Supplementary Information for Bushong (2025) *Glossa Psycholinguistics*

1 Further Model Details

The models as presented in the main text leave some room for interpretation in how to parameterize them when fitting to behavioral data. In particular, it is likely that different participants will vary in their underlying subjective probability distributions for $p(t|VOT)$, $p(t|context)$, etc., and that these values also likely vary by sentence stimulus. Thus, for our models we include random effects for these parameters by participant and by sentence item. In this section, we present the parameterizations of each model as fitted in `R`. Our data, code, and fitted models can be found in our OSF repository at https://osf.io/rszfg/.

In this section, we also more fully explore the quantitative and qualitative descriptions of each model's predictions. Finally, we discuss plausible alternative formalizations of the ambiguity-dependent model.

1.1 Ideal integration

1.1.1 Predictions

Figure S1(b, g) shows the predictions of the ideal integration model under four plausible combinations of VOT and context effect sizes. Here and for the other models, when we refer to the strength of contextual evidence, we mean the subjective (to the listener) amount of information conveyed by the subsequent context. We define this as $log(p(t|context)) - log(p(d|context))$—i.e., the hypothetical effect of context in the absence of VOT information. Similarly, the VOT effect refers to the subjective amount of information conveyed by VOT in the absence of other information. The steeper the VOT slope (moving left to right in the facets of Figure S1(b)), the more categorical listeners' responses are (i.e., most VOTs are perceived as /t/ or /d/ close to 100% of the time, while only a small range of VOTs near the category are judged more ambiguously).

The most important takeaway of the ideal integration model is shown in Figure S1(g): the effect of context is entirely independent of VOT. This is a result of the additivity of the two cues in log-odds space: regardless of how ambiguous or unambiguous the VOT value is, context always influences categorization to the same degree. Because the ideal integration model predicts additive effects of VOT and context, the behavioral effects of VOT and context are equivalent to their underlying subjective probabilities (i.e., the parameters in the model).

1.1.2 Parameterization

The ideal integration model does not require the use of non-linear parameters since it predicts that VOT and context are independent and additive in log-odds space. Thus, we fit a mixed-effects logistic regression of the form:

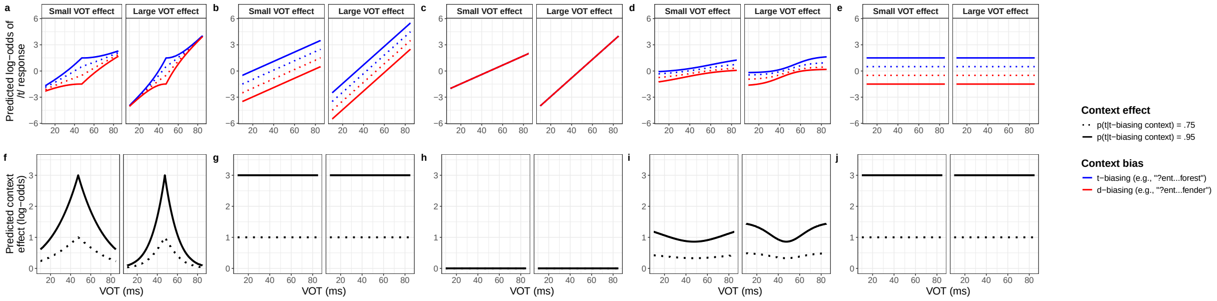

Figure S1: Illustration of the qualitative predictions of the (**a, f**) ambiguity-dependent, (**b, g**) ideal integration, (**c, h**) categorize-&-discard, (**d, i**) categorize-discard-&-switch, and (**e, j**) context-only models on the (**a-e**) log-odds of /t/-responses by context condition and VOT; and (**f-j**) log-odds of difference in /t/-responses between context conditions. Predictions are shown depending on the strength of the contextual evidence (line type) and effect of linear VOT (facets). We set the point of maximal ambiguity to the center of the displayed VOT range, and assume that the contextual evidence for either response (*tent* vs. *dent*) is symmetric around a neutral categorization function that would result in a neutral context (not shown). These choices make it easiest to see the influence of VOT and context on the predictions of the models. For the evaluation of the models, we do not make these assumptions.

$$\begin{aligned} \text{/t/ response} &\sim \text{VOT} + \text{VOT}^2 + \text{Context} + \\ &(\text{VOT} + \text{VOT}^2 + \text{Context}|\text{Participant}) + \\ &(\text{VOT} + \text{VOT}^2 + \text{Context}|\text{Sentence}) \end{aligned} \tag{1}$$

where effects of VOT and context are additive, and are allowed to vary between participants and sentence items.

1.2 Ambiguity-dependent

1.2.1 Predictions

The particular formalization of the ambiguity-dependent model presented in the main text was inspired by Connine et al.'s conception of how subsequent context may be used—namely, that context is not used at all for unambiguous stimuli and is used most for completely ambiguous stimuli (Connine et al., 1991). Thus, by design, the minimum context effect in our ambiguity-dependent model is 0 and occurs as the log-odds of a /t/ response based on VOT alone approaches $+/-\infty$ (0 and 1 in probability space). This results because the weight of the VOT-only component of the mixture model is 1 and the weight of the ideal-observer component is 0. The maximum context effect, then, occurs at the most ambiguous point, when the weights between these two components of the mixture model are equal. Given that the ideal observer model is the basis of the context effect at the most ambiguous point, the maximum context effect of the ambiguity-dependent model is equal to the (constant) context effect of the ideal integration model. Notice that the maximum and minimum of the context effect are independent of the slope of the VOT effect.

The slope of the context effect across the VOT spectrum is determined jointly by the contextual evidence and the VOT effect. The mixture weight α depends on $p(t|VOT)$. Specifically, the steeper the VOT slope, the larger the change in mixture weights between VOT steps (see difference between left-to-right panels in Figure S1(a)). Further, since the model always has a zero context effect at the theoretical endpoints, increases in contextual evidence always result in a steeper slope of the context effect along VOT. Finally, the slope of the VOT effect is predicted to be identical to the underlying subjective probabilities, as in the ideal integration model.

1.2.2 Plausible alternative formalizations

There are several plausible alternative ways to instantiate the ambiguity-dependent model than the formalization presented here. In the present work, α is a linear function of $p(/t/|VOT)$. Rather than treating it as linear, one could apply a softmax function, which would produce steeper changes in context use across the continuum. The most extreme version of this would be to describe a step function in context effect space where context is only used at the most ambiguous points and is never used anywhere else (i.e., $\alpha = 1$ for a defined ambiguous VOT range and $\alpha = 0$ everywhere else). Both of these alternatives are plausible interpretations of the fundamental ideas set forth by (Connine et al., 1991), and all three possibilities make broadly similar qualitative predictions.

In general, the class of models described above constitute a fairly strong version of the ambiguity-hypothesis, where subcategorical information maintenance approaches zero at acoustic-perceptual endpoints. However, ambiguity-dependent maintenance can take many forms, as long as the maximum context effect is at the most ambiguous point and decreases toward the endpoints. Here we assume that context does not affect categorizations at all at the endpoints, but a softer version of ambiguity-dependent maintenance may predict an attenuated, but not zero, effect. One way to instantiate such a model would be to fit α as a free parameter. We choose to evaluate a version of this model with a fixed α parameter as a starting point, because allowing α to be a free parameter would give this model the power to fit almost any behavioral data pattern—including patterns where context is actually smaller for more ambiguous VOTs, which we do not see as aligning with the spirit of the conceptual proposal of ambiguity-dependent maintenance (Connine et al., 1991).

1.2.3 Parameterization

The ambiguity-dependent model can be described as a mixture model where the mixture weights vary, and are determined, by $p(t|VOT)$ as described in the main text:

$$\begin{aligned} \text{/t/ response} \sim\ & \mathit{inv\_logit}(2|\mathit{inv\_logit}(v) - 0.5|) \times \mathit{inv\_logit}(v) + \\ & (1 - \mathit{inv\_logit}(2|\mathit{inv\_logit}(v) - 0.5|)) \times \mathit{inv\_logit}(v + c), \\ c \sim\ & 0 + \text{Context} + (0 + \text{Context}|\gamma_1|\text{Participant}) + \\ & (0 + \text{Context}|\gamma_2|\text{Sentence}), \\ v \sim\ & \text{VOT} + \text{VOT}^2 + (\text{VOT} + \text{VOT}^2|\gamma_1|\text{Participant}) + \\ & (\text{VOT} + \text{VOT}^2|\gamma_2|\text{Sentence}) \end{aligned} \tag{2}$$

Like the ideal integration model, we link this to responses using logistic regression (transformation not displayed here).[1][2]

[1] We take the log of the first line of Equation 2 and subtract from it the formula corresponding to predicted /d/-responses.

[2] γ represents a link between random effects groups.

1.3 Categorize-&-discard

1.3.1 Parameterization

Since the categorize-&-discard model predicts that listeners only use VOT in their responses, we can again employ a simple logistic regression like in the case of the ideal integration model:

$$\begin{aligned} \text{/t/ response} \sim \text{VOT} + \text{VOT}^2 + \\ (\text{VOT} + \text{VOT}^2|\text{Participant}) + (\text{VOT} + \text{VOT}^2|\text{Sentence}) \end{aligned} \tag{3}$$

1.4 Categorize-discard-&-switch

1.4.1 Predictions

As can be seen in Figure S1(d, i), the size of the context effect under the categorize-discard-&-switch model is driven jointly by the probability of /t/-responses based on the VOT and the contextual evidence. The maximum context effect is determined by the contextual evidence. As $p(t|VOT)$ approaches 0 or 1, the difference between the tent- and dent-biasing conditions reduces to half the difference in their log probabilities $\frac{1}{2}(log(p(t|context) - log(p(d|context))$. Also of note, the theoretical maximum is never in practice reached for the categorize-discard-&-switch model, since it occurs at $+/-\infty$ log-odds of /t/ responses (an inverse of the ambiguity model, where context effects never reach their theoretical minimum of zero). While the maximum context effect is unaffected by the size of the VOT effect, the slope of the context effect across the VOT spectrum *does* depend on it. The steeper the VOT slope, the more quickly the context effect reaches its maximum. Note that, unlike the ambiguity-dependent model, the minimum context effect is *not* zero (with the notable exception of $log(p(t|context)) = 0$, when the context effect is 0 across the VOT spectrum). The minimum is reached at $p(t|VOT) = 0.5$ and, derived from Equation 7 in the main text, is $2log\frac{0.5 + 0.5p(t|context)}{1 - 0.5p(t|context)}$ in log-odds space. Notice that this too affects the slope of the context effect in VOT space: as contextual evidence becomes weaker, the predicted minimum context effect is a higher percentage of the maximum context effect than for stronger contextual evidence, resulting in a shallower slope. Thus, like the ambiguity-dependent model, weaker VOT and context effects correspond to shallower context effect slopes across the VOT spectrum.

The most drastic difference in predictions between the prior three models and the categorize-discard-&-switch is in the predicted slope of the VOT effect in behavior. The categorize-discard-&-switch model predicts a much smaller average VOT effect than its underlying subjective probability—the more unambiguous a VOT is, the more trials where subjects make switching responses (e.g., strongly biased /d/ paired with /t/-biasing context) which dampens the effect of VOT.

1.4.2 Parameterization

Like the ambiguity-dependent model, the categorize-discard-&-switch model is also a mixture model whose weights are determined by the evidence provided by VOT:

$$
\begin{aligned}
\text{/t/ response} \sim\ & inv\_logit(v)+ \\
& (1 - inv\_logit(v)) \times inv\_logit(c), \\
c \sim 0 + \text{Context} + (0 + \text{Context}|\gamma_1|\text{Participant})+ & \\
(0 + \text{Context}|\gamma_2|\text{Sentence}), & \\
v \sim \text{VOT} + \text{VOT}^2 + (\text{VOT} + \text{VOT}^2|\gamma_1|\text{Participant})+ & \\
(\text{VOT} + \text{VOT}^2|\gamma_2|\text{Sentence}) &
\end{aligned} \tag{4}
$$

Like the ambiguity-dependent model above, we link this to responses using logistic regression (transformation not displayed here).[3]

1.5 Context-only

1.5.1 Parameterization

Since the context-only model predicts that listeners only use context in their responses, we can again employ a simple logistic regression like in the case of the ideal integration model:

$$
\begin{aligned}
\text{/t/ response} \sim \text{Context}+ & \\
(\text{Context}|\text{Participant}) + (\text{Context}|\text{Sentence}) &
\end{aligned} \tag{5}
$$

1.6 Model Fits to Experiments 1-4

Below we show the predictions of each model fitted to each of our behavioral experiments. The left panel of each plot shows predictions in proportion space (lines and shaded regions) with the empirical means and 95% confidence intervals (points). The center panel shows predictions in log-odds space. The right panel shows the predictions for the context effect across VOT space (i.e., the subtraction of the two lines in the center panel). In all plots, dashed lines and shaded regions are mean and 95% highest-density continuous interval (HDCI) of model predictions drawn from 1,000 random posterior samples.

2 VOT Norming Study for Experiments 3-4

As mentioned in the main text (Section 3.2), we created new stimuli for Experiments 3-4 (the properties of these stimuli are described further in Bushong & Jaeger, 2019). To ensure we chose appropriate VOT steps, we conducted a norming study on these new stimuli. 32 participants were recruited from Amazon Mechanical Turk and paid $1.00 for their participation. Participants listened to the target words in isolation and responded whether they heard the word "tent" or "dent". We presented VOTs from 10ms to 85ms in steps of 5ms for a total of 16 tested continuum steps. Participants heard each step five times for a total of 80 trials (trial order was fully randomized).

We fitted a mixed-effects logistic regression model using `R`'s `lme4` package (Bates et al., 2014), predicting /t/ responses from (scaled) VOT with random intercepts and slopes by subject. As predicted, we found a significant effect of VOT on /t/-responses ($\hat{\beta} = 14.66, z = 10.43, p < .001$).

Figure S6 shows the mean /t/ responses across the tested VOT continuum. Empirically, participants' responses ranged from .006-.97 (this was computed by first averaging /t/ responses by VOT within-subject,

[3] Again, we can do this simply by taking the log of the first line of Equation 4 and subtracting from it the same formula corresponding to predicted /d/-responses.

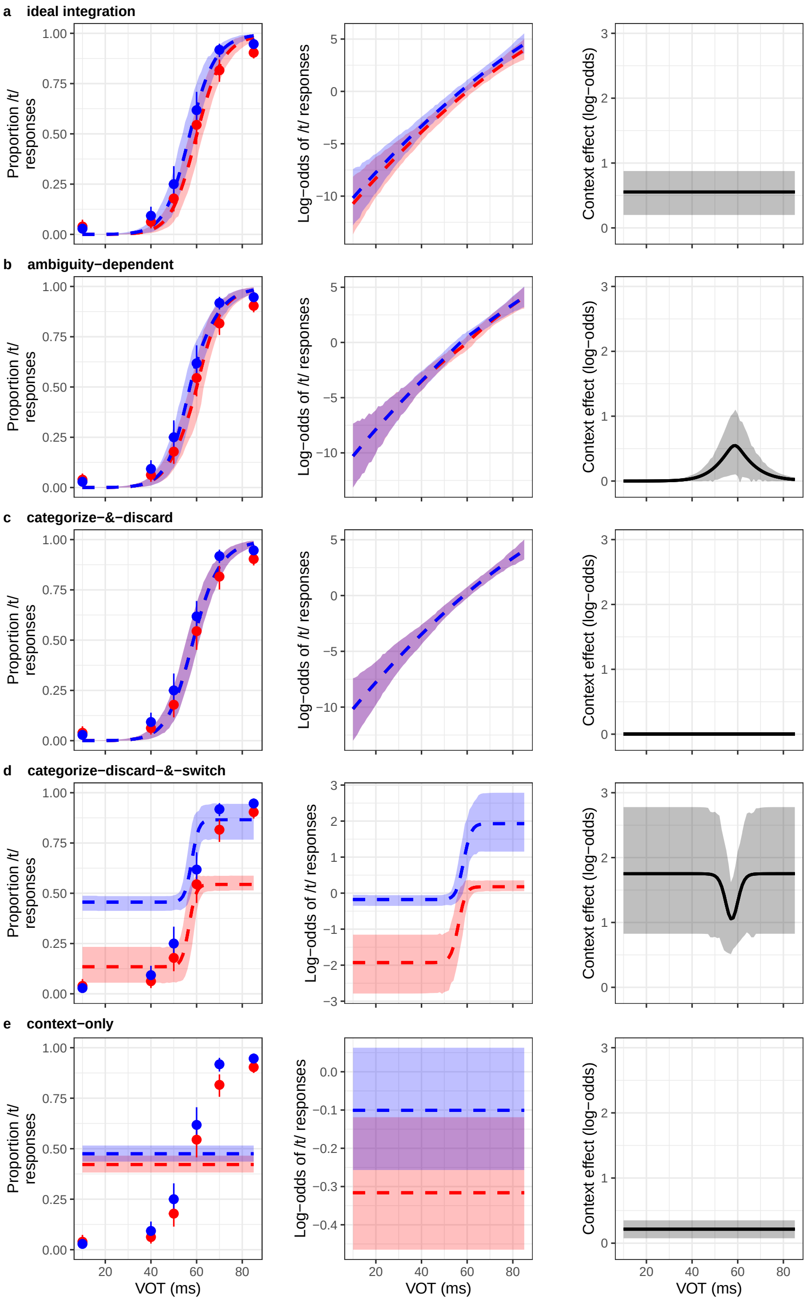

Figure S2: Model fits to Experiment 1.

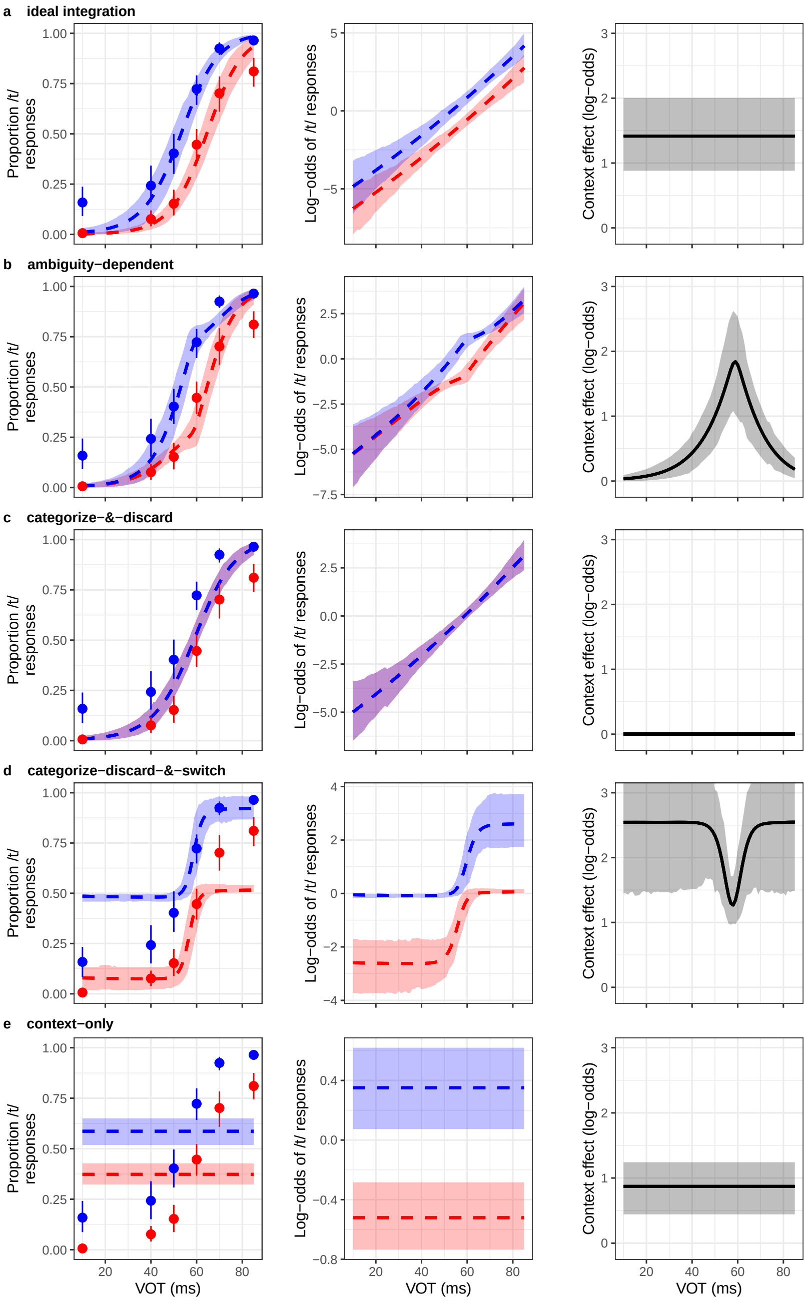

Figure S3: Model fits to Experiment 2.

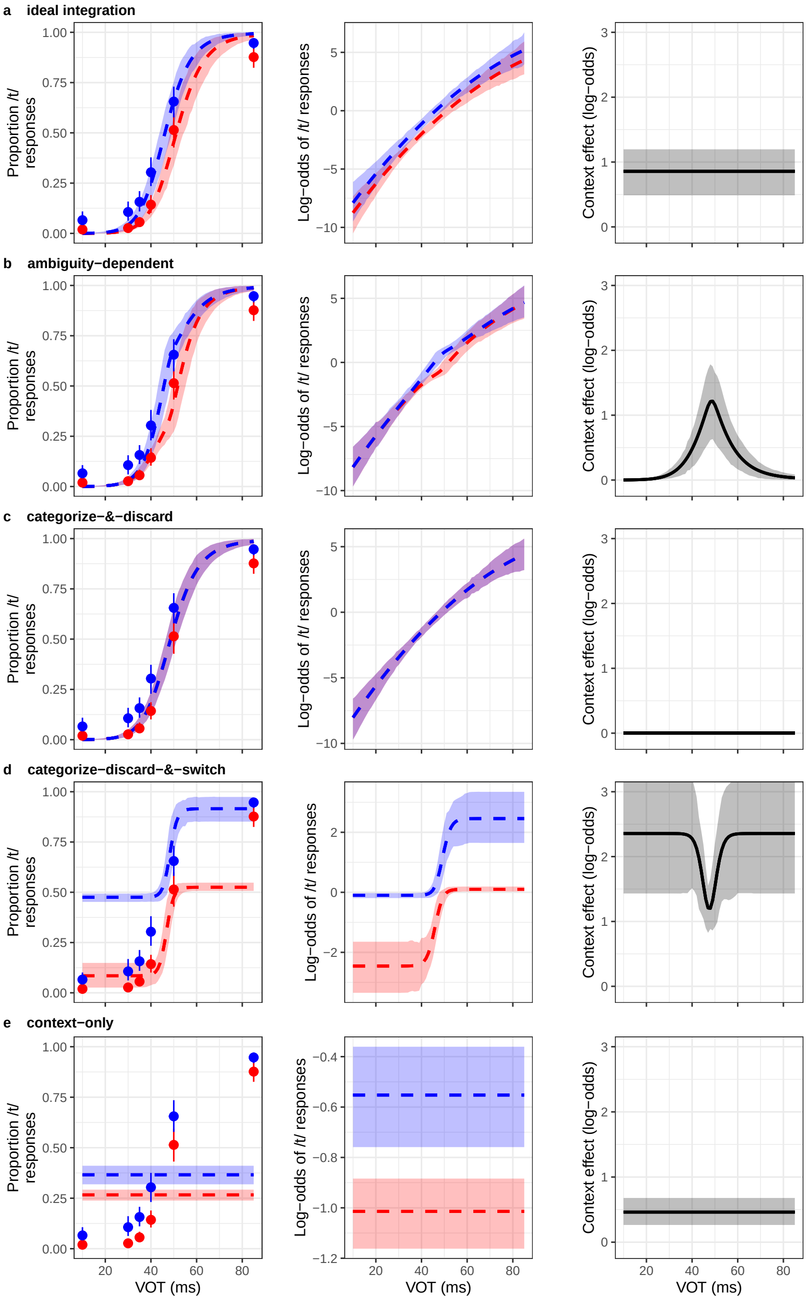

Figure S4: Model fits to Experiment 3.

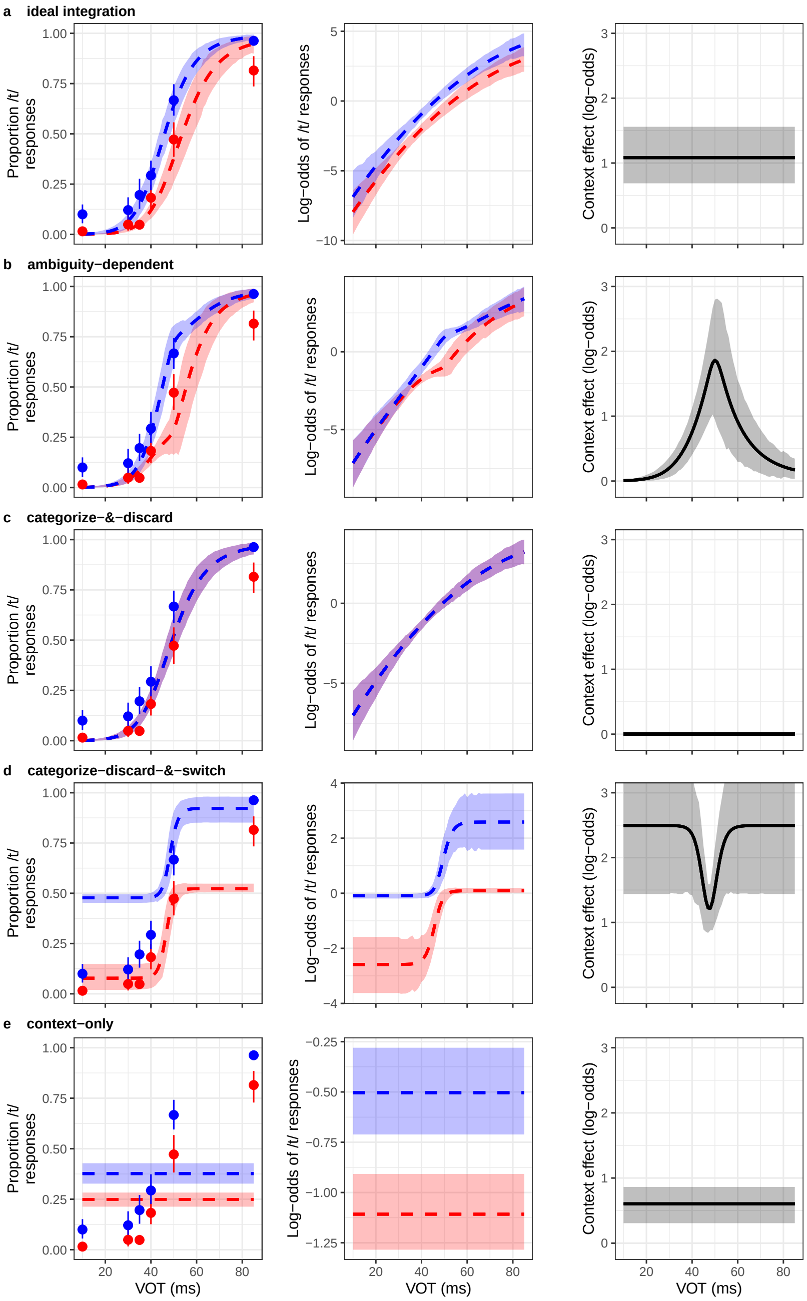

Figure S5: Model fits to Experiment 4.

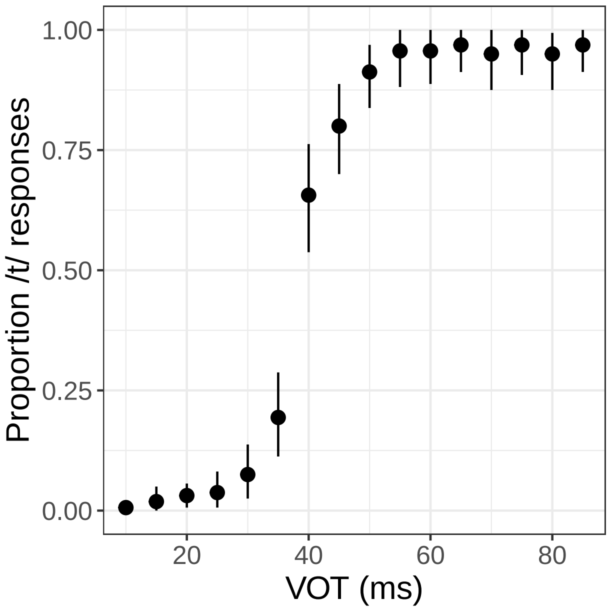

Figure S6: Mean proportion /t/ responses by VOT in the norming study for Experiments 3-4. Error bars 95% confidence intervals bootstrapped over subject means.

Experiment	**Min prop. /t/**	**Max prop. /t/**
Experiment 1	.034	.925
Experiment 2	.085	.888
Experiment 3	.042	.912
Experiment 4	.053	.879
Exp 3-4 Norming	**.006**	**.97**

Table 1: Empirically observed floor and ceiling proportion /t/ responses in Experiments 1-4. Estimates were derived by first averaging /t/ responses within subject and VOT; and then aggregating over subject. This creates a more realistic ceiling/floor estimate than directly aggregating over all subjects.

then averaging over VOT). The model-estimated category boundary (i.e., the VOT at which the logistic regression predicts 50% /t/ responses) was 38.9. These results drove our choice of 10, 30, 35, 40, 50, and 85 as the presented VOT steps in Experiments 3-4 since they represent the most and least ambiguous points, with some points of intermediate ambiguity.

One point of note is that the categorization curve obtained in this norming study is much steeper (more categorical) than observed in Experiments 3-4 (the empirical minimum and maximum proportion of /t/ responses for Experiments 1-4 is shown in Table 1). There are a few potential reasons for this. One possibility is that participants are more likely to have attentional lapses and miss the target word when listening to it in the context of a full sentence as compared to in isolation, which would result in noisier/shallower categorization curves. Secondly, the sentence contexts might also give other relevant acoustic details to the listener: for example, the (limited) presence of voicing contrasts on other words in the sentence stimuli might drive some limited learning and adaptation, producing slightly different categorization curves. Finally, it is possible that presenting a higher number of VOT steps affects participants' perception in unexpected ways.

Experiment	**VOT**	**Context**
Experiment 1	5.62 [3.69, 7.62]	3.01 [2.21, 3.83]
Experiment 2	2.72 [1.19, 4.33]	3.16 [2.17, 4.16]
Experiment 3	7.95 [6.17, 9.81]	2.38 [1.63, 3.13]
Experiment 4	6.03 [4.33, 7.84]	1.85 [1.04, 2.67]

Table 2: Estimates and 95% credible intervals of acoustic and contextual effects at the first trial of each experiment.

3 Trial analyses

One concern with the paradigm used in our experiments is that subjects may learn task-specific strategies over time, given the high degree of repetition present in our stimuli. One way to address this concern is to estimate listeners' behavior on the first trial of the experiment, before they have had any experience with the task and stimuli. If listeners show effects of both VOT and context from the first trial of the experiment, that suggests that they do not learn to use these cues exclusively through exposure to our task (though it does not rule out that there are some sorts of task-specific learning throughout the course of the experiment).

In order to assess effects on the first trial of the experiment, we fit Bayesian mixed-effects logistic regression models predicting /t/ responses as a function of VOT, context, as well as the interactions of each factor with trial number (logged); we also included random effects by subject and item for each of these factors (we excluded interactions as they resulted in convergence issues). We used the same priors as for the models described in the main text. With the inclusion of trial interacting with the main factors, the fixed effects estimates of VOT and context can then be interpreted as the magnitude of the effect on the first trial of the experiment. We classify an effect as significant if the estimate was in the correct direction (positive) and its 95% credible intervals do not include 0.

For each experiment, we found significant effects of VOT and context from the first trial. These results are summarized in Table 2.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). Lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7), 1–23. https://doi.org/10.32614/cran.package.lme4

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

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