.

Experiment 2

The primary goal of Experiments 2 and 3 is to replicate Experiment 1. Instead of an exact replication, however, we used a different set of VOT steps. Specifically, we aimed to choose VOT steps that increase the statistical power to test the log-odds additivity prediction of the ideal observer through the combined analysis. To this end, Experiment 2 includes VOT steps that are expected to fall into the midrange between the most and least ambiguous VOTs. At these intermediate points, the ambiguity hypothesis predicts decreasing effects of right context, whereas the ideal observer does not. Power simulations presented in the SI confirm that this goal was achieved (Figure S2): we estimate the power to detect the predicted decrease in the context effect in Experiment 2 at about 85 % (compared to 75 % in Experiment 1; the power to detect a main effect of context was estimated to be close to 100 %).

Experiment 2 was originally reported as part of a study on task effects on subcategorical information maintenance (non-archival, Bushong and Jaeger, 2017). Both experiments in that study exhibit the effects reported here. In choosing Experiment 2 for presentation in the present article, our choice was solely driven by a preference to keep the design across Experiments 1–3 as similar as possible, avoiding the need to introduce additional manipulations not relevant to the present purpose. In the general discussion, we return to Bushong and Jaeger (2017).

Method

Participants

Forty-eight workers on Amazon Mechanical Turk participated in the experiment between 11/16–11/17/2016. Nine participants were removed because they did not exhibit significant effects of VOT.

Materials

Stimuli were identical to Experiment 1 except for the difference in VOTs. Based on power simulations over Experiment 1 (reported in the SI), we decided to use VOTs of 10, 40, 50, 60, 70, and 85 ms instead of those used in Experiment 1 (10, 40, 45, 50, 55, 85). We included the endpoints because of their relevance to the hypotheses of interest, despite power being lowest at those VOTs. The other four VOT steps were chosen to cover the whole range of the *perceptual* continuum from strongly dent-biased to strongly tent-biased (see also Fig. 12A below for demonstration that this goal was indeed achieved). Additional power simulations that guided the design of Experiment 2 are presented in the SI (Figure S4).

Procedure

Procedure was identical to Experiment 1 with one exception. Four pseudorandomized lists were created instead of each participant hearing an individually randomized order.

Results

Analyses were identical to those in Experiment 1. The response data are visualized (in proportion space) in Fig. 12A.

Independent analysis

Bayes factors, posterior probabilities, and 95 % credible intervals for the effect of right context in log-odds space for each continuum step are plotted in Fig. 12B. With one exception, the results of Experiment 1 replicate closely. We find "moderate" to "very strong" evidence in support of positive context effects for all four VOT steps near the category boundary ($p_{posterior}$s $\geq$.93). We also replicate the "strong" support for a positive context effect at the /t/ endpoint ($p_{posterior}$ = .99). At the /d/ endpoint, however, we find "anecdotal" evidence *against* a positive context effect (BF = 2.0, $p_{posterior}$ = .68). In line with power simulations (see Figure S3 in the SI), the credible intervals of this effect indicate large uncertainty. In particular, the 95 % credible intervals also include

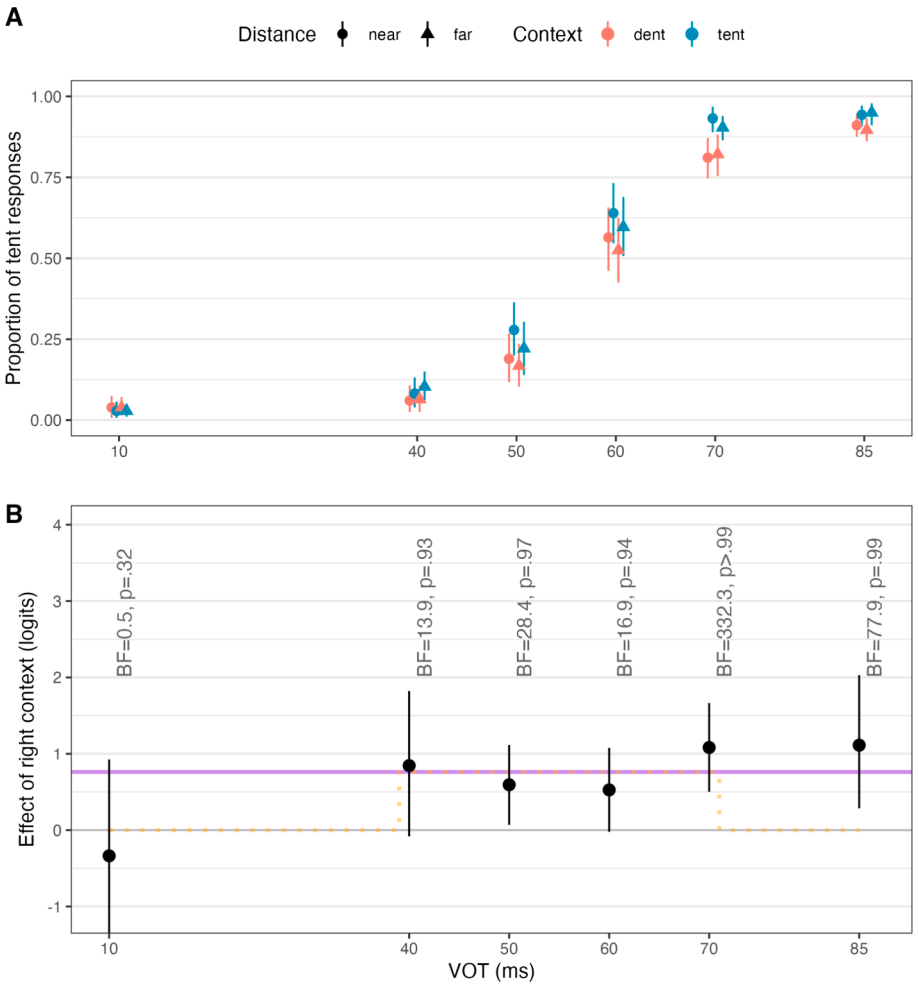

Fig. 12. Panel A: Proportion of 'tent' responses in each condition of Experiment 2. Error bars show 95% confidence intervals, bootstrapped over participant means. **Panel B:** Bayes factor and posterior probability of a positive context effect obtained from the independent Bayesian logistic mixed-effects regressions in Experiment 2. Point ranges show 95% CIs. Also plotted are schematic predictions from the ideal observer (solid purple) and the strong ambiguity hypothesis (dotted orange).

the predictions of the ideal observer (purple line).

While these results are more mixed than those of Experiment 1, the "very strong" support for a context effect at the /t/ endpoint argues against the strong ambiguity hypothesis. Next, we turn to the combined analysis to test the weak ambiguity hypothesis.

Combined analysis

The full model output is reported in the SI. Table 3 summarizes the Bayesian hypothesis tests for the effects of interest.

Replicating Experiment 1, we find both "very strong" evidence for both a positive linear effect of VOT ($\widehat{\beta} = 9.17$, BF ≥ 19999, $p_{posterior} > .999$) and a positive main effect of right context ($\widehat{\beta} = .80$, BF $= 216.4$, $p_{posterior} > .995$). Of note, the context effect was somewhat smaller than in Experiment 1, though the 95 % CIs overlapped. This is also reflected in Fig. 14: while the majority of participants exhibit some evidence for both effects, this was not the case for all participants (additionally, unlike in the two data sets analyzed so far, we do not see that trade-off in the participant-specific magnitude of VOT and right context effects). Critically, Experiment 2 replicates the lack of support for the weak

Table 3
Summary of the Bayesian hypothesis tests conducted over the combined analysis of Experiment 2. The first part of the table summarizes the effects of the phonetic continuum and right context on participants' responses. The second part summarizes tests assessing the predictions of the ambiguity and ideal observer hypotheses. The third part summarizes the test of whether the effects of the phonetic continuum and right context decrease at longer distances. See SI for full model summary.

Hypothesis	Est.	SE	CI_L	CI_U	BF	$p_{post}(h)$	
VOT>0	9.17	0.908	7.73	10.70	>19000.0	1.000	*
Context > 0	0.80	0.315	0.29	1.32	216.4	0.995	*
Ctxt:VOT^2 < 0	0.87	0.663	−0.19	1.98	0.1	0.088	
Ctxt:Dist < 0	0.04	0.379	−0.57	0.66	0.8	0.459	
Ctxt at near Dist > 0	0.78	0.369	0.19	1.39	57.7	0.983	*
Ctxt at far Dist > 0	0.82	0.366	0.23	1.42	81.6	0.988	*
Dist:VOT<0	−0.32	0.452	−1.06	0.40	3.4	0.772	
VOT at near Dist > 0	9.49	1.028	7.85	11.21	>19000.0	1.000	*
VOT at far Dist > 0	8.85	0.999	7.26	10.52	>19000.0	1.000	*

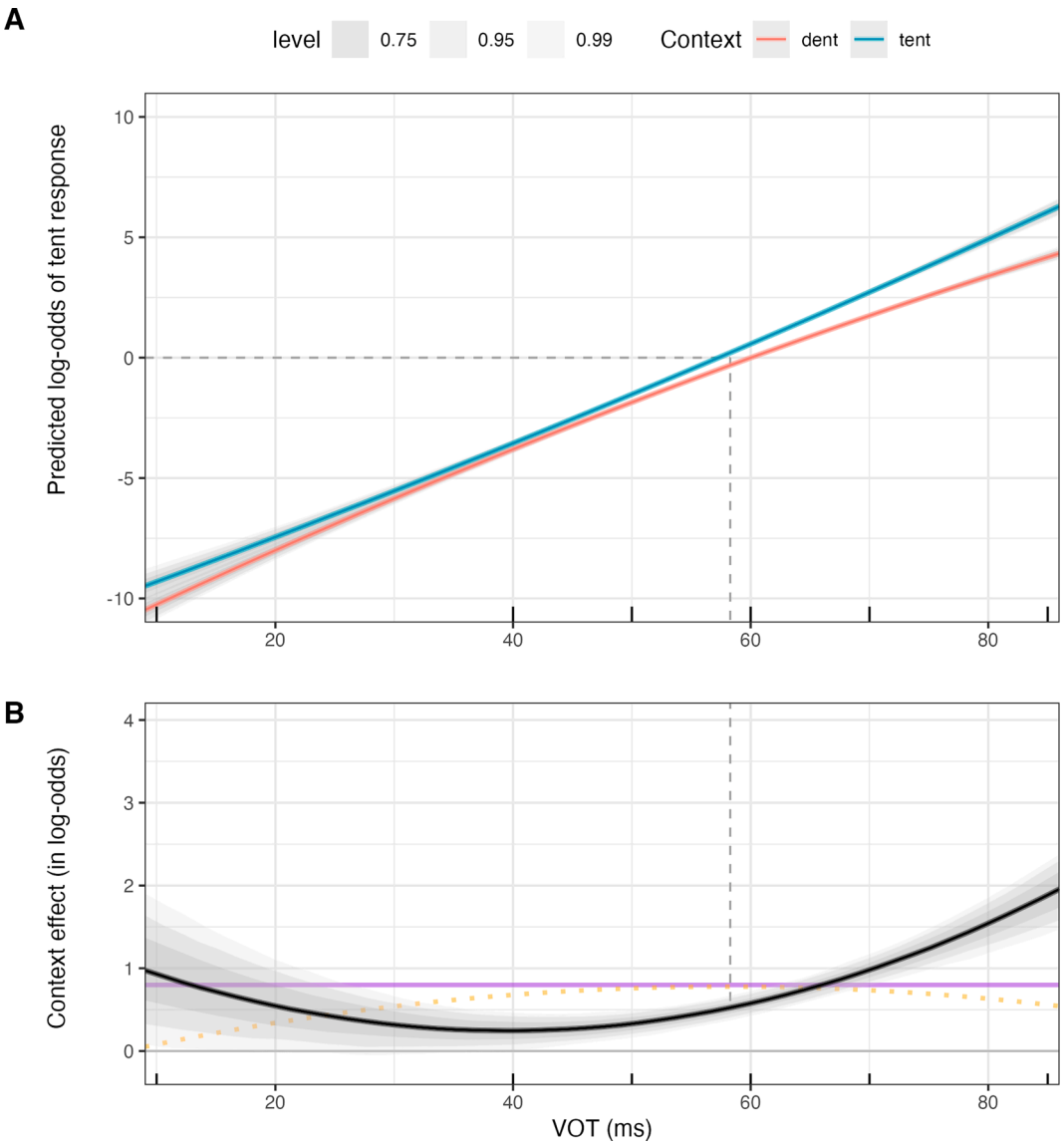

Fig. 13. Panel A: Marginal effects of continuum—including its linear and quadratic effect, and all their interactions with other predictors in the model—on participants' categorization responses in the combined analysis, shown for both context conditions in Experiment 2. The dashed gray lines indicate the point of maximal ambiguity. **Panel B:** Marginal effect of context in the same combined analysis—i.e., the difference between the two lines in panel A. Also plotted are the qualitative predictions from the ideal observer (solid purple) and the weak ambiguity hypothesis (dotted orange), as implemented for the hypothesis tests over the combined model. Shading in both panels shows credible intervals. The continuum steps participants heard during the experiment are indicated by the upwards ticks along the x-axis. Shaded intervals show 75–99% CIs.

ambiguity hypothesis ($\widehat{\beta} = .87$, BF $= .1$; $p_{posterior} < .09$). The support *against* the ambiguity hypothesis was "moderate" ($BF_{\neg H1,\, H1,} = 10.4$), somewhat weaker than in Experiment 1. Of interest, this was again driven by a positive quadratic trend in the context effect (Fig. 13B).

With regard to our second goal, we find "anecdotal" support *against* the hypothesis that effects of right context decrease with increasing distance ($\widehat{\beta} = .04$, BF $= 1.2$, $p_{posterior} > .54$). Replicating Experiment 1 and the Szostak and Pitt data, the effect of right context was robust at both lags: there was "strong" evidence for a positive effect of right context at both the near and the far distance. Finally, there was only "moderate" evidence that the effect of VOT was weaker at far distances ($\widehat{\beta} = -.32$, BF $= 3.3$, $p_{posterior} > .77$), and there was "very strong" evidence for an effect of VOT at both distances.

Discussion

Experiment 2 provides a partial replication of Experiment 1, while extending the results to previously untested continuum steps. As in Experiment 1, we found strong evidence for a right context effect at upper continuum endpoint (85 ms). However, at the lower endpoint (10 ms), the effect of context was consistent with the null predicted by the strong ambiguity hypothesis. As in Experiment 1, this was the endpoint at which responses were particularly close to categorical (see Fig. 13A), resulting in particularly small power (as we showed in Fig. 4; confirmed also specifically for Experiment 2 in Fig. S2B in the SI). When the data from all VOT steps were combined, we again found "moderate" evidence against even the weak ambiguity hypothesis. Experiment 2 does, however, also replicate the positive interaction between right context and the quadratic effect of VOT. This makes it the third data set, in which we find this effect—contrary to the predictions of the ideal observer.

For Experiment 2, the positive quadratic identified by the combined analysis trend might initially appear surprising in light of the independent analyses. In particular, if one aims to connect a quadratic curve through the points in Fig. 12B, the /d/ endpoint would not fall into that

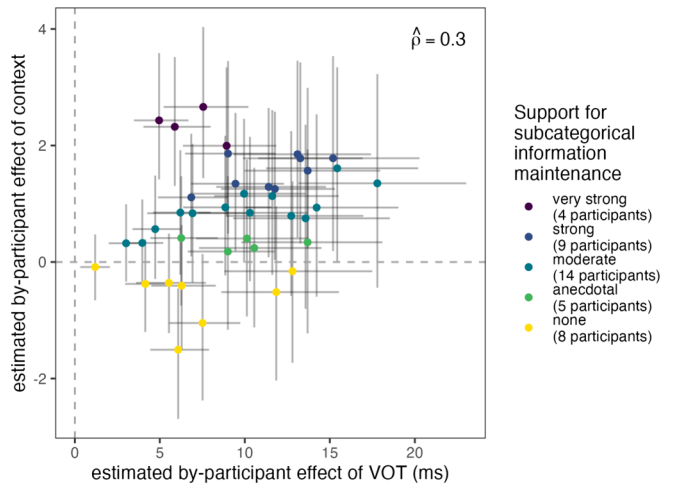

Fig. 14. Summary of participant-specific effects of VOT and right context in Experiment 2, as well as the correlation between these effects, derived from the Bayesian mixed-effect logistic regression for the combined analysis. Pointranges show 95% CIs. Support indicates the lower level of support between the effects of VOT and right context. Note that participants without a VOT effect were removed prior to analysis.

curve. However, unlike the independent analysis, the combined analysis does not consider each VOT step on its own. Instead, the combined analysis asks what type of quadratic trend (positive, negative, or null) best describes the context effect across *all* steps along the VOT continuum. It does so while accounting for *uncertainty* about the true context effect at each VOT step (which is inevitably largest at the endpoints). With this in mind, we make two observations. First, in all three data sets analyzed so far, the context effect inferred by the combined analysis always went through the 95 % CIs of all effects estimated by the independent analyses. Second, the most ambiguous continuum step always had the smallest context effect in the independent analysis (step 22 in Szostak & Pitt's data; VOTs of 55 and 60 in Experiments 1 and 2, respectively). These two considerations make apparent that the combined and independent analyses are not in conflict (rather, they complement each other).

We again emphasize that the combined analysis leaves open whether the positive quadratic trend is primarily driven by the continuum midpoints or endpoints, or equally by both. This also means that the results of the combined analyses do not necessarily entail that the effects of right context continue to increase indefinitely for less and less ambiguous continuum steps, or even that the effects of context are necessarily larg*est* at the continuum endpoints. All that can be concluded from the combined analyses is that, *on average*, the effects of right context increase as one moves from the continuum midpoint towards perceptually

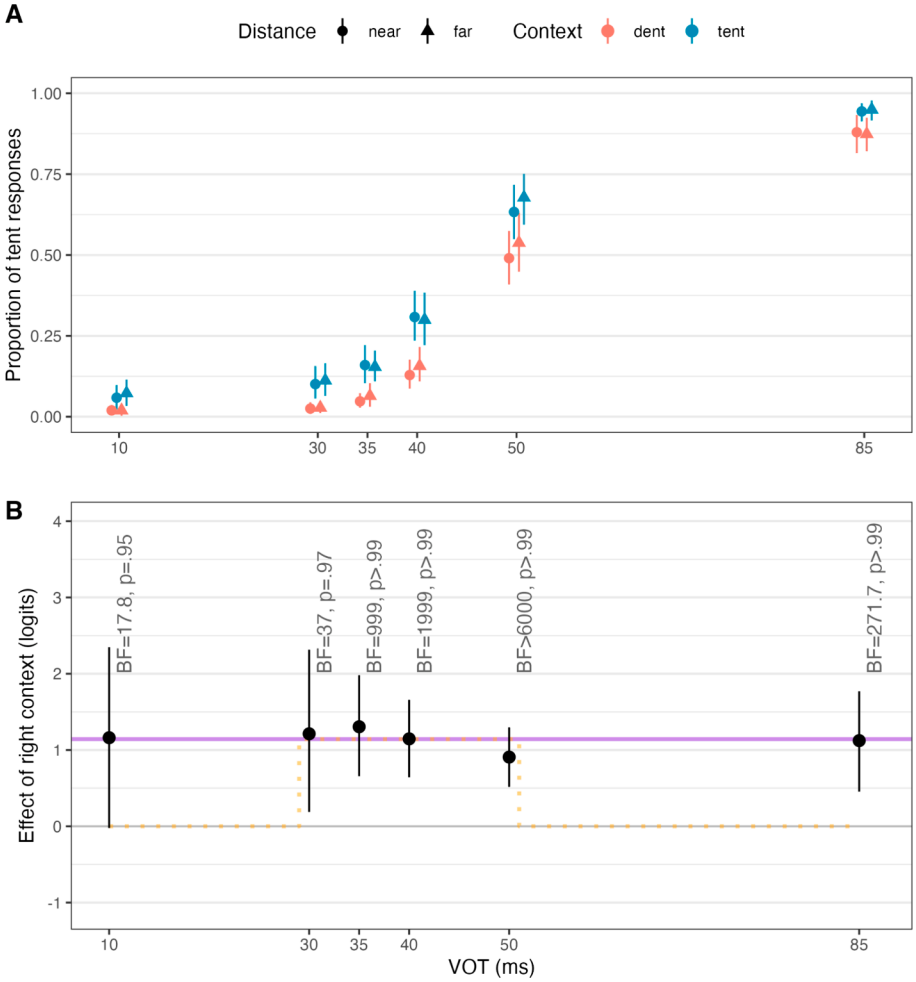

Fig. 15. Panel A: Proportion of 'tent' responses in each condition of Experiment 3. Error bars show 95% confidence intervals, bootstrapped over participant means. **Panel B:** Bayes factor and posterior probability of a positive context effect obtained from the independent Bayesian logistic mixed-effects regressions in Experiment 3. Point ranges show 95% CIs. Also plotted are schematic predictions from the ideal observer (solid purple) and the strong ambiguity hypothesis (dotted orange).

less ambiguous steps—it is possible, for example, that the positive quadratic trend observed in the combined analyses is primarily driven by the continuum midpoints. Regardless of the specific reason for the quadratic trend, it is unexpected under not only the ambiguity hypothesis but also the ideal observer hypothesis. We return to this point in the general discussion.

Finally, Experiment 2 also replicated strong support for an effect of right context for stop contrast stimuli even in the far condition, providing further evidence that comprehenders can maintain subcategorical information about, not just fricatives, but also stop contrasts for at least 6–8 syllables. Whereas Experiment 1 found "moderate "support for a decrease of the context effect with increasing lag, Experiment 2 found "anecdotal" support against this hypothesis.

Experiment 3

Experiment 3 is a re-analysis of data originally reported in Bushong and Jaeger (2019a), as part of two experiments on the role of cue conflict, summarized in more detail in the general discussion. Both experiments in that study exhibit the effects reported here. In presenting Experiment 3 here, our choice was solely driven by a preference to keep the design across Experiments 1–3 as similar as possible.

Compared to Experiments 1 and 2, Experiment 3 increases the number of items across participants (to further increase power), and slightly increases the number of participants. We again chose slightly different VOT steps, this time increasing our resolution towards the /d/ endpoint. This provides additional power to detect whether the effect of context decreases, stays constant, or increases towards the /d/ endpoint—the side of the VOT continuum for which we have consistently observed lower power (see Figure S3 in the SI). Post-hoc power analyses for Experiment 3 estimate the power to detect the decreasing context effect towards the endpoint predicted by the weak ambiguity hypothesis at about 85 % (the power to detect the main effect of right context was close to 100 %, Figure S2).

Method

Participants

Sixty workers on Amazon Mechanical Turk participated in the experiment between 11/07–11/08/2017. Nine participants were removed because they did not exhibit significant effects of VOT.

Materials

Stimuli were similar to those used in Experiments 1 and 2 except that we created 13 new sets of sentence frames, for a total of 20 sets (each in its four context x distance conditions yielding 80 unique sentences). The recording speaker was the same as the speaker used for Experiments 1 and 2. We created a new VOT continuum following the same procedure as in Experiments 1 and 2 but used different VOT steps. Specifically, we selected six VOT values: the same two unambiguous endpoints as in Experiments 1 and 2 (10 and 85 ms) and four values around the category boundary (30, 35, 40, 50 ms), based on a norming study conducted on a separate set of participants. The 80 sentence frames were combined with the 6 VOTs to create 480 recordings.

Procedure

Procedure was identical to Experiment 2 with one exception. To keep the experiment reasonably short, we distributed the 20 sentence frames across 20 pseudorandomized lists such that each list contained 7 of the sentence frames and all sentence frames were sampled equally often across participants.

Results

Analyses were identical to those in Experiments 1 and 2. The response data are visualized (in proportion space) in Fig. 15A.

Independent analysis

Bayes factors, posterior probabilities, and 95 % credible intervals for the effect of right context in log-odds space for each continuum step are plotted in Fig. 15B. We again find strong to "very strong" evidence in support of a positive context effect for the four VOT steps near the category boundary ($p_{posterior}$s $\geq$.97). Replicating Experiment 1, we also find support for a context effect at the /d/ and /t/ endpoints. This support was "moderate" at the /d/ endpoint ($p_{posterior} = .95$) and "very strong" at the /t/ endpoint ($p_{posterior} = .99$). Both BFs are more decisive than in Experiments 1 and 2. These results are predicted by the ideal observer, and incompatible with the hypothesis that subcategorical information is only maintained for highly ambiguous tokens. Finally, we note that the most ambiguous VOT step again exhibits the smallest context effect, replicating the previous three experiments we analyzed.

Combined analysis

The full model output is reported in the SI. Table 4 summarizes the Bayesian hypothesis tests for the effects of interest.

Replicating Experiments 1 and 2, we find both "very strong evidence" for both a positive linear effect of VOT ($\widehat{\beta} = 8.44$, BF $\geq$ 19999, $p_{posterior} > .999$) and a positive main effect of right context ($\widehat{\beta} = 1.36$, BF = 19999, $p_{posterior} > .999$). In line with the larger effects for right context, compared to Experiment 2, Fig. 17 shows that most participants exhibit evidence for both effects. We also replicate the trade-off in the magnitude of VOT and right context effects (observed in Experiment 1 and the Szostak and Pitt data, but not Experiment 2). Critically, Experiment 3 provides a third replication of the lack of support for the weak ambiguity hypothesis ($\widehat{\beta} = 1.11$, BF $< .1$; $p_{posterior} < .009$). This time the support *against* the ambiguity hypothesis was "strong" ($BF_{\neg H1,\ H1,} = 127.2$). As in Experiments 1 and 2, this was again driven by a positive quadratic trend in the context effect (Fig. 16B).

With regard to our second goal, Experiment 3 closely replicates Experiments 1 and 2, providing "anecdotal" evidence *against* the hypothesis that effects of right context decrease with increasing distance ($\widehat{\beta} = .11$, BF $= 2.0$, $p_{posterior} > .67$) and the hypothesis that the effects of VOT decrease with increasing distance ($\widehat{\beta} < .01$, BF $= 1.0$, $p_{posterior} > .50$). Replicating and further strengthening the results of Experiments 1 and 2, there was "very strong" evidence for positive VOT and context effects at both lags.

Discussion

Experiment 3 provides a close replication of all the crucial results from Experiments 1 and 2 and adds even stronger support for an effect of right context at both VOT endpoints. Unsurprisingly, this support was

Table 4
Summary of the Bayesian hypothesis tests conducted over the combined analysis of Experiment 3. The first part of the table summarizes the effects of the phonetic continuum and right context on participants' responses. The second part summarizes tests assessing the predictions of the ambiguity and ideal observer hypotheses. The third part summarizes the test of whether the effects of the phonetic continuum and right context decrease at longer distances. See SI for full model summary.

Hypothesis	Est.	SE	CI_L	CI_U	BF	$p_{post}(h)$	
VOT>0	8.44	0.712	7.29	9.63	>19999.0	1.000	*
Context > 0	1.36	0.283	0.90	1.83	>19999.0	1.000	*
Ctxt:VOT^2 < 0	1.11	0.505	0.33	1.98	0.0	0.008	
Ctxt:Dist > 0	0.11	0.253	−0.31	0.53	0.5	0.328	
Ctxt at near Dist > 0	1.30	0.313	0.79	1.82	>19999.0	1.000	*
Ctxt at far Dist > 0	1.41	0.307	0.92	1.93	>19999.0	1.000	*
Dist:VOT<0	0.00	0.364	−0.60	0.59	1.0	0.495	
VOT at near Dist > 0	8.43	0.806	7.15	9.79	>19999.0	1.000	*
VOT at far Dist > 0	8.44	0.794	7.16	9.76	>19999.0	1.000	*

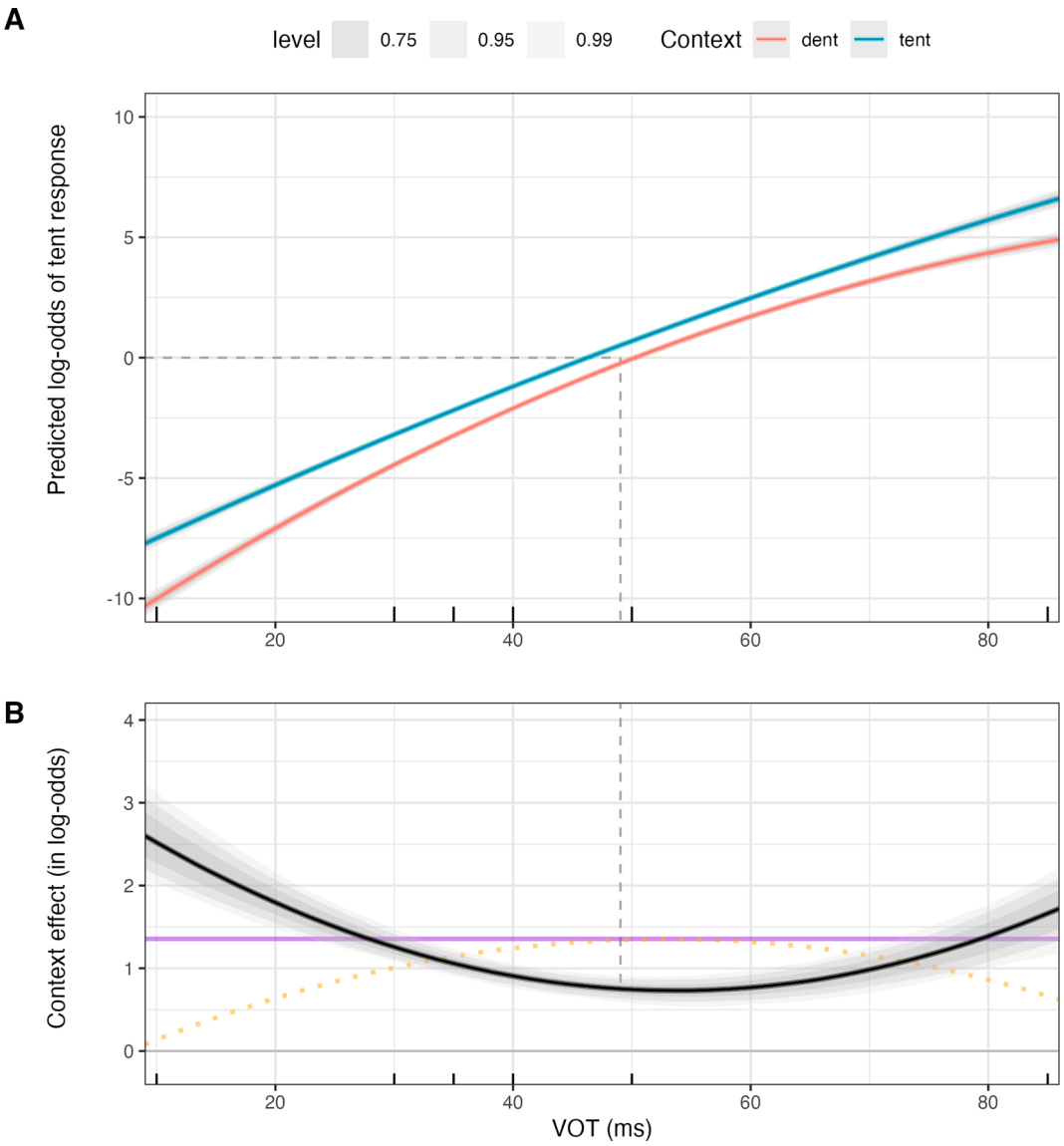

Fig. 16. Panel A: Marginal effects of continuum—including its linear and quadratic effect, and all their interactions with other predictors in the model—on participants' categorization responses in the combined analysis, shown for both context conditions in Experiment 3. The dashed gray lines indicate the point of maximal ambiguity. **Panel B:** Marginal effect of context in the same combined analysis—i.e., the difference between the two lines in panel A. Also plotted are the qualitative predictions from the ideal observer (solid purple) and the weak ambiguity hypothesis (dotted orange), as implemented for the hypothesis tests over the combined model. Shading in both panels shows credible intervals. The continuum steps participants heard during the experiment are indicated by the upwards ticks along the x-axis. Shaded intervals show 75–99% CIs.

again less strong at the /d/ endpoint, which again was the endpoint at which responses were more categorical (Fig. 16A). We also replicate the lack of credible reductions in either context or VOT effects at the longer lag.

Experiment 3 is the fourth dataset for which we find a *positive* interaction between the context effect and the quadratic effect of the phonetic continuum. Indeed, the support for this effect was strongest in Experiment 3, compared to all other datasets. The presence of this trend is incompatible with the ambiguity hypothesis but is also unexpected under the ideal observer hypothesis. We discuss this effect further below.

General discussion

We began this paper by pointing out two potential limitations on the maintenance of subcategorical information for sub-phonemic detail for phonemic contrasts that distinguish between alternative words suggested by previous work: (1) subcategorical information may be only, or primarily, maintained for the most ambiguous tokens near a category boundary (the ambiguity hypothesis) and (2) subcategorical information may be maintained for less than 6–8 syllables, at least for some types of phonological contrasts. We begin by discussing our findings regarding the latter limitation, and then turn to the ambiguity hypothesis.

Do listeners maintain information about fricatives and stops differently?

With respect to the second potential limitation, prior studies, differing in stimuli and procedure, yielded different results. Connine et al. (1991), who used stop contrasts and allowed participants to respond before hearing the relevant right context, found no evidence that subcategorical information was being maintained 6–8 syllables downstream. Szostak and Pitt (2013), who used fricative contrasts and required participants to hear the relevant right context before

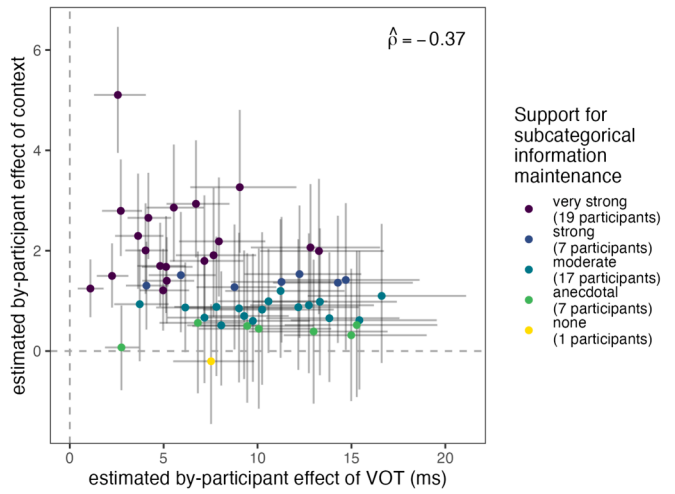

Fig. 17. Summary of participant-specific effects of VOT and right context in Experiment 3, as well as the correlation between these effects, derived from the Bayesian mixed-effect logistic regression for the combined analysis. Pointranges show 95% CIs. Support indicates the lower level of support between the effects of VOT and right context. Note that participants without a VOT effect were removed prior to analysis.

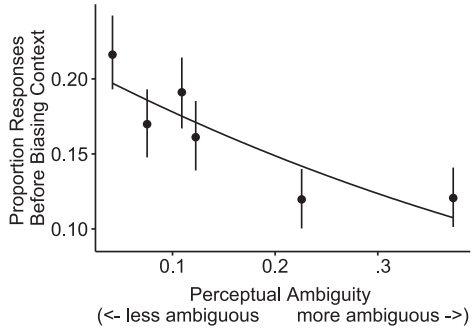

Fig. 18. When listeners are allowed to respond before the end of the sentence recording, they are more likely to respond early in the trial, before encountering right context, when perceptual evidence is less ambiguous. (Reprinted Fig. 4 from Bushong & Jaeger, 2017).

responding, found evidence that subcategorical information was being maintained even 8–9 syllables downstream. Szostak and Pitt conjectured that the differences in the results were likely caused by differences in the stimuli. For example, one specific hypothesis entertained by Szostak and Pitt was that the perceptual memory for fricatives might be longer lasting than that for stops.

Two later studies also focused on fricative contrasts (Brown-Schmidt & Toscano, 2017, Exp. 3; Falandays, Brown-Schmidt & Toscano, 2020). Both studies crossed an acoustic continuum (in their case, a fricative continuum from he to she) and right context (strongly biasing towards an interpretation of either he or she). The disambiguating right context occurred 6–7 syllables (5 words) downstream (Brown-Schmidt & Toscano, 2017)—resembling the far condition used in the present and previous work—or up to 35 syllables downstream (Falandays et al., 2020). However, unlike previous work on right context effects beyond word boundaries, Brown-Schmidt and colleagues analyzed right context effects during online language processing—specifically, eye-movements in a visual world experiment (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy, 1995). Following the logic of the McMurray et al. (2009) study described earlier, Brown-Schmidt and colleagues analyzed recovery time, defined as the amount of time between the onset of the disambiguating right context (e.g., fender in "The t/dent in the fender …") and the first subsequent time that the participant fixated the referent consistent with that right context.[10] The results suggest that recovery time was a function of both the acoustic continuum step and right context: the more consistent the original acoustics were with the interpretation indicated by the right context, the faster the recovery time. This suggests that listeners can maintain some gradient information (at least uncertainty about the intended referent of the pronoun he/she) about the speech input for much longer than previously thought, and that they can integrate this information with subsequent context during online language understanding. Given that these studies focused on fricatives, these findings are compatible with Szostak and Pitt's conjecture that listeners might be able to maintain information about fricatives longer than information about stops.

The three new experiments reported here used stop contrasts like those used by Connine et al., but with a procedure like Szostak and Pitt, which required participants to hear the relevant right context before responding. The results of all three experiments found clear evidence for effects of right context in the 6–8 syllable condition, and little to no evidence that the strength of the context effect decreased with distance (see also additional replications in Bushong & Jaeger, 2017, 2019b). These results replicate, for stop contrasts, the finding that Szostak and Pitt (2013) obtained for fricative contrasts. This suggests that the failure to find evidence of subcategorical information maintenance at far distances in Connine et al. (1991) might have been due to the task: if participants are allowed to respond before the end of the sentence, they are more likely to have responded before the critical context if that context occurs later.

This interpretation is supported by an additional experiment reported in Bushong and Jaeger (2017). That experiment was identical to our Experiment 2, except that participants were allowed to respond before the end of the sentence (as in Connine et al., 1992). With this small change in procedure, the patterns reported in Connine et al. (1991) replicate: the context effect is significantly smaller at the long (6–8 syllables) distance. This suggests that the two experimental tasks—letting participants respond whenever they want to vs. only after the end of the recording—measure different aspects of listeners' behavior, and they inform theories of speech perception in different ways.

Specifically, we submit that the task used by Connine and colleagues measures when participants feel sufficiently confident in their categorization decision to report it. This interpretation is supported by additional analyses reported in Bushong and Jaeger (2017), which found that participants are more likely to respond before hearing right context as the perceptually ambiguity of the VOT decreases (see Fig. 18). Critically, participants' decision to respond leaves open whether they maintain subcategorical information beyond that moment. Additionally, it is unclear whether response time decisions in a (monotonous) experiment are a good indicator of when listeners would categorize speech input during everyday speech perception, which differs in its task demands, contextual affordances, and incentives.

In contrast, the experimental task employed on the present study and by Szostak and Pitt (2013, Experiment 2) measures participants' in-principle capacity to maintain subcategorical information. Findings like ours thus show that listeners can in principle maintain subcategorical information for at least 6–9 syllables—including for stop consonants. Additionally, we found that most participants do seem to maintain subcategorical information: at least when participants have to wait until the end of a recording anyway, the clear majority of our participants exhibited effects of both VOT and subsequent context.

[10] This analysis included only trials in which the participant was not already fixating the correct referent at the onset of the disambiguating word but did fixate it within 5 s. This removed a substantial proportion of trials from analysis (between 62–93% in Brown-Schmidt and Toscano, 2017; 70-73% in Falandays et al., 2020).

We emphasize that this leaves open the extent to which maintenance of subcategorical information is typical during everyday speech perception. Both the paradigm used in the present study and that used by Connine and colleagues arguably differ from everyday speech perception in that listeners categorize the same word pair many times while knowing in advance that there are only two alternatives. Further, both paradigms share with most other works on this topic that the relevant phonetic information (a) always utilized the same phonetic contrast with word pairs and (b) always occurred in the same position within the sentential frame (for further discussion, see Burchill et al., 2018). This raises the questions as to whether results like ours are the consequence of participants developing experiment-specific strategies.

Some recent works have begun to address this question. One line of work has provided evidence that subcategorical information maintenance can be detected on the very first trial of experiments (Bushong & Jaeger, 2019a, 2024). Additionally, these studies found that the effects of subsequent context were highest at the beginning of the experiment and tended to decrease throughout the course of the experiment. This line of work thus seems to suggest that standard analyses of subcategorical information maintenance under-, rather than over-, estimate the true context effects (since analyses aggregate across trials).[11]

Other recent works have developed paradigms with improved ecological validity: for instance, by embedding the phonetic manipulations of interest in longer discourses, rather than short isolated sentences, and while reducing item repetition (Brown-Schmidt & Toscano, 2017; Falandays et al., 2020; described above); by masking the critical manipulation (Burchill, 2023; Caplan et al., 2021); or by avoiding the repetition of specific binary phonetic contrasts altogether (Burchill et al., 2018). The fact that most of these studies have found evidence of subcategorical information maintenance beyond word boundaries lends further credence to the idea that such maintenance also occurs during everyday speech perception (but see Caplan et al., 2021). We emphasize, however, that none of these paradigms perfectly approximates everyday speech perception, leaving an interesting space to be explored by future paradigm developments. For now, we conclude that listeners can in principle maintain subcategorical information for several syllables beyond word boundaries, that listeners seem to do so from the very first trial of experiments (at least for simple 2AFC tasks, as used in the paradigms we have focused on here), and that these observations continue to hold under existing attempts to increase ecological validity.

Is maintenance of subcategorical information limited to ambiguous inputs?

With respect to the first potential limitation mentioned above, we argued for the utility of using an ideal observer model as a null hypothesis of rational information integration. We derived the predictions of such a model and showed that: (a) the model predicts that subcategorical information from the preceding speech input and the right context should combine additively in log-odds space and (b) this prediction is qualitatively consistent with results of prior work that had previously been understood to support the first limitation. We then compared the predictions of the ideal observer and the ambiguity hypothesis against participant's responses across four experiments (including one re-analysis of previous work). We found that the evidence argues against both strong and weak versions of the ambiguity hypothesis. Although power at the continuum endpoints is (inevitably) very low, we found (and replicated) evidence that right context can have significant effects even at continuum endpoints. This result is incompatible with a strong version of the ambiguity hypothesis in which subcategorical information is maintained only for maximally ambiguous stimuli. Additionally, we found little evidence that right context effects decreased towards the continuum endpoints. To the contrary, the combined analyses for all four data sets—which would theoretically be well-suited to detect evidence for the ambiguity hypothesis and ameliorate issues with reduced statistical power at the continuum endpoints—consistently show a trend in the opposite direction. This is unexpected even under a weak ambiguity hypothesis according to which subcategorical information is less likely to be maintained for unambiguous stimuli (rather than never being maintained). Overall, then, our results reject the ambiguity hypothesis.

This conclusion is further supported by findings that have been published since we first conducted Experiment 1 ten years ago. Both Brown-Schmidt and Toscano (2017, p. 1223–4) and Falandays et al., (2020) present evidence that the effects of the previous speech input in their study were not limited to the most ambiguous inputs. One caveat to these findings is that it is less clear what predictions the ambiguity hypothesis makes for recovery times—the measure of primary interest in these previous studies. At first blush, however, the results of both Brown-Schmidt and Toscano (2017) and Falandays et al., (2020) would seem to provide further evidence against the ambiguity hypothesis.

What do our results convey about ideal information maintenance and integration during speech perception?

Between the ambiguity hypothesis and the ideal observer hypothesis, the four data sets we analyzed here favor the latter. This would suggest that integration of right context with preceding speech input proceeds optimally or near-optimally. However, the combined analyses also consistently found that right context were, on average, larger for perceptually less ambiguous continuum steps. The evidence for this trend ranged from "anecdotal" (Szostak and Pitt re-analysis) to "strong" (Experiment 3). Without further assumptions, this trend is not expected under either hypothesis we have entertained so far. What does this mean for subcategorical information maintenance?

One question that arises is whether we can reject the hypothesis that listeners cannot maintain and rationally integrate subcategorical information with subsequent context. To address this question, it would be necessary to directly compare the predictions of the ideal observer against alternative hypotheses that might explain larger effects of right context for perceptually less ambiguous tokens (unlike the weak or strong ambiguity hypothesis).[12] One potential approach is to compare the null predictions of the ideal observer—that the effect of right context interacts with neither the linear nor the quadratic effects of the phonetic continuum—against the alternative that those effects are not null. This is an approach we initially considered for the present study but then dismissed: an informative test of the nulls would require the specification of prior expectations about the alternative effect sizes; however, in its present form the ambiguity hypothesis is not sufficiently specific to deliver such quantitative predictions (for discussion, see also Bushong, 2020).

An alternative, more informative, test of the hypothesis that listeners maintain and integrate subcategorical information optimally for at least 6–8 syllables would require the specification of models that make specific quantitative predictions. In ongoing work, we have begun to

[11] Bushong and Jaeger (2019a) present evidence that this decrease is caused by the presence of conflicting cues—the fact that all VOT steps, including the continuum endpoints, occur equally often with both types of subsequent context. When these cue conflicts are removed, context effects remained stable across the experiment (see also Giovannone & Theodore, 2021).

[12] Note that the hypothesis tests presented in the combined analyses do not directly address this question. Rather, those tests compare the hypothesis that the effect of right context interacts with the quadratic effects of the continuum in the direction predicted by the weak ambiguity hypothesis against the alternative hypothesis that this interaction goes in the opposite direction (rather than the alternative of this interaction being null).

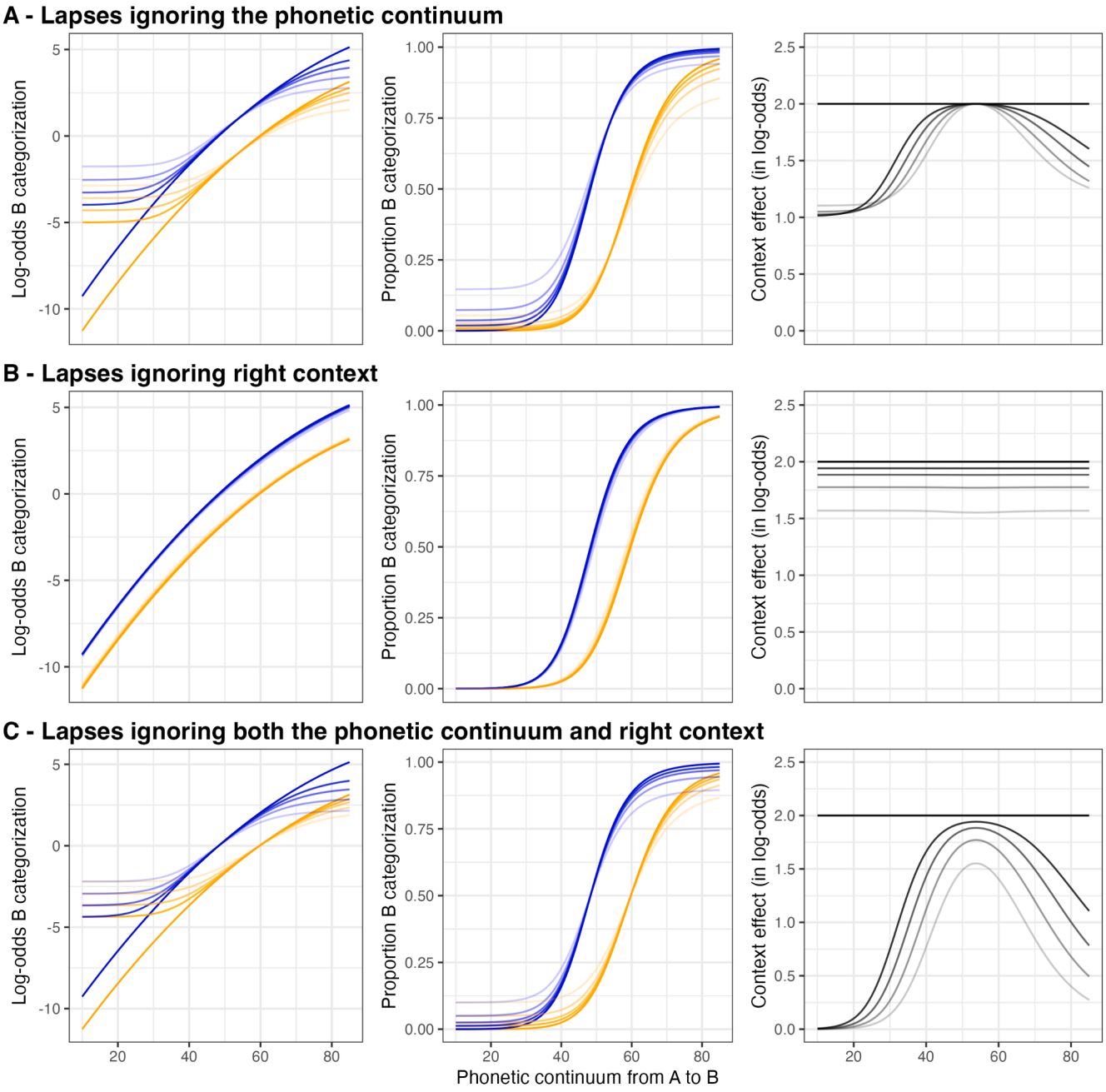

Fig. 19. Predictions of an extended ideal observer model that integrates attentional lapses on which listeners ignore A) the phonetic continuum, B) the right context or C) both. As in Fig. 2, we show predictions for a 2AFC task, in which tokens along the phonetic continuum are categorized into category A or B. **Left:** hypothetical effects of context (blue: B-biasing and orange: A-biasing) and phonetic continuum on log-odds of categorizing the stimulus as belonging to category B. Opaqueness of the lines indicates the percentage of trials on which acoustic information is ignored (from 0% for the solid lines to 2.5%, 5%, 10%, and 20% for the most transparent lines). The effects of context and phonetic continuum are modeled after Experiment 3 but with a larger effect of context to more clearly illustrate the predictions. **Middle:** The same but in proportion space. **Right:** Predicted context (differences between blue and orange line in left panel). The solid line (0% ignoring of acoustics) corresponds to the prediction of the naïve ideal observer tested in this paper—additivity in log-odds. Note that the right column shows a much smaller range of log-odds (y-axis) than the left column, in order to make the effects of context more visually apparent. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

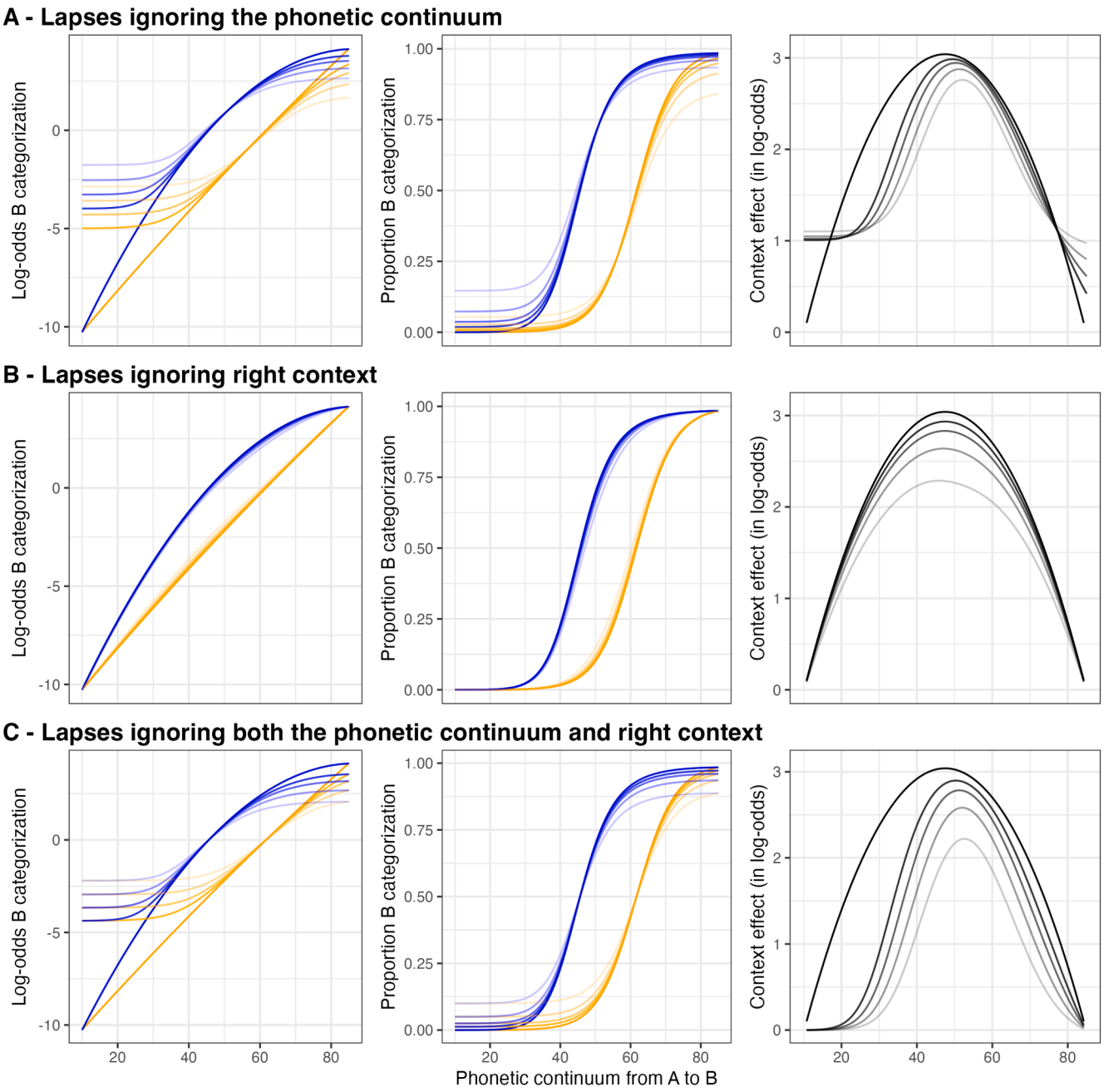

Fig. 20. Same as Fig. 19 but for the weak ambiguity hypothesis, as implemented in the present study. Unlike the ideal observer hypothesis, the ambiguity hypothesis predicts a qualitatively more similar pattern regardless of the type and rate of attentional lapses.

develop such predictive models (in the sense of Yarkoni & Westfall, 2017), integrating competing hypotheses about information maintenance into a general (offline) model of perceptual decision-making.[13] We have fit competing perceptual models—each implementing a different hypothesis about subcategorical information maintenance—to the data collected here as well as other data sets. These initial tests confirmed that the ambiguity hypothesis provides a bad fit against listeners' responses, and that a modified version of the ideal observer provides the best fit of all candidate models we have considered so far (for initial results, see Bushong & Jaeger, 2019c; Bushong, 2020). Specifically, our initial results would seem to suggest that listeners might be able to maintain and integrate subcategorical information optimally whenever they are not attentionally lapsing. We elaborate on this possibility.

Listeners have attentional lapses—trials on which they do not process parts or all of the sentence recording. Such lapses can reflect poor engagement in the experimental task, which is not necessarily reflective of the decisions listeners would make during everyday speech perception (see, e.g., evidence discussed above that experiments like ours underestimate subcategorical information maintenance). In some

[13] We consider the use of perceptual decision-making models more promising than the addition of higher-order polynomials (e.g., cubic polynomials) or other general curve-fitting methods (e.g., generalized additive mixed-effects models) to our combined analysis. Unlike these alternatives, models of perceptual decision-making are constrained by theory. This limits the types of interaction between right context and the preceding speech input that these models can account for, reducing the risk of over-fitting (for details, see Bushong, 2020). This is particularly important when analyzing non-linear effects over a small number of continuum steps (here: 5 to 6 steps).

situations, attentional lapses can even be a rational consequence of the task that listeners aim to optimize (which is not necessarily the same task as the task intended by the experimenter, cf. Pisupati et al., 2019). For example, participants might decide to ignore the early parts of the recording, for example, because they misunderstand the task of the experiment, or to strategically reduce task demands. Recall that the paradigm employed here and in previous work is highly repetitive. Participants repeatedly hear a small number of sentence frames. And with these sentence frames, participants also hear the same pairs of right context keywords (e.g., "in the fender" or "forest"). It is thus possible that participants start focusing just on the right context, and that participants vary in whether/how early in the experiment they do so (see also discussion in Burchill et al., 2018).

Whether intentionally or unintentionally, participants might thus respond A) solely based on the right context, B) solely based on the phonetic continuum (as they did, e.g., in much of the far condition of Connine et al., 1991), or C) while ignoring both sources of information. This raises the question whether such behavior would affect the predictions of the ambiguity hypothesis and/or the ideal observer. Fig. 19 illustrates the predictions of the ideal observer for each of these three 'lapsing' scenarios, and for different lapse rates (for reference, lapse rates < 10 % are not uncommon in speech perception experiments of similar length, Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; Tan & Jaeger, 2024). For example, when listeners occasionally ignore the phonetic continuum (Fig. 19A), the ideal observer predicts that the effects of context decrease towards the continuum endpoints—resembling predictions that one would expect from the weak ambiguity hypothesis! (The subtle asymmetry of the predicted decrease in the context effect is a consequence of the small quadratic effect of the phonetic continuum shown in the left panel of the same row.) This trend is even more pronounced when listeners occasionally ignore both the phonetic continuum *and* the right context, and respond by guessing. However, when listeners occasionally ignore right context while paying attention to the phonetic continuum, the ideal observer hypothesis makes rather different predictions (Panel B of Fig. 19): it predicts a small *positive* quadratic trend in the center of the continuum. This pattern is, however, predicted to be subtle even for large lapse rates of 20 % (see right panel of Fig. 19B). Fig. 20 shows the predictions of the ambiguity hypothesis under the same three lapsing scenarios. Unlike the ideal observer hypothesis, the ambiguity hypothesis *always* predicts a negative quadratic trend of the context effect, with decreasing effects of context towards the endpoints of the phonetic continuum.

A question for future research is thus whether an implementation of the ideal observer hypothesis that incorporates attentional lapses—in particular attentional lapses that ignore the right context (as in Fig. 19B)—provides a better fit to listeners' responses than the combined analyses we presented here. At first blush, this might seem unlikely given how subtle the quadratic trend in Fig. 19B is, compared to the quadratic trends we seem to have detected in our combined analyses. However, two considerations suggest that it would be premature to dismiss the possibility that an ideal observer with attentional lapses would fit participants' data well. First, the predictions shown in Fig. 19 are derived for specific parameter settings for the effects of context and the phonetic continuum, and the rate at which listeners (incidentally or intentionally) ignore right context. Consider, for example, the finding mentioned above, that right context effects tend to decrease throughout the course of the experiment when context frequently conflicts with the preceding phonetic cues (Bushong & Jaeger, 2019a, 2024). Such conflicts were present in all experiments analyzed here. If these conflicts lead participants to increasingly ignore right context, we might expect even larger rates of right context 'lapsing' than considered in Fig. 19B, and thus even more pronounced quadratic trends. Second, our combined analyses can only accommodate effects like those in Fig. 19B by inferring a positive quadratic trend across the entire phonetic continuum (which is what we observed). Combined with the inevitably high uncertainty about the effects of context at the continuum endpoints, this can mean that the combined analysis substantially *over*-estimates the effect at the continuum endpoints (as also pointed out by an anonymous reviewer).

In short, our analyses do not definitively answer whether listeners maintain and integrate subcategorical information rationally once incidental or intentional lapses are considered. They do, however, show that (1) ideal information maintenance and integration remains a viable candidate hypothesis if such 'right context lapsing' is taken into account, and (2) without this or alternative considerations, ideal information maintenance and integration is not compatible with our data. Readers interested in these questions are pointed to our ongoing efforts to implement the competing hypotheses in models of perceptual decision-making (the mathematical framework, model formulations, and initial results are presented in Bushong & Jaeger, 2019c; Bushong, 2020).

Directions for future work

One important question for future research is the grain at which listeners maintain subcategorical information about preceding speech input: do they maintain phonetic or even acoustic details, or do they 'merely' track their degree of uncertainty about the phonemic category or the word (e.g., $p(\text{tent} \mid \text{acoustics}) = 0.4$)? The results of the present analyses are equally consistent with maintaining uncertainty about phonetic features (e.g., voicing), phonemes, and words. Maintaining uncertainty about phonemes or words may tax sensory memory less, while still allowing optimal information integration during online language understanding (see the ideal observer derivations). There are, however, also reasons to believe that listeners might maintain more detailed subcategorical information despite the additional memory demands. For examples, listeners seem to be able to store phonetic or acoustic information about specific talkers and talker groups over extended periods (e.g., Eisner & McQueen, 2006; Goldinger 1996; Johnson et al., 1999; Liu & Jaeger, 2018; Walker & Hay, 2011; reviewed in Hay, 2018; Weatherholtz & Jaeger, 2016). Maintaining this information allows listeners to deal with inter-talker variability in the realization of phonological categories (for discussion, see Kleinschmidt & Jaeger, 2015) and facilitate inferences about talker's social identity (for review, see Foulkes & Hay, 2015; Kleinschmidt, Weatherholtz, & Jaeger, 2018). Notably, optimal *adaptation* to unfamiliar talkers—unlike optimal information integration during the comprehension of speech from familiar talkers—would require listeners to maintain more than uncertainty about preceding speech input (Burchill & Jaeger, 2024). Put differently, uncertainty maintenance is sufficient for optimal information integration as long as listeners already have an adequate model of the talker's usage of phonetic cues.

An important direction for future research is thus to integrate and, if necessary, reconcile findings from research on subcategorical information processing during processing and findings from long-term storage of subcategorical information (for relevant discussion and evidence, see Burchill et al., 2018; Falandays et al., 2020; Gwilliams et al., 2018). Recent studies have begun to test, for example, whether phonetic or richer information is available for sufficiently long time periods to facilitate adaptation to talker-specific pronunciations, with partially conflicting results (Burchill et al., 2018; Caplan, Hafri, & Trueswell, 2021; for a potential reconciliation of these findings, see Burchill & Jaeger, 2024). Data from brain imaging studies might provide one way to resolve this question. In a ground-breaking study, Gwilliams et al. (2018) used MEG to investigate neural responses within the auditory cortex to within-word effects of right context (i.e., the type of stimuli investigated behaviorally in McMurray et al., 2009). Gwilliams and colleagues find that subcategorical information—including phonetic information such as VOT and place of articulation—seems to be maintained in superior temporal regions throughout the duration of a word, and indeed repeatedly reactivated "even while subsequent phonemes are being received" (p. 7597). Future brain imaging studies could assess for how long this type of information is maintained (if at all) beyond

word boundaries.

Conclusions

Classic work by Connine et al. (1991) is widely cited as evidence that maintenance of subcategorical phonetic information is: (1) restricted to highly ambiguous segments close to a category boundary; and (2) only possible for a few syllables beyond the word boundary. We revisited these putative limitations using a combination of analyses of existing data and data from new experiments, power simulations, and Bayesian hypothesis tests. We conclude that the evidence is inconsistent with either hypothesized limitation. The data are, in fact, more consistent with a basic ideal observer model that maintains probabilistic evidence based on the subcategorical information and rationally integrates it with later information, e.g., semantic information that occurs later in the sentence. However, we also identify a consistent pattern in the data that is unexpected under any existing model, including the basic ideal observer model.

We now turn to the contributions we believe this work makes to the literature, with all the caveats discussed throughout this paper. Our first contribution is to formulate an ideal observer model for the influential paradigm pioneered by Connine et al. (1991). This ideal observer formalizes subcategorical information maintenance as a classic cue integration problem. Second, we show that this simple model is sufficient to explain the qualitative data pattern that had previously been taken to support the ambiguity hypothesis (reduced or no effects of right context for perceptually unambiguous continuum endpoints). Third, we derive that the two hypotheses make different predictions if analyzed in a more informative space (the log-odds of listeners' responses). Our fourth contribution is to test this stronger prediction against data from four different experiments. We present novel Bayesian analyses that quantify the evidence for each hypothesis both separately at each continuum step, and across continuum steps. Both approaches find no support for either of the two limitations hypothesized by Connine et al., favoring the ideal observer model instead. An additional contribution of our analyses is that they support the same conclusion regardless of whether a stop or a fricative contrast is used. This suggests that it is *not necessary* to postulate that listeners maintain subcategorical information differently for different phonetic contrasts (contrary to hypotheses advanced by Szostak and Pitt, 2013). Such differences might exist, but existing data does not provide evidence for them. The Bayesian analyses we present also quantify, for the first time, individual differences in the reliance on subcategorical information. This allowed us to validate an important assumption that is often made, but rarely tested, in research on subcategorical information maintenance: that the majority of participants in all four experiments are sensitive both to the phonetic information of the target word *and* information in the right context (see also Bushong & Jaeger, 2024).

Finally, we identify a quadratic trend in the data that is not predicted by a simple ideal observer, cue-integration model, and, thus, requires an additional mechanism. This raises the question about whether any model that can accommodate this new data pattern must abandon the assumption of rational information integration. Our final contribution is to show that incorporating attentional lapses into a model of rational information maintenance (but not the weak or strong forms of the ambiguity hypothesis) *might* be able to account for the data, thus identifying an important avenue for future research.

Author contributions: KB identified the issues affecting the interpretation of previous work and derived the predictions of the ideal observer model. KB and MKT designed Experiment 1. WB and TFJ designed Experiments 2 and 3. KB and WB created stimuli. KB conducted Experiment 1, WB conducted Experiments 2 and 3. All authors conceptualized and discussed analyses. TFJ conducted the Bayesian analyses and visualizations. WB developed the arguments against tests of the null. TFJ and WB conducted the power simulations. All authors wrote the paper.

CRediT authorship contribution statement

Klinton Bicknell: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Wednesday Bushong:** Writing – review & editing, Visualization, Resources, Methodology, Data curation, Conceptualization. **Michael K. Tanenhaus:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **T. Florian Jaeger:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data are shared on OSF.NIHG

Acknowledgment

This research was supported by an NIH post-doctoral fellowship to KB (T32-DC000035 to the Center for Language Sciences at the University of Rochester), NSF CAREER IIS-1150028 and NIH grant HD075797 to TFJ, and NIH grant HD073890 to MKT. The authors thank Meredith Brown and Bozena Pajak for technical assistance in stimuli creation, and Dave Kleinschmidt for technical assistance in preparing the web-based experiment. The authors thank Evan Hamaguchi for assistance in stimulus creation and recording. The authors are especially grateful to Christine Szostak and Mark Pitt for sharing their raw data with us. We also benefitted from feedback from Jim Magnuson, Bob McMurray, and especially Delphine Dahan. Experiment 2 was originally presented at the 2017 CUNY and CogSci conferences. Experiment 3 was originally presented at the 2018 CUNY and AMLaP conferences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2024.104565.

References

ten Bosch, L., Boves, L., & Ernestus, M. (2022). DIANA, a process-oriented model of human auditory word recognition. *Brain Sciences, 12*, 681.

Brown, M., Tanenhaus, M. K., & Dilley, L. A. (2021). Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science, 13*(1), 351–398.

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

Bürkner, P. (2017). Advanced Bayesian multilevel modeling with the R package brms. arXiv preprint arXiv:1705.11123.

Burchill, Z., & Jaeger, T. F. (2024). *Probing the nature of information listeners maintain about recent speech input: Initial results from a model-guided paradigm*. Manuscript: University of Rochester.

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PLoS One1, 13*(8), e0199358.

Bushong, W. (2020). *Maintenance of subcategorical information in spoken word recognition*. University of Rochester.

Bushong, W. & Jaeger, T. F. (2017). Maintenance of Perceptual Information in Speech Perception. In Gunzelmann, G., Howes, A., Tenbrink, T. & Davelaar, E. J. *(eds.) Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci), 186-191*. Austin, TX: Cognitive Science Society.

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

Bushong, W. & Jaeger, T. F. (2019b). Memory maintenance of gradient speech representations is mediated by their expected utility. In A.K. Goel, C.M. Seifert, & C.

Freksa (eds.) *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*, 1458-1463. Austin, TX: Cognitive Science Society.
Bushong, W. & Jaeger, T. F. (2024). Maintenance of subcategorical representations in spoken word recognition is modulated by recent experience. Submitted for review.
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.
Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., Riddell, A., et al. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 20*(2), 1–37.
Christiansen, M. H., & Chater, N. (2016). The Now-or-Never Bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences.*
Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.
Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language, 30*(2), 234–250.
Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6*, 84–107.
Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science, 19*(2), 121–126.
Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950–1953.
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*(May2019), 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.
Foulkes, P., & Hay, J. B. (2015). The emergence of sociophonetic structure. In B. MacWhinney, & W. O'Grady (Eds.), *Handbook of Language Emergence* (First Edit, pp. 292–313). JohnWiley & Sons, Inc.
Geisler, W. S. (2003). Ideal Observer Analysis. In L. M. Chalupa, & J. S. Werner (Eds.), *The Visual Neurosciences* (pp. 825–837). MIT Press.
Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine, 27*(15), 2865–2873.
Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360–1383.
Giovannone, N., & Theodore, R. M. (2021). Individual differences in the use of acoustic-phonetic versus lexical cues for speech perception. *Frontiers in Communication, 6*, Article 691225.
Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, memory, and cognition, 22*(5), 1166–1183.
Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(1), 69–78.
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
Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLoS One1, 14*(2), e0210793.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.
Jeffreys, J. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
Johnson, K., Strand, E.a., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*(4), 359–384.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.
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.
Kleinschmidt, D., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *TopiCS.* [10.1111/tops.12331].
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.
Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition, 174*, 55–70.
Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–523.
Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science, 189*, 226–228.
Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interations and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.
Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*, 298–421.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.
McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review, 15*(6), 1064–1071.
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.
Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395.
Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2021). Lapses in perceptual decisions reflect exploration. *eLife, 10*, e55490.
R Core Team (2019). R: A language and environment for statistical computing. *Foundation for Statistical Computing*, Vienna, Austria. URL 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.
Tan, M., & Jaeger, T. F. (2024). *Unravelling incremental adaptation to an unfamiliar talker*. Manuscript: University of Rochester.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). *Science, 268*(5217), 1631–1634.
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*(5), 779–804.
Walker, A., & Hay, J. (2011). Congruence between "word age" and "voice age" facilitates lexical access. *Laboratory Phonology, 2*(1), 219–237.
Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics.*
Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of non-native speech: a large-scale replication. *Journal of Experimental Psychology: General.*
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.