Listeners Optimally Integrate Acoustic and Semantic Cues Across Time During Spoken Word Recognition

Wednesday Bushong (wednesday.bushong@wellesley.edu)
Department of Psychology and Program in Cognitive & Linguistic Sciences, Wellesley College

Daniel LaMarche (dlamarch@buffalo.edu)
Department of Psychology, University at Buffalo

Elayna A. Espinal (eespinal@hartford.edu) and Zachary J. Longo (longo@hartford.edu)
Department of Psychology, University of Hartford

Abstract

Understanding spoken words requires listeners to integrate large amounts of linguistic information over time. There has been considerable debate about how semantic context preceding or following a target word affects its recognition, with preceding semantic context often viewed as a constraint on possible future words, and following semantic context as a mechanism for disambiguating previous ambiguous input. Surprisingly, no studies have directly compared whether the timing of semantic context influences spoken word recognition. The current study manipulates the acoustic-perceptual features of a target word, a semantic cue elsewhere in the sentence biasing toward one interpretation, and the location of the semantic context. We find that the two cues are additively integrated in participants' word identification responses, and that semantic context affects categorization the same regardless of where it appears relative to the target word. This suggests that listeners can optimally integrate acoustic-perceptual and semantic information across time.

Keywords: speech perception, spoken word recognition, cue integration, semantic context, acoustic cues

Introduction

Spoken language is a fluid, continuous signal that unfolds over time. In order to understand spoken language, listeners must integrate different sources of information (cues) to infer the speaker's intended meaning. This is especially apparent in speech perception: when a listener is trying to identify a sound category, there are acoustic cues that arrive on the segment itself; however, there are many other sources of additional information that arrive both before and after (Liberman et al., 1967). These time-disjoint cues can include acoustic information (see Schertz & Clare, 2019, for review), but also semantic and contextual information. Consider the sentence "I don't mind ?eas, but I hate squash." If the listener's goal is to categorize the "?" sound, the later word *squash* is a cue that the speaker's intended word was *peas* (as opposed to an acoustically similar phonological neighbor like *bees*). Exactly how listeners integrate low-level acoustic cues with high-level semantic cues across time during spoken word recognition is not very well understood.

It has long been known that semantic information influences the processing of subsequent words in a sentence in spoken, written, and signed language (Altmann & Kamide, 1999; Kutas & Hillyard, 1980; Kutas et al., 1987, inter alia). Preceding semantic context has often been viewed as a mechanism for listeners to constrain the hypothesis space of future potential words. For example, sentences like "I like my coffee with cream and dog" produce considerable processing difficulty and large N400 effects in EEG studies (Kutas & Hillyard, 1980), suggesting that listeners use semantic cues to predict contextually appropriate words like *sugar*. To what degree preceding sentence context pre-activates semantic or lexical features of upcoming words, or functions as a mechanism for ruling out candidate words (like *dog*) entirely, has been a matter of debate (for review, see Van Petten et al., 1999). What has received comparatively less attention in the literature is how semantic information affects perception of the target word itself. In order to investigate this question, it is useful to examine situations where the target of recognition has an acoustically similar minimal pair. For example, Connine (1987) presented participants with sentences like "She wanted to wear the ?oat", where the "?" stimulus was acoustically manipulated to vary between /k/-/g/. The semantic information biases toward a *coat* interpretation, but the perceptual evidence varies between *coat* and *goat*. Connine (1987) found that preceding semantic context and acoustic-perceptual information are both used in word recognition, though semantic effects were particularly pronounced for more perceptually ambiguous target words. This suggests, contrary to proposals that the semantic context restricts the hypothesis space of future words (Altmann & Kamide, 1999), that semantic context may function as an additional piece of information when acoustic-perceptual cues are unclear.

A parallel line of work has investigated whether semantic context *following* a target word can influence its recognition. Unlike preceding semantic information, following context cannot act as a constraint on future words; rather, semantic information could function either as simply another cue which listeners integrate with earlier acoustic-perceptual information, or as a repair mechanism for disambiguating previous words. There is another notable difference with following semantic information: to successfully use it in spoken word recognition, listeners would need to maintain gradient subcategorical information about previously encountered input in memory to integrate with potential future cues like semantic context (Christiansen & Chater, 2016). A series of studies dating back to Connine et al. (1991) have shown that semantic context can in fact affect spoken word recognition even when it follows the target word (Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2017, 2019a; Connine et al., 1991; Falandays et al., 2020; Szostak & Pitt, 2013);

Semantics and timing	Semantic cue distance (syllables)	Semantic cue distance (words)	Non-target words with b/p onset
b-biasing/before	4 (1.46)	3.84 (1.27)	0.42 (0.56)
b-biasing/after	4.19 (1.83)	3.84 (1.37)	0.48 (0.57)
p-biasing/before	3.65 (1.4)	3.48 (1.21)	0.45 (0.57)
p-biasing/after	4.26 (1.67)	3.9 (1.08)	0.55 (0.62)

Table 1: Statistics of sentence stimuli. Mean in main text of cell, standard deviation in parentheses.

like with preceding context, some studies find that the semantic cue effect is more pronounced for perceptually ambiguous target words. Taken together, these results suggest that listeners are in principle capable of keeping track of acoustic and semantic cues over the course of a sentence and integrating them together to achieve successful word recognition.

A fundamental problem in the literature on the influence of semantic context on spoken word recognition is that there are few formal theories of its effects with clear quantitative tests. Proposals on the basis of empirical studies employing different methods have variously posited that semantic information acts as a constraint (to varying degrees) on the interpretation of future words, a mechanism for repairing misinterpretation of previous input, an influence only in the special case of ambiguous perceptual information, or a cue treated with the same status as acoustic-perceptual information. What is needed is a clear test of a formal theory of how semantic information affects spoken word recognition. Notably, many computational theories of speech perception and spoken word recognition make predictions that can be easily tested in simple perception experiments. Several theories, for example, treat semantic context as simply another cue available to listeners which is subject to similar cue integration processes as multiple acoustic-perceptual cues would be (Magnuson et al., 2020; Norris & McQueen, 2008). The purpose of the current work is to clarify the role of semantic context in spoken word recognition. One approach to this problem is to test listener behavior against an optimal baseline (like that provided by Bayesian models of speech perception, e.g., Norris & McQueen, 2008). If listeners integrate semantic and acoustic-perceptual cues in a statistically optimal fashion, this would suggest that humans can keep track of relevant cues in memory over the course of a sentence, integrating them together whenever a new piece of information arrives; semantic context is one piece of the puzzle, but does not play a special or privileged role in spoken word recognition. If, on the other hand, listeners employ a *non-optimal* strategy, this would suggest that the role of semantic information is more complex. It may function, as previously suggested, as a mechanism for constraining the hypothesis space of future words, or as a repair mechanism for interpreting past ambiguous input.

A simple way to test optimal cue integration is to manipulate the acoustics of a target word, a semantic cue elsewhere in the sentence, and the timing of the semantic cue (before vs. after the target word). If listeners optimally integrate these cues, then word categorizations should show additive effects of acoustic-perceptual and semantic information, and the timing of semantic cues should not make a difference.[1] Surprisingly, no study has tested these factors together in a single study; the present experiment manipulates all three of these factors to make an explicit comparison.

Methods

Participants

61 subjects were recruited via the FindingFive web-based experiment platform and were compensated $5.00 for their participation. All participants were native speakers of English currently living in the United States.

Materials

We developed a novel set of sentence stimuli varying in semantic context and its timing relative to an acoustically manipulated target word with an onset varying between /b/-/p/. This resulted in twenty-eight sentence quadruplets like the following:

1(a) I don't mind **[bees/peas]**, but I hate **squash** more than anything. (p-biasing, semantics-after)

1(b) I don't mind **squash**, but I hate **[bees/peas]** more than anything. (p-biasing, semantics-before)

1(c) I don't mind **[bees/peas]**, but I hate **wasps** more than anything. (b-biasing, semantics-after)

1(d) I don't mind **wasps**, but I hate **[bees/peas]** more than anything. (b-biasing, semantics-before)

Sentences were constructed to avoid two major confounds. First, we wanted to avoid differences in the distance of the semantic cue from the target between the semantics-before vs. semantics-after condition, both in number of syllables and number of words. Secondly, we wanted to minimize the number of other words in the sentence with [b/p] onsets, as this may influence the use of acoustic cues over the course of the experiment. There were no statistically significant differences between semantic cue or timing conditions on any of these three features (semantic cue distance in words, $ps > .35$, semantic cue distance in syllables, $ps > .47$, non-target words with b/p onset $ps > .66$). Table 1 shows the statistics of the stimuli. We hope these stimuli will be useful to other

[1] See following section Deriving Predictions from an Optimal Model.

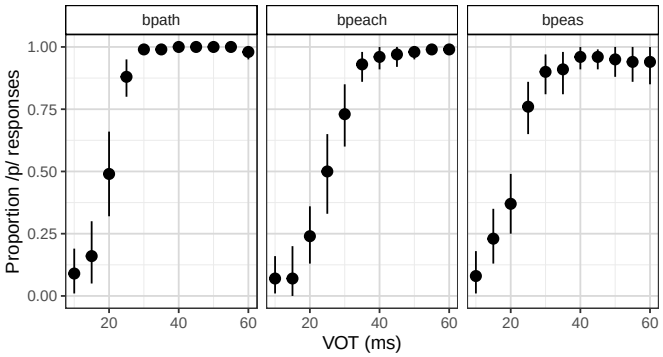

Figure 1: Categorization responses by VOT for each critical word (bath/path, beach/peach, bees/peas) in the norming study. Error bars represent 95% confidence intervals bootstrapped over subject means.

researchers studying the relationship between semantic and perceptual cues.[2]

We acoustically manipulated the voice-onset time (VOT) of the first sound of the target word to vary between /b/ and /p/. VOT is the main acoustic cue distinguishing voicing in American English stops; lower VOTs are more likely to be perceived as voiced (i.e., /b/), and higher VOTs as voiceless (i.e., /p/; Lisker & Abramson, 1970). We followed the VOT manipulation procedure developed by Winn (2020). To confirm successful acoustic manipulation, we conducted a norming study (described below).

We additionally created sixteen sentence quadruplets with critical words containing [l/r] distinctions (e.g., lake/rake). These critical words were not acoustically manipulated and were included in the perception experiment as filler trials. This was done as an effort to reduce repetitiveness in the experiment. On the whole, participants hear six different sets of target word pairs they need to make judgments about across the experiment.

Norming Study

To confirm that our VOT acoustic manipulation was successful, we conducted a norming study before the main experiment. 20 subjects were recruited via the FindingFive web-based experiment platform and were compensated $2.50 for their participation. All participants were native speakers of English currently living in the United States. Each participant heard each target word in 11 acoustically manipulated VOT steps (10-60ms in steps of 5ms), repeated five times each, for a total of 165 trials. Participants heard the word in isolation and were asked if the word started with /b/ or /p/. We fit a mixed-effects logistic regression model predicting the probability of /p/ responses from standardized VOT with random intercepts and slopes by subject and target word.[3] We found a significant effect of VOT ($\hat{\beta} = 10.81, z = 6.58, p < .001$). This suggests that as VOT increased, /p/-responses also increased, as expected. The model-estimated category boundary (i.e., 50% /p/ responses) was 22.08ms VOT, with some variation between target words (bath/path: 20.15, beach/peach: 24.74, bees/peas: 21.77). Figure 1 shows empirical categorization responses for each target word.

Procedure

Participants listened to sentences like (1a-d) above where the acoustics of a target word are manipulated to range between two possibilities (e.g., *bees-peas*), and a semantic cue either preceding or following the target biases toward one interpretation or the other (e.g., *wasps/squash*). All three factor were manipulated within participants. After each sentence, participants were prompted to identify a word in the sentence between two options; on critical trials, this was the target word with a [b/p] contrast (beach/peach, bees/peas, or bath/path), and on filler trials this was a word containing an [l/r] contrast (lake/rake, lock/rock, or lace/race).

We used a 7-step VOT continuum of 10, 20, 25, 30, 35, 40, and 50ms. We chose these VOT steps to range from unambiguous /b/ to unambiguous /p/ with intermediate steps in between, particularly concentrated around the category boundary of $\sim$22ms identified in the norming study described above.

A full crossing of all experimental factors with all twenty-eight critical items would result in an overly long experiment. Thus, participants were randomly assigned to one of seven experimental lists which allowed for more variety and less repetition within a single participant, while ensuring coverage of all conditions crossed with all items across participants. Each participant heard each unique sentence at only one VOT step; VOTs were then rotated according to a Latin square design to create seven experimental lists, so that each unique sentence item was paired with each VOT across lists. Participants also heard each filler sentence item in each of its four variants. This resulted in each participant hearing 176 total sentences (112 critical trials and 64 filler trials).

Deriving Predictions from an Optimal Model

Optimal models of spoken word recognition predict that listeners should use cues additively regardless of time. Generally, the statistically optimal solution to multiple cue integration, under the assumption that the perceiver's goal is accuracy, in any domain is to add all relevant sources of information weighted by their relative reliabilities (see, e.g., Ernst & Banks, 2002). The optimal solution to any cue integration problem where the goal is to infer some category c from sources of information $s_1 ... s_n$ is:

$$p(c|s_1, ..., s_n) \propto \prod_{i=1}^{n} p(c|s_i) p(s_i) \quad (1)$$

[2] All stimuli, data, and analyses can be found at https://osf.io/vbyag/.

[3] We standardized VOT by subtracting its mean and dividing by twice its standard deviation, a variant of z-scoring (Gelman, 2008).

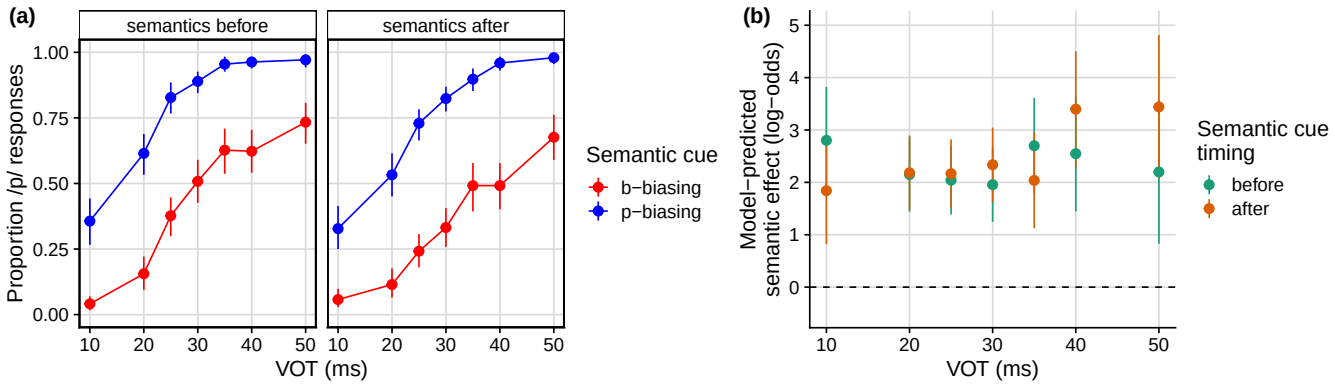

Figure 2: (a) Categorization responses by VOT, semantic cue, and semantic cue timing. Error bars are 95% confidence intervals bootstrapped over subject means. (b) Model-predicted semantic cue effect across the VOT spectrum by timing condition. Error bars are 95% confidence intervals derived from standard errors under assumption of normality.

In probability space, this describes a multiplicative relationship between cues. However, when we convert from probability space to log-odds space, these effects become additive.[4] Take, for example, a listener deciding between two categories c_1 and c_2 using two sources of information s_1 and s_2 (assuming that $p(s)$ is uniformly distributed):

$$\begin{aligned} log\frac{p(c=c_1|s_1,s_2)}{p(c=c_2|s_1,s_2)} &\propto log\frac{p(c_1|s_1)\times p(c_1|s_2)}{p(c_2|s_1)\times p(c_2|s_2)} \\ &\propto log\frac{p(c_1|s_1)}{p(c_2|s_2)} + log\frac{p(c_1|s_1)}{p(c_2|s_2)} \end{aligned} \tag{2}$$

There are a number of assumptions to acknowledge here. Firstly, additivity of cues holds only if cues are assumed to be conditionally independent; this is almost certainly not the case in natural speech (see Bushong & Jaeger, 2019a, for more on this point). Secondly, additivity predictions may not follow under a different objective function—for example, if a listener's goal is to categorize as *quickly* as possible rather than as *accurately* as possible.[5] Nevertheless, normative models like this provide a useful baseline against which to compare human behavior.

In the context of the present experiment, the relevant sources of information are acoustic and semantic evidence. The above framework of optimal cue integration would predict that listeners should show additive effects of VOT and semantic cues in their categorization responses in log-odds space (i.e., the space of logistic regression). Because cues are additive, the relative timing of the cues should not play any role in categorization responses.

Predictions and Analyses

Based on previous work, we would expect to see significant main effects of VOT and semantics on categorization responses (Bushong & Jaeger, 2019a; Connine, 1987; Connine et al., 1991, inter alia). Furthermore, these results should be reflected at the individual as well as group level: if some subjects use only VOT in their responses and others only semantic context, on average the results may appear as though listeners integrate both cues in their responses. Thus, a secondary goal of the current study is to estimate the effects of acoustic and semantic cue usage at the individual level.

The primary focus of the current study is the interaction between semantics, timing, and VOT. If listeners engage in optimal cue integration of semantics and acoustics, then the two cues should be used additively and timing should not matter. This would predict three major results. Firstly, we should find that semantic cues affect categorization responses regardless of VOT; that is, there should be a significant effect of semantics at each of the seven VOT steps tested. Second, this effect should be of similar magnitude across the VOT continuum—i.e., there should be no interaction between semantics and VOT. Finally, semantic cues should affect categorization responses regardless of timing; that is, there should be no significant interaction between semantics and timing.

To test these predictions, we conduct two types of logistic regression analyses in `R` (Bates et al., 2014; R Core Team, 2016). First, we fit a mixed-effects logistic regression predicting the likelihood of /p/-responses from standardized VOT, semantic cue (sum-coded; b-biasing = -.5, p-biasing = .5), timing (sum-coded; semantics-before = -.5, semantics-after = .5), and their interactions. This allows us to test for main effects of VOT and semantics, and interactions between semantics and VOT/timing. We included the maximal random-effects structure that resulted in successful convergence, which was random intercepts and slopes for scaled VOT, context, and timing by subject and item, with no interactions or random correlations. We extracted the by-subjects random effect estimates and standard errors from this model as a measure of the individual-level effects of VOT and semantic cue. Participants were classified as having a signifi-

[4] See also Norris and McQueen (2008) and Bicknell et al. (under review).

[5] Thanks to an anonymous reviewer for raising these points.

cant effect if the estimate was in the predicted direction and 95% confidence intervals derived from the standard errors did not include 0. These 'significance' estimates should not be taken at face value, as they are based on the limited data available from each participant and are influenced by group-level effects; rather, they give us a general idea of whether the majority of participants use both acoustic and semantic cues in categorization as expected, or if there is a mixture of participants who exclusively use one cue or the other.

In addition, we also conduct a simple effects analysis[6] predicting the likelihood of /p/-responses from semantic cue, timing, and their interaction, at each of the seven VOT levels tested. This allows us to estimate whether semantic cues affect categorizations across the VOT (acoustic-perceptual) spectrum, and whether there are any interactions between semantics and timing at specific VOT steps, which might not be picked up on in the overall interaction.

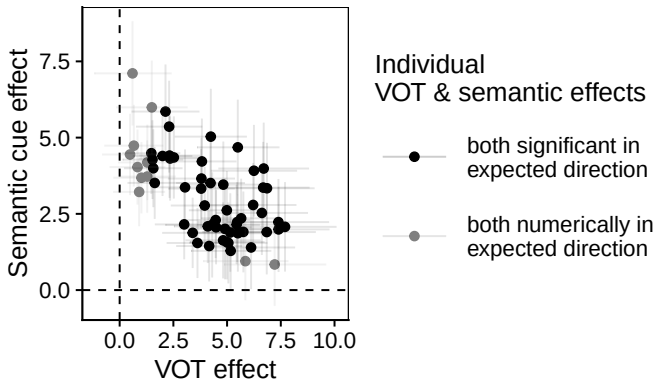

Figure 3: By-participant VOT and semantic cue effects, estimated from random effects. Error bars are 95% confidence intervals derived from standard errors under assumption of normality.

Results

Figure 2a shows the empirical categorization results by VOT, semantic cue, and semantic cue timing. As predicted, we found significant effects of VOT ($\hat{\beta} = 4.26, z = 12.66, p < .001$) and context ($\hat{\beta} = 3.22, z = 12.59, p < .001$). There was a main effect of timing ($\hat{\beta} = -.69, z = -4.63, p < .001$), such that participants were more likely to respond /p/ in the context-before condition, possibly reflecting a change in response bias. There was also an interaction between VOT and context ($\hat{\beta} = .69, z = 3.28, p = .001$), such that context influenced responses more toward the /p/ VOT endpoint than /b/. There was also a marginal interaction between VOT and timing ($\hat{\beta} = -.36, z = -1.89, p = .06$). Critically, there was no significant interaction between context and timing ($\hat{\beta} = .25, z = 1.41, p = .16$). Figure 3 shows the estimated VOT and semantic cue effects for individual participants. All participants exhibited effects in the numerically predicted direction, and the vast majority (50/61, or 82%) exhibited significant effects of both VOT and context.

Figure 2b shows the simple effects model-estimated context effect across the VOT continuum for each timing condition. In the simple effects analysis, semantic context was significant at every step on the VOT continuum ($\hat{\beta}s = [2.1 - 2.97], zs > 9, ps < .001$). There were no significant interactions between semantics and timing at any VOT step ($|z|s < 1.1, ps > .3$).

Discussion

We found significant effects of acoustic-perceptual cues (here, VOT) and semantic context on word categorization responses, replicating a large literature on the use of these cues in spoken word recognition (Bushong & Jaeger, 2019a; Connine, 1987; Connine et al., 1991, inter alia). Furthermore, we found that semantic context affected word categorizations at every step along the VOT continuum, not just the most perceptually ambiguous points. Critically, we found no evidence of an interaction between semantic context and timing, either overall or at any individual VOT step. Together, these results suggest that listeners optimally integrate semantic context and acoustic-perceptual information.

Two aspects of our results warrant further discussion. Firstly, it is striking that we find no evidence that semantic context only affects spoken word recognition for the most perceptually ambiguous stimuli, since this was a major finding of early work in this area (Connine, 1987; Connine et al., 1991) and is a generally accepted result (Dahan, 2010). There are two potential reasons for this discrepancy. Firstly, when analyzing binary categorization data, differences between conditions are difficult to detect when response proportions are close to 0 or 1, particularly when analyzed in proportion rather than log-odds space (Jaeger, 2008), which was the technique employed in the studies by Connine and colleagues. Secondly, in Connine et al. (1991), where semantic context appeared after the target word, participants were allowed to respond anytime after hearing the target word. This resulted in a substantial proportion of trials, particularly for perceptually unambiguous stimuli, where semantic context did not influence categorization because participants simply did not hear it. In many studies of the effect of subsequent semantic context on spoken word recognition, the perceptual ambiguity effect has not been replicated when participants must categorize after hearing the entire sentence (e.g., Bushong and Jaeger, 2019a, 2019b; see Bushong and Jaeger, 2017 for a more detailed ambiguity analysis).

Secondly, contrary to our expectation that semantic context and VOT should be used additively, we did find an interaction between these two variables. However, this appears to be pri-

[6] To fit the simple effects models, we use the nesting functionality for general linear models in R. This allows us to fit simple effects without needing to subset the data and re-fit models for each VOT level. The complexity of the model resulted in convergence issues when including random effects, so we conducted a simple logistic regression model for this analysis.

marily driven by the context effect being somewhat larger at the /p/-endpoint of the VOT spectrum than the /b/-endpoint, particularly in the semantics-after condition (see Figure 2b). Such a pattern would not be predicted if semantic context affects perceptually ambiguous stimuli more strongly, since we find here that semantic context is actually slightly larger for the *less* perceptually ambiguous /p/ endpoint of the VOT continuum. Similar patterns have been observed in other experiments where semantic context follows a target word (see, e.g., Bushong & Jaeger, 2019b). What exactly to make of this pattern is not clear. We need more well-specified theories of cue integration to explain this observation.

Overall, the results of the current study are most compatible with the view that low-level (here, acoustic-perceptual) and high-level (here, semantic) linguistic information is treated equally during real-time spoken word recognition. That is, semantic context is just another cue to word identity which is integrated with other cues encountered throughout an utterance. We find no evidence that semantic context preceding a target word functions to constrain the space of possible candidate words; instead, it is added with following acoustic-perceptual information. Similarly, context following a target word does not seem to be a mechanism for disambiguating previous ambiguous input, but rather is an additional cue integrated into listeners' representations of word candidates alongside the acoustic-perceptual evidence.

These findings provide support for theories of language processing which posit that speech perception and spoken word recognition are optimal processes, making use of all possible information across time (Norris & McQueen, 2008). This is consistent with proposals that semantic information does not provide a hard constraint on the interpretation of spoken words; rather, semantic information is matched with incoming acoustic-perceptual signals, which together provide evidence for a word candidates that best represent the sum of the information (Van Petten et al., 1999). This study provides the first direct evidence that the relative timing of acoustic-perceptual and semantic cues does not affect their integration, which would in principle be predicted by some recent models of spoken word recognition (in both Bayesian and connectionist frameworks; Magnuson et al., 2020; Norris & McQueen, 2008). Somewhat surprisingly, there are few formal models of spoken word recognition that explicitly take its temporal aspect into account.[7] Thus, the clear next step for future work is to formally model how spoken word recognition progresses across time with information provided by both low- and high-level linguistic cues.

Acknowledgments

This research was supported by internal grants at University of Hartford (Dean's Research Award to W.B.; Student Research & Creativity Award to D.L.).

[7] Shortlist B, for example, takes sentential context into account, but does not explicitly model how candidate word probabilities evolve incrementally over time (Norris & McQueen, 2008).

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

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.

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

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

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptual information in speech perception. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society*, 186–181.

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). Modeling long-distance cue integration in spoken word recognition. *Proceedings of the 2019 Workshop on Cognitive Modeling and Computational Linguistics*, 62–70.

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

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, *26*(5), 527–538.

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

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

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

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, 104088.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and clinical neurophysiology supplement*, *39*(1), 325–330.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

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

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

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

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

Schertz, J., & Clare, E. J. (2019). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1521.

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

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394.

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852–866.