Listeners store category identity and uncertainty in memory during spoken word recognition, but not acoustic detail

Wednesday Bushong (wb104@wellesley.edu)[1,2]
[1]Program in Cognitive & Linguistic Sciences, Wellesley College
[2]Department of Psychology, Wellesley College

Abstract

During spoken language processing, listeners maintain gradient information about past speech input in memory, but it is unclear how detailed these representations are. We test this question in an experiment using the speech discrimination paradigm. We focus on two perceptual decision-making effects: (1) more distant stimuli are more accurately discriminated; and (2) people are more confident about accurate choices, and this relationship is stronger as stimulus distance increases. We consider three types of stimulus 'distance' as a proxy for level of representational detail: (a) distance in acoustic cue space; (b) distance in category certainty space; and (c) difference in category identity. We assess which of these levels drive decision-making effects, finding that only category identity and category certainty are significant; difference in acoustic space was not a driver of perceptual decisions. This suggests that during real-time processing, listeners can maintain uncertainty about past speech, but cannot access its acoustic details.

Keywords: psycholinguistics; speech perception; perceptual decision-making; perceptual discrimination

Humans near-constantly make perceptual judgments: a pedestrian waiting to cross the street estimates the distance and speed of oncoming traffic; a pharmacist tries to decipher the names and doses of medications written in a doctor's messy handwriting; a bridesmaid assesses whether the color of her dress is a match for the couple's request; etc. The questions of how humans construct perceptual representations, and use them to make perceptual decisions, are thus of major interest to cognitive scientists.

One perceptual domain where these issues arise is in real-time spoken language processing. Speech is a continuous signal that listeners must transform into abstract linguistic representations (of sounds, words, grammatical structures, etc.) For instance, the distinction between the sound categories /b/ and /p/ is cued by many different continuous acoustic cues like voice onset time (VOT) and fundamental frequency (F0), among many others (Lisker & Abramson, 1970). Listeners have implicit probabilistic knowledge of how their language maps these acoustic cue values onto different categories, which they can use to make a judgment about which category the sound belongs to—i.e., how we consciously perceive the sound.[1] Figure 1 (top panel) shows a schematic of this process of transforming the signal into a perceptual representation of the stimulus, a distribution over likely categories (which throughout this paper we call *category certainty*), and finally an identified category.

This intrinsic pressure to categorize is reflected in early theories about speech perception which claimed that it is fully *categorical* in nature: after initial processing leading to the categorization, the listener *discards* all gradient information about the signal. Categorical perception was initially supported by evidence from AX speech discrimination tasks, where participants judge whether a stimulus X is identical to, or different from, a reference stimulus A. Classic studies found that when listeners judge two stimuli which are acoustically different but belong to the same category, performance is very poor; however, when two stimuli have the same magnitude of acoustic difference but *cross* a category boundary, performance spikes (Liberman et al., 1961; Liberman, 1957).

However, the strong claim for categorical perception has been chipped away over the years. Firstly, listeners do not always perform poorly at within-category discrimination (Gerrits & Schouten, 2004), and can improve with training (Carney et al., 1977); alternative tasks can also often reveal gradient within-category sensitivity (Apfelbaum et al., 2022; Barclay, 1972; Massaro & Cohen, 1983; Pisoni & Tash, 1974). Furthermore, the literature on the effect of subsequent context on spoken word recognition suggests that representations remain gradient and malleable for perceptually long stretches of time (Andruski et al., 1994; Bicknell et al., 2025; Bushong, 2025; Connine et al., 1991; Ganong, 1980; McMurray et al., 2002; for reviews, see Dahan, 2010; Muegge et al., 2025). Neuroimaging work has also suggested that neural representations encode gradient within-category information over longer timescales than traditionally thought (Gwilliams et al., 2018, 2025; Toscano et al., 2010). It is now widely accepted that speech perception is not entirely categorical (see McMurray, 2022, for a recent review).

What remains unclear is what level of representational detail listeners *do* have access to over time. Returning to Figure 1, do listeners have access to their perceptual encoding of acoustic features (panel (b)), or merely to a more abstract (but still gradient) probability distribution over likely categories (panel (c))? Attempts to answer this question, largely in the perceptual learning literature, have yielded conflicting results (see, e.g., Burchill, 2023; Burchill et al., 2018; Bushong et al.,

[1]Different theories of speech perception differ in the precise details of this process, but all agree that listeners follow these general stages (for a recent discussion of this issue and a fleshed-out processing model of speech perception, see Xie et al., 2023).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

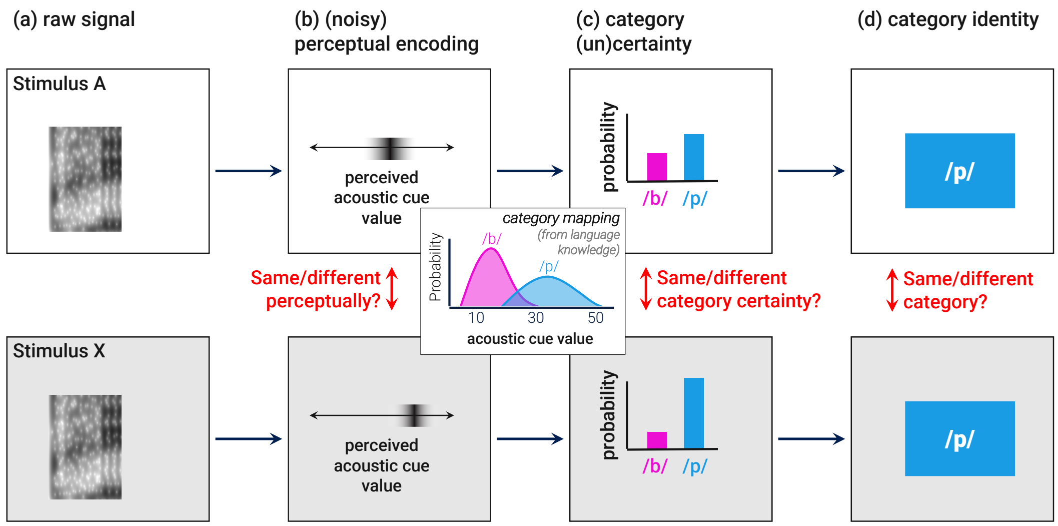

Figure 1: Levels of speech processing and potential comparison bases for perceptual decisions as tested in standard decision-making paradigms.

under review; Caplan et al., 2021).[2] Here, we propose that returning to classic perceptual decision-making tasks, like the AX discrimination tasks historically used in speech perception, can help us address this question. Interestingly, the field of speech perception has largely abandoned the speech discrimination task.[3] But AX (and other) discrimination tasks are alive and well in other perceptual domains, leading to rich theories of perceptual representation and decision-making that the speech perception literature can benefit from. Next, we review some basic findings in the non-language perceptual decision-making literature, and propose a link to speech perception.

Perceptual decision-making tasks generally ask participants to make a comparison between stimuli: for instance, the AX discrimination task described above. Experimenters manipulate various aspects of the stimuli in these tasks, most commonly stimulus distance (also referred to as stimulus discriminability), i.e., how near/far stimuli *A* and *X* are along some objectively measurable stimulus dimension. What is measured is choice accuracy and reaction time (which throughout this paper we will call *decision time*), and sometimes subjective confidence.

Through many decades of perceptual decision-making work, some clear empirical facts have emerged involving the relationship between stimulus distance, accuracy, and confidence: (1) the role of stimulus discriminability: the more distant the compared stimuli along the manipulated dimension, the more accurate the decisions; (2) the speed-accuracy tradeoff: longer decision times tend to result in more accurate choices (Garrett, 1922; Ratcliff & Rouder, 1998; J. Schouten & Bekker, 1967; Wickelgren, 1977); and (3) choice accuracy and confidence are correlated, and the correlation is stronger as stimulus discriminability increases (Baranski & Petrusic, 1994, 1998; Pleskac & Busemeyer, 2010; Sanders et al., 2016). These findings have inspired the most influential framework in perceptual decision-making, *evidence accumulation*: the perceiver aims to reach an evidence threshold for one choice or the other (i.e., 'same' or 'different' in an AX task), and they drift toward that bound on the basis of continued (perceptually noisy) evidence sampling, which accumulates over time until the threshold for a decision is met (and beyond—confidence effects have been interpreted as continued post-decision evidence accumulation; Pleskac & Busemeyer, 2010).

Taking evidence accumulation as a basic framework for perceptual decisions, we can now return to our question about what level of detail listeners maintain about past speech input. In particular, the critical question is *what evidence* listeners are accumulating during their decision process. In fully continuous domains, this is assumed to be a (perceptually noisy) encoding of the stimulus—e.g., returning to Figure 1, the comparison of stimuli would be at the perceptual level (panel (b)). However, given the memory limitations inherent in processing a transient acoustic signal, and the categorical structure of speech, it is possible that listeners instead maintain, and accumulate evidence over, a more abstract level of representation like category certainty or category identity. While speech discrimination experiments regularly assess whether differences at the acoustic-perceptual level or category identity level drive

[2] See also Xie et al. (2023) for arguments on why perceptual learning studies may not be able to definitively address this question.

[3] This is an exaggeration—however, it is true that speech discrimination experiments are very rarely used for the theoretical purpose of understanding the nature of listeners' speech representations (but see, e.g., Feldman et al., 2009; B. Schouten et al., 2003). And when these tasks are used, they are often *category* discrimination tasks in the vein of Pisoni and Tash (1974), which have no objective correct answers and are thus difficult to link back to theories from perceptual decision-making.

decision-making, no experiments have attempted to systematically test, on a participant-specific level, intermediate levels of representation like category certainty. The goal of the present study is to estimate to what degree each of the levels of representation drive listeners' decision-making.

The Present Study

The present study aims to test which level(s) of representational detail drive the broad perceptual decision-making effects described above. In particular, we consider three measures of stimulus distance. The first is the objective difference between stimuli in acoustic cue space, which serves as a proxy for listeners' perceptual encoding of the stimuli. The second is the difference in the log-odds of category probabilities between the stimuli, estimated individually for each participant based on a phonemic categorization task; this serves as a proxy for category certainty. Finally, we consider category contrast (whether two stimuli come from the same or different categories) again estimated from individual participants' categorization results; this serves as a proxy for category identity. We can then assess which of these levels drive characteristic decision-making findings while controlling for the other levels of representation.

Several outcomes are possible. One possibility is that listeners show characteristic decision-making effects only at the category identity level, which would suggest that listeners only have access to categorical information during processing, and cannot access 'lower-level' representations like perceptual encoding or category certainty. Conversely, if we see all three levels driving stimulus-discriminability effects, this would suggest listeners have access to all levels of information during processing.

Methods

All materials, data, and analyses are available at https://osf.io/3mdk9/overview.

Ethical approval

Experimental procedures were approved by the Brandeis University Human Research Protection Program (Protocol 26007R-E).

Participants

We recruited 120 participants using the Prolific platform, who then completed the experiment on Finding Five, a web-based experiment platform. We compensated participants $5.00, based on an expected completion time of 25 minutes at a $12/hr compensation rate. Using Prolific filters, all participants were native speakers of English who were born and currently living in the United States. All participants self-reported completing the experiment in a quiet space using earbuds or headphones.

We removed from analysis all participants with below-chance (50%) accuracy on the task, suggesting unusually poor discrimination ability or accidental reversal of choice options. This resulted in the removal of 10 participants.

Materials

We recorded a female native speaker of American English speaking full sentences (full-sentence stimuli were used in other work in our lab unrelated to the present project). For this experiment, stimuli were individual words extracted from the sentences; participants listened to a continuum of stimuli varying between *beach* and *peach*. We acoustically manipulated the voice-onset time (VOT) of the onset to vary between /b/ and /p/. VOT is one of the primary acoustic cues which distinguishes 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).[4] We confirmed successful VOT manipulation in a norming study reported elsewhere (Westwig et al., under review).

Procedure

Figure 2 shows the general procedure of the experiment, which proceeds in two phases. In the first phase, participants complete an AX discrimination task on speech stimuli. They hear two recordings, separated by 250ms, and are asked to judge whether the stimuli were the same or different. The instructions for the task explicitly instruct participants to focus on the acoustics of the sounds, ignoring category. Participants are instructed to respond that two sounds are different, even if they came from the same category, if they were "pronounced differently". Participants responded by pressing the 'M' or 'X' key on their keyboard to ensure quick reaction times (we balanced which key corresponded to which option across participants). Immediately after making a discrimination judgment, participants were asked to rate their confidence in their decision on a scale from 1 (least confident) to 5 (most confident).

We used VOT steps of 10, 20, 30, and 40ms. We chose these steps based on a norming study previously conducted in our lab (Westwig et al., under review), which found an average category boundary at 25ms; thus, we had stimuli that would fall approximately equally on either side of the category boundary for the average participant. Half of trials were identical, and the other half were different. For the different trials, we equally balanced each unique VOT pair, which results in an inherent imbalance in VOT-distance space; i.e., there are three possible VOT combinations resulting in a distance of 10ms, two possible combinations resulting in a difference of 20ms, and only one combination resulting in a difference of 30ms. We thus decided to value balancing unique combinations over balancing stimulus distance across the experiment. For each of the different pairs, half of trials presented the lower-VOT stimulus

[4] This procedure also includes modulations of fundamental frequency (F0) so that the edited result sounds more natural; thus, while we refer to our manipulation as targeting VOT, we are targeting voicing cues more generally.

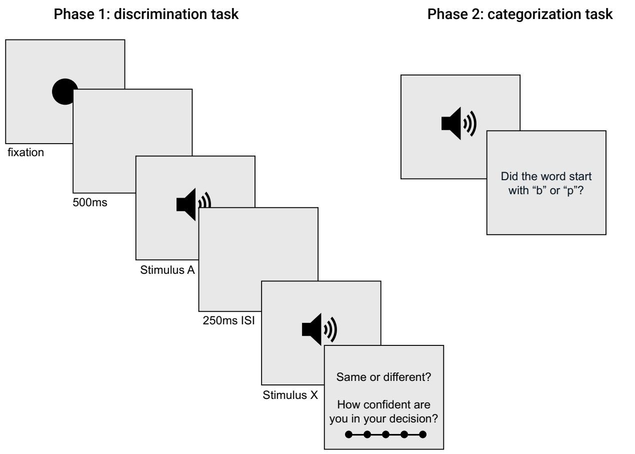

Figure 2: Trial structure of each experimental phase.

first, and half of trials presented the higher-VOT stimulus first. Trial order was fully randomized for each participant. There were a total of 96 trials per participant.

In the second phase of the experiment, participants complete a basic categorization task on the same stimuli as the discrimination task. In this task, participants hear one recording at a time and are asked if the word started with a “b” or a “p”. We sampled additional VOT steps (10, 15, 20, 25, 30, 35, and 40ms) so that we could achieve higher-resolution estimates of participants' category spaces. Each VOT step was repeated 10 times for a total of 70 trials in this phase.

Measuring participants' category structure

It is easy to measure the objective differences between stimuli in acoustic (i.e., VOT) space. But how do we measure participants' category certainty and whether a contrast is within- vs. between-category? We utilize the categorization data from the second phase of the experiment. Figure 3 shows an example of this process for one participant.

We first fit a mixed-effects logistic regression model predicting /p/ responses from VOT and squared VOT,[5] and fit random intercepts and slopes for each predictor by participant. Using the model, we then predict the log-odds of a /p/ response at each VOT continuum step for each participant. VOT contrasts are considered within-category if they shared the same sign (e.g., if 10ms VOT corresponded to -3 log-odds /p/ response and 20ms corresponded to -.5, they would both be assumed to belong to the /b/ category since they are lower than 50% /p/ responses), and between-category otherwise.

We can also compute the estimated difference of each VOT pair in the participants' category space. For instance, suppose that for some participant, the predicted log-odds /p/ response for 10, 20, and 30ms VOT is -3, -.5, and 1.5, respectively. A discrimination trial comparing 10ms vs. 20ms would be considered a distance of 2.5 in category certainty space, and a within-category contrast; and 20ms vs. 30ms would be a distance of 2 in category space and a between-category contrast. Generally, the distance in category certainty space will correlate with within- vs. between-category status, but not always—in the example above, the within-category contrast is more distant in certainty space than the between-category contrast.

Analyses and predictions

Prior to analysis, all trials with decision times or confidence rating reaction times under 200ms or over 5000ms were removed. This resulted in the removal of 280 trials (2.64%).

We fit a mixed-effects logistic regression with choice accuracy as the dependent variable, and VOT difference (z-scored), category certainty difference (z-scored), contrast type (sum-coded: -.5 = within-category, .5 = between-category), decision time (z-scored), and confidence rating (z-scored) as fixed effects. We also included in the fixed effects the two-way interactions between confidence rating and each of the stimulus distance variables (VOT difference, certainty difference, and category contrast). We included random intercepts and slopes for each variable by participant; we did not include any interactions between variables in the random effects, as they did not result in successful model convergence.

We predict that we will observe the characteristic results of perceptual decision-making discussed in the introduction:

1. The stimulus distance effect: more distant stimuli are more accurately discriminated.
2. The confidence-accuracy relationship: people are more subjectively confident about accurate choices; this relationship is stronger as stimulus distance increases.

Of critical interest here is at *which level* of stimulus distance—acoustic-perceptual (VOT), category certainty, or category identity—these effects surface.

we additionally assess whether there is an overall speed-accuracy tradeoff in our task: this is an effect that is consistently observed in decision-making tasks in non-categorical perceptual domains (Garrett, 1922; Ratcliff & Rouder, 1998; J. Schouten & Bekker, 1967; Wickelgren, 1977). Interestingly, the speed-accuracy tradeoff has not been assessed in speech discrimination experiments. Many experiments have manipulated interstimulus interval (ISI), finding (predictably) that accuracy decreases for longer ISIs (Crowder, 1971, 1982; Pisoni, 1973); however, we were unable to find instances of speech discrimination experiments in the literature which assess whether longer *decision times* are associated with higher or lower accuracy.

Results

We found significant main effects of stimulus distance at the category certainty ($\hat{\beta} = 1.873, z = 10.096, p < .001$) and category contrast levels ($\hat{\beta} = 1.975, z = 9.181, p < .001$),

[5]VOT distributions for /b/ and /p/ are of unequal variance, and thus we generally expect to observe quadratic effects in categorization space (see Lisker & Abramson, 1967).

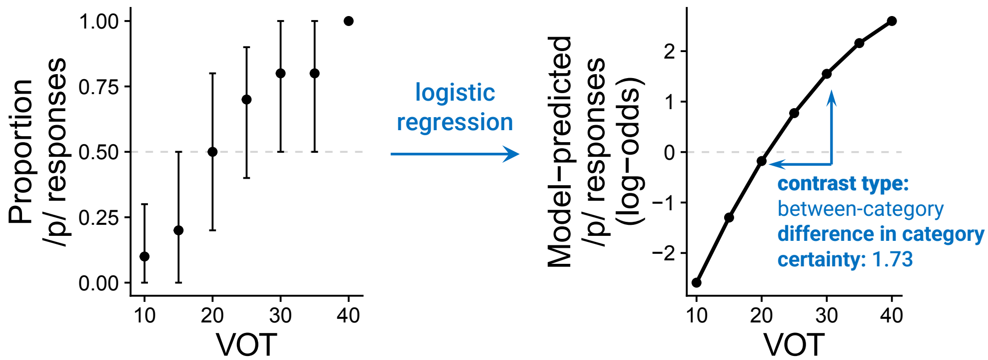

Figure 3: Left panel: example categorization curve from one participant. Right panel: calculation of contrast type and category certainty. A logistic regression model is fit to categorization responses; the VOT at which the model predicts 0 log-odds of /p/ response is deemed the category boundary. Difference in predicted log-odds of /p/ response is used as the difference in category certainty.

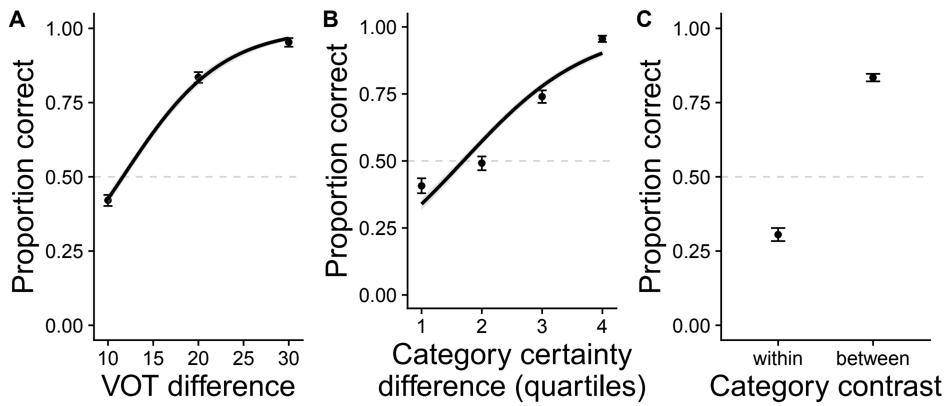

Figure 4: The empirical effect of stimulus distance (x-axis) on discrimination accuracy (y-axis). Panels show different ways of measuring stimulus distance (**A**: VOT difference, **B**: Difference in category certainty, **C**: difference in category identity). Note: for ease of presentation, difference in category certainty is rounded to its nearest quartile; however, it is treated as a continuous variable in the statistical models.

but not at the VOT difference level ($p > .7$, see also Figure 4). We also observed the expected main effect of confidence, such that confidence was higher for more accurate decisions ($\hat{\beta} = .281, z = 2.504, p = .012$). There were also significant two-way interactions between confidence and both category certainty ($\hat{\beta} = .566, z = 4.615, p < .001$) and category contrast ($\hat{\beta} = .514, z = 4.264, p < .001$), such that the confidence-accuracy relationship was stronger for more distant stimuli (see also Figure 5). There was no significant interaction between confidence and VOT difference ($p > .3$). Finally, we did not observe a speed-accuracy tradeoff ($p > .2$).

Discussion

This study builds on recent work in speech perception and spoken word recognition which suggests that listeners can maintain gradient information about past linguistic input in memory beyond the word boundary (Andruski et al., 1994; Bicknell et al., 2025; Bushong, 2025; Connine et al., 1991, inter alia). Past work has been unable, however, to definitively address how *detailed* this information is.[6] The goal of the present study was to characterize what level of representational detail listeners can use to make perceptual decisions. We found that, for major decision-making effects, category certainty and category identity drive listeners' responses, while acoustic-perceptual information does not. To our knowledge, this is the first study to link participant-level category certainty to speech discrimination. This is an important choice, since our key result only surfaces when we control for all factors simultaneously: acoustic-perceptual information did empirically drive listeners' decisions, but these effects disappear when accounting for category certainty and category identity. This highlights the importance of accounting for intermediate levels of representation in speech perception research.

[6] For further discussion of efforts in this vein in the perceptual learning literature, see Burchill et al. (2018), Caplan et al. (2021), Burchill (2023) Chapter 4, and Bushong et al. (under review).

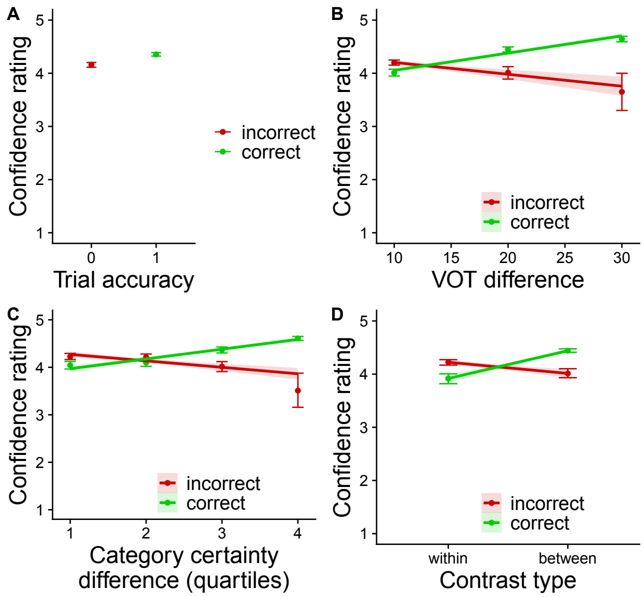

Figure 5: Confidence (y-axis) as a function of discrimination accuracy (color) in the empirical data. Panel A shows overall effect, and panels B-D show the confidence-accuracy relationship by our three major measures of stimulus distance. Note: category certainty difference is again rounded to its nearest quartile for ease of presentation.

Implications for psycholinguistic theories

As discussed in the introduction, categorical perception has fallen out of favor (McMurray, 2022). Notably, however, most theoretical alternatives that have arisen in its place assume, implicitly or explicitly, that listeners' representations of past speech are intimately shaped by category structure. In TRACE, for instance, gradient speech perception effects arise from lingering activation at the phonemic category level, not at the acoustic feature level (McClelland & Elman, 1986). Bayesian models like Shortlist B assume that gradient information about past speech is available indefinitely, but that information need not be more detailed than a probability distribution over possible categories; that is, listeners can compute $p(category|acoustics)$ once and maintain this value in memory, without having to store the precise value of all the acoustic features over time (Norris & McQueen, 2008). Feldman and colleagues' account of the perceptual magnet effect also explains effects in speech perception as arising from category structure, rather than a veridical encoding of the stimulus (Feldman et al., 2009). Caplan et al. (2021) make this assumption explicit in their activation over categories (AOC) formulation, which posits that listeners discard the details of the signal but can maintain gradient activation over discrete categories. Our present results are compatible with this view. As psycholinguists, it is important to ask ourselves what we mean when we use the term "perception"—true perceptual representations may decay incredibly quickly, despite the long-lasting gradient effects frequently observed in the literature.

However, there is an important caveat to keep in mind when interpreting these results. AX discrimination tasks are highly artificial and do not reflect the task demands of everyday spoken language processing.[7] It is possible that listeners are capable of maintaining more detailed representations of past speech that we are not able to capture in this paradigm. More generally, experiments utilizing isolated words (or nonwords) may underestimate the timescale of listeners' memory representations of past input. This is rarely discussed in the literature, but effects of subsequent context on spoken word recognition are typically much shorter-lived in priming (Andruski et al., 1994, observes a decay in effects after 250ms) compared to listening behavior in full sentences or discourses (Falandays et al., 2020, find effects up to 35 syllables after target word presentation). Thus, it is possible that the present task underestimates listeners' ability to maintain perceptual-level representations over time.

Links to perceptual decision-making

This is the first study to assess subjective confidence effects in a speech discrimination paradigm (though confidence has been investigated in the context of category discrimination tasks; see Schoenherr et al., 2012). We replicate the major findings of previous work in non-speech domains that confidence, accuracy, and stimulus distance are related (Pleskac & Busemeyer, 2010; Sanders et al., 2016); however, this is just a starting point and there are many more confidence effects in the perceptual decision-making literature that should be investigated in speech.

Finally, in contrast to non-categorical perceptual domains (Garrett, 1922), we observed no evidence of a speed-accuracy tradeoff. On the one hand, this may not seem surprising: most perceptual decision-making studies occur in domains where participants can continually sample evidence directly (e.g., visual discrimination tasks where stimuli are visible for the duration of participants' decision times). In speech-discrimination studies, by contrast, the signal is not continuously available to listeners. On the other hand, if category certainty-level representations are stable, one might *a priori* expect to observe a speed-accuracy tradeoff, particularly benefiting contrasts that are closer in category certainty space. Future studies should investigate the speed-accuracy tradeoff in this context in more detail.

Overall, this work represents a first step linking the speech perception and spoken word recognition literature to the broader perceptual decision-making literature. Immediate future work should aim to connect these results to models of evidence accumulation to further investigate where speech processing is similar to, or different from, perceptual decision-making in other domains.

[7] Thanks to an anonymous reviewer for raising this point.

References

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187. https://doi.org/10.1016/0010-0277(94)90042-6

Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, *152*(6), 3728–3745. https://doi.org/10.1121/10.0015201

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. https://doi.org/10.3758/bf03205299

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945. https://doi.org/10.1037/0096-1523.24.3.929

Barclay, J. R. (1972). Noncategorical perception of a voiced stop: A replication. *Perception & Psychophysics*, *11*(4), 269–273. https://doi.org/10.3758/bf03210376

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

Burchill, Z. (2023). *The reliability of standard reading time analyses and understanding the nature of maintained information in speech processing*. University of Rochester.

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

Bushong, W. (2025). Strong evidence for maintenance of gradient representations during language processing. *Glossa Psycholinguistics*, *4*(1), 1–29. https://doi.org/10.5070/G6011.40229

Bushong, W., Chang, A. J., Doherty, M., & Mercado, L. (under review). The role of sentence context in perceptual learning of speech.

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

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America*, *62*(4), 961–970. https://doi.org/10.1121/1.381590

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

Crowder, R. G. (1971). The sound of vowels and consonants in immediate memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(6), 587–596. https://doi.org/10.1016/s0022-5371(71)80063-4

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(2), 153–162. https://doi.org/10.1037//0278-7393.8.2.153

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

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

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

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

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, *56*, 1–105.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–376. https://doi.org/10.3758/BF03194885

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

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2025, March). Hierarchical dynamic coding coordinates speech comprehension in the human brain. https://doi.org/10.1101/2024.04.19.590280

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, *61*(5), 379–388. https://doi.org/10.1037/h0049038

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

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

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

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*(1), 15–35. https://doi.org/10.1016/0167-6393(83)90061-4

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

McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, *152*(6), 3819–3842. https://doi.org/10.1121/10.0016614

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

Muegge, J. B., Kim, H., & McMurray, B. (2025). Decoupling speech processing from time. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.05.017

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

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260. https://doi.org/10.3758/BF03214136

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290. https://doi.org/10.3758/BF03213946

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Schoenherr, J., Logan, J., & Winchester, A. (2012). Subjective confidence of acoustic and phonemic representations during speech perception. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 971–976.

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*(1), 71–80. https://doi.org/10.1016/S0167-6393(02)00094-8

Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153. https://doi.org/10.1016/0001-6918(67)90054-6

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

Westwig, A., Kim, Y., & Bushong, W. (under review). Listeners near-optimally integrate acoustic and semantic cues in spoken word recognition: Evidence from experiments manipulating cue order and reliability.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

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. https://doi.org/10.1121/10.0000692

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, *166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003