**Main points:**

1. **Length of manuscript.** We share the impression that that the manuscript is long, and that this will have a somewhat limiting effect on its impact. At the same time, like the reviewer, we don’t see any major ways to further cut the content. (We found one paragraph in the method part of Section 4 that seemed an unnecessary aside and we have removed it. We have also edited it through to minimize redundancies and to increase readability.) We had the manuscript read by about 20 people at this point. Our general impression is that anything that can be cut from the perspective of one reader, causes misunderstandings for another reader with a different background. We thus have not made any major cuts. Instead, we hope to give this manuscript its due impact through a series of follow-up studies (several of which are already in preparation) that demonstrate the advantages of the ASP framework.
2. **Removal of animations from journal article.** Both reviewers asked for this. Additionally we have set the animations to stop at the end of each loop, as suggested by R3.

**Overview of revisions**

The main message of the reviewers’ comments, as we understood, was two-fold. First, the

Finally, we made minor edits to improve accessibility of the text. Detailed responses to the remaining points of the reviewers are presented below.

**Responses to the remaining comments of Reviewer 1** (reviewers’ comments highlighted in blue)

We thank the reviewer for the encouragement and balanced critique. We focus here on R1’s two remaining comments.

- p. 8 clarification of "parsimony" at this point in the paper, the authors have claimed that both normalization and post-perceptual mechanisms are computationally more parsimonious than changes in representations. Does this not depend on the specific assumptions made for how each of the three systems is instantiated?

This is a great question. We have added a footnote on p. 7 that attempts to address it. Our claims about parsimony are about the ‘inherent’ complexities of the different mechanisms, and they should hold regardless of the specific assumptions made about the models, as long as one compares like with like.

For example, in the most general case, normalization requires the computation of a set of statistics (e.g., the mean, or the mean and standard deviation, or …) for at least each cue (e.g., VOT or f0). Changes in representations, on the other hand, require the same *but separately for each category* (e.g., the mean VOT for the /d/ category and the mean VOT for the /t/ category).

Of course, one could design a model of changes in category representations (e.g., only storing the mean of each cue for each category and doing so only over a moving time window of 5 seconds) that is less parsimonious than some other model of normalization (e.g., storing the first 100 statistical moments of each cue and doing so separately for each talker). But that is what we mean by comparing like with like. i.e., For each model of normalization, there is a parallel model of changes in representations that is less parsimonious.

Line 541: I am not sure that it is useful to send people to the end of the paper at Fig 35 at this point. I think the relevant details could be emphasized here, perhaps with a simple schematic, rather than pointing to this more complicated figure that can better be appreciated after reading more of the text.

Agreed. We have removed the reference to Figure 35. In the interest of brevity, we have not added further detail about the updating to the main text. We hope that the existing text and figures conveying the gist of the how this change model reacts to new input.

**Responses to the remaining comments of Reviewer 3**

[summary omitted] The work addresses an important need in the field, and I believe it will be a very influential paper when published. However, I have some major concerns about how the authors' modeling approaches as well as some suggestions to improve the presentation of the content, including a slightly expanded discussion of relevant work. I believe these concerns can be addressed in a revision. At the end of the review, I include a list of additional references that I mention in this review that were not included in the manuscript.

We thank the reviewer for engaging with our proposal and the constructive feedback. We are glad that the manuscript got another set of eyes that hadn’t seen the previous version. This clearly demonstrated a need to further clarify and revise some of our presentation.

Major Comments

1. The authors write that "the signature results from two influential lines of research—often taken to lend support to changes in category representations—are actually compatible with computationally more parsimonious change mechanisms (pre-linguistic signal normalization and changes in post-perceptual decision-making)" (p. 60, lines 1147-1151) and make several similar claims throughout the manuscript. However, I think this conclusion (at least as articulated) is premature.

For instance, can a normalization-based or post-perceptual mechanism account for the fact that some phonetic contrasts (e.g., /s/-/∫/) are learned in a talker-specific manner and others (/d/-/t/, the contrast focused on in this manuscript) are not? The finding that not all phonetic contrasts are learned in a talker-specific way (Kraljic & Samuel, 2007) is not addressed in the manuscript — but I think it's a critical point, and one that might even be problematic for a purely normalization-based account. (Is there a reason that normalization would apply to some cues and not others?) I'd love the authors to discuss this point, or at least acknowledge it as an avenue for future inquiry. [… additional point discussed separately below …] Without a clear account of how normalization or decision-level mechanisms would explain key findings in these literatures, I think it's premature to say that the data are consistent with other mechanisms. I'd be more precise in stating specifically that condition-specific changes in category boundary can be accounted for by any one of three mechanisms, and so normalization-based and decision-level accounts should be seriously considered by scientists in the field, but that additional work would be needed to assess these alternative mechanisms more clearly (and to see whether they can account for other key findings in these literatures).

This is a great point. However, we indeed intend our claims to hold at the strength that we stated them in the manuscript. There are, of course, many parts of the literature that we have not explicitly addressed (despite the length of the manuscript). But talker-specificity is a particularly nice example to demonstrate *why* we think our claims hold. The short of it is that there is nothing in existing theories that prevents normalization or changes in decision-making from being talker-specific. Quite to the contrary, in research on normalization, **talker-specificity is routinely taken to be a property of normalization** (e.g., Barreda, 2012, building on Magnusson & Nusbaum, 2007; or works on C-CuRE, which normalize by *talker*, assuming learning and storage of talker-specific marginal cue statistics for an unspecified amount of time—very much like talker-specific accounts of changes in representations). **We now clarify this as part of footnote XXX on p. 7.**

**We also hope that our general discussion makes clear that we do in fact think that there is existing evidence that could distinguish between the mechanisms.** But it’s far and few between, often not yet replicated, and most of these pieces of evidence (all discussed in SI XXX) arguably only imply that at least two of the mechanisms are required, not that a specific mechanisms definitely must be involved. For instance, Norris et al., 2003 rules out most simple normalization accounts as the sole explanation for their final experiments, but it does not necessarily rule out changes in decision-making in combination with normalization as an alternative to changes in representations.

We hope the reviewer doesn’t find it too obnoxious that we take this comment—at the risk of having misunderstood it—and use it to make a more general point about the field (without at all meaning to put anyone on the spot). We read the comment of the reviewer to assume that talker-specificity essentially implicates changes in representations as the mechanism. And, based in many conversations we had in the context of presenting our work, we think that the reviewer is not alone in this assumption. But *why?* We think that this is precisely because of the point we raise in the paper: separate lines of work on adaptive speech perception (incl. some of ours!) have gone on *for decades* without actually bothering to conduct *contrastive* (non-confirmatory) tests about the fundamental nature of the mechanisms we all study (we have collectively documented many *properties* of these mechanisms—such as talker-specificity—but almost all of these properties are compatible with all of the three mechanisms we discuss).

How could that happen and continue to happen? We think it is because of a tradition of confirmatory testing (perhaps too harsh a term: what we mean is testing that does not contrast the three mechanisms against each other). Specifically, we hypothesize that the causal chain of events that allows e.g., talker-specificity to become associated with changes in representations in one part of the field but with cue normalization in another part of the field works roughly like this:

* we conduct an experiment that we motivate through a particular mechanism. E.g., we might ask whether “phonetic learning” (presumably meant to refer to changes in representations) is talker-specific.
* we use a paradigm that we think—along with other researchers—taps into “phonetic learning” (e.g., perceptual recalibration or perhaps even a distributional learning paradigm).
* we find that—at least for the particular contrasts and/or cues studied in the experiment—adaptation seems to be talker-specific (e.g., Kraljic & Samuel’s finding for “s” vs. “sh”).
* we publish our paper, concluding that phonetic learning is talker-specific, or at least can be.
* others in our lab/field follow us, using the same framing, demonstrating further talker-specificity.
* New researchers entering the field read these papers, and over time we start seeing papers (e.g., recent reviews) that talk about talker-specificity as a property of phonetic learning.
* From there it is a small step to believing that talker-specificity is *evidence* for phonetic learning.

But, taking a step back, we realize that the results used to motivate talker-specificity were never shown (and in some cases, not even *argued*) to be due to phonetic learning. End of unsolicited reflection on the mechanics of the field.

[…] I'm also a bit puzzled as to why the authors focus on a /d/-/t/ example in this manuscript, given that the focus of the article is understanding how listeners condition phonetic identity on talker information and /d/-/t/ adaptation (at least in the lexically guided perceptual learning paradigm) seems to generalize across talkers.

[we cut this point out of the statement above to address it separately here]

Agreed. We had related conversations within our team. We opted to stick with this contrast because of how comparatively well its phonetics are understood, and because of the quality of available databases that provide information about the relevant phonetic distributions.

As an aside, we note that issues of talker-specificity are a bit more complicated than they are usually discussed (e.g., generalization across talkers does *not* necessarily argue against talker-specificity, e.g., if the speech rate of the two talkers is similar, as it was in Kraljic’s experiment). As discussed in Kleinschmidt & Jaeger (2015) talker-specificity and cross-talker generalization can be productively understood as inferences, rather than either being there or not.

2. I'm not sure I'm convinced by the authors' approach to modeling the post-perceptual decision-making stage of speech processing — I could have used a bit more explanation for why the authors formalize it in the way that they do. Consider, for instance, Section 2.2.3, where the authors appeal to Sohoglu and Davis (2016), writing that those authors "describe adaptation to degraded speech as changes in decision making" (p. 27, line 567); this work is characterized similarly elsewhere in the manuscript (p. 75, lines 1505-1512). However, I'm not sure if this is a fair characterization of Sohoglu and Davis's position — certainly, it seems inconsistent with how they've described this phenomenon elsewhere (e.g., Sohoglu et al., 2014; Davis & Sohoglu, 2020), where they've argued that this adaptation does \*not\* occur at a post-perceptual decision stage but instead involves a \*perceptual\* adjustment.

The **qualitative limitations that we identify for changes in decision-making should hold generally**, regardless of the specific implementation of our model**.** That’s because they mostly depend on the fact that any such model affects only the decision biases, rather than on the specific way that we *model* those changes. **We now clarify this on p. XXX.** Changes in decision-making largely can only lift/lower the overall categorization function (with some caveats once lapse rates are non-zero but even those changes are still very limited in nature).

We also note that our first submission used a different (less cognitive plausible) change model. That model also updated decision biases (as any change model for decision-making has to, by definition) but it did not employ any prediction error. The only effect this had is on the time course (across trials) of adaptation, not the types of changes that the model can explain.

Turning to Sohoglu and Davis (2016, SD16), we thank the reviewer for raising this point. We have re-read the SD16 and the other papers, and now realize that **we indeed misunderstood the account provided by Sohoglu & Davis**. **We have removed any mention of SD16 from the section on decision-making, and now discuss it in the preceding section on changes in representations.** We have also adjusted our general discussion. In case, it is of interest, we describe our revised understanding of the proposal made in SD16 below, and how it relates to our proposal.

The only description of an actual model that we found in this and the other two papers mentioned by the reviewer was in the SI of SD16. This description is sparse but the following quotes describe the *representations* assumed, and *computations* performed, by the model:

1. “Feature and phonological levels of representation were both modeled by assigning activation values to a set of units that represent a probability density function (PDF) as depicted in the bar graphs of Fig. 6A and Fig. S4).” 🡨 **This assumption matches those of Kleinschmidt & Jaeger (2011, 2012, 2015) and ASP’s categorization model.**
2. “In simulating perceptual learning, reductions in prediction error were attributed to changes in the variance or precision of predictions for sensory features.” 🡨 This describes a specific subset of events that can occur from perceptual learning model like that in **Kleinschmidt & Jaeger (2011, 2012, 2015) integrated in ASP as the change model for category representations. That is, the reviewer is correct: SD16 assume a model of changes in representations.**
3. “We therefore simulated perceptual learning by contrasting perceptual outcomes and prediction errors, during a pretraining period in which the distribution of sensory features was more precise than predicted, with a post-training period in which predictions were made with an increased precision that matched the sensory input (i.e., we used identical parameters for the SD of the category-to-feature weights and the sensory input in Table S1)” 🡨 **In short, SD16 does notpresent any actual learning/change model.** Instead, they contrast a state assumed to be the starting state of the model (researchers’ degrees of freedom) to the end state of having correctly acquired the actual precision of the degraded speech (leaving open whether/when an actual learning model would arrive at that precision). For a proof of concept, this makes sense, and we have employed the same approach, skipping the need for an actual learning model, in several of our own studies (e.g., Xie, Buxo-Lugo, & Kurumada, 2021; Tan, Xie, & Jaeger, 2021). We mention it here as context for the wording we chose to describe the SD model.

As such, I could have used a bit more explanation as to why prediction error is viewed as a signal to guide \*post-perceptual\* decisions (p. 28). Prediction error need not be conceptualized as relating to post-perceptual changes; indeed, phenomena like phonetic recalibration has often been described as a perceptual learning phenomenon (even dating back to the papers where the paradigm was introduced, such as Norris et al., 2003), which suggests a low-level perceptual locus rather than a post-perceptual locus (Goldstone, 1998). Furthermore, some work suggests that prediction error may be functionally equivalent to top-down feedback (i.e., may have a perceptual locus; Luthra et al., 2021, PB&R). What, then, leads the authors to describe this phenomenon in terms of changes to post-perceptual biases? If this is just one possible view of how such learning should occur, I'd encourage the authors to say so explicitly. For instance, the sentence "Participants can use this prediction error—operationalized here as the surprisal (|) of the category label given the acoustic input—to adapt the biases for all categories" (lines 582-584) could be preceded with a clause like "Under the view that talker adaptation reflects changes at a post-perceptual stage of speech processing."

**We now clarify that *all* of the three change models are sensitive to prediction errors (fn XXX on p. XXX).** Kleinschmidt & Jaeger (2015), for example, is a model of *perceptual learning* that is sensitive to prediction errors (see Jaeger et al., 2019 for discussion and demonstration of how Bayesian belief-updating without ever referring to prediction errors is actually sensitive to prediction errors). We have also removed the paragraph with links to the prediction error literature, in order to avoid that this aspect is seen as particular to this change model.

Prediction errors can—and probably do—exist at many levels of representations. And, in line with the reviewer’s comment, for prediction errors to ‘make sense’, they need to encode the information that would correspond to ‘top-down’—or to be more cautious ‘context’—effects. How exactly prediction errors are coded (e.g., whether the narrow interpretation of predictive coding a la Friston holds up to scrutiny; or whether it’s encoded ‘laterally’—i.e., inherent in the neural coding of e.g., a specific cortical layer—or vertically through top-down feedback, see Kuperberg & Jaeger, 2016) remains a topic of debate in computational neuroscience but that it is a *theoretical quantity* that is important in understanding processing and learning across the cognitive sciences is pretty uncontroversial.

In fact, it’s hard to conceive of any reasonable learning account that does not directly or indirectly refer to prediction errors (see also Jaeger & Snider, 2013 on discussion of the term “error-based learning” or Qian, Jaeger, & Aslin, 2012 on the role of prediction error in general theory of learning).

3. On a somewhat related note, it may be worth addressing (albeit briefly) some of the literature on whether there is potentially feedback from higher stages of processing to lower stages. As currently discussed in the manuscript, there appears to be an implicit assumption that context is integrated with phoneme-level information post-perceptually (e.g., p. 6, lines 165-167). But many prominent models of speech perception (e.g., the TRACE model) assume some degree of top-down feedback, a claim that has some support in the literature (Elman & McClelland, 1986; Magnuson et al., 2003; Luthra et al., 2021, Cognitive Science; but see, e.g., Norris et al., 2016). Footnote 2 (which alludes to the question of how higher-level information might be integrated into phonetic categorization) might be one place where it might be helpful to discuss this; what will be the important considerations to keep in mind as we move to thinking about phonetic categorization in context?

We thank the reviewer for raising this point. The original submission had a footnote that clarified that we do *not* assume discrete (non-cascading) feedforward models. That footnote was deleted as part of the revisions. But we **now clarify again that the three mechanisms should not be understood as discrete information-encapsulated processes** (footnote XXX, p. XXX). ASP is not a processing model. That is, we do not (yet) aim to capture the temporal dynamics for which modeling of feedforward and feedback information is critical.

Rather, we model the *outcome* of those processes. In ASP, decision-biases capture what *can* be integrated algorithmically (and neuronally) as top-down feedback during *processing*. During learning, each of the prior parameters (the kappas, nus, and betas) serves as top-down information.

4. I also have a few general suggestions regarding presentation. First, I wonder if it's worth showing just the end state of the simulations in the figures and then posting the animations online. The animations kept crashing my Adobe Reader, which was frustrating because what I was ultimately most interested in in is how the categorization functions differ in the end state. The order in which the distributions are sampled is random, so I think it's not helpful to show the intermediate states. If the authors decide to keep the animation, I'd suggest including a pause at the end state; currently, the animation loops immediately from end state to beginning state, meaning that the reader doesn't actually get a chance to see how it ends.

Thank you. This point has been addressed above.

Additionally, there are \*a lot\* of variables for the reader to keep track of throughout the manuscript. I'd strongly encourage the authors to provide a glossary in a Table, where one can quickly look up what, for instance, ,0 refers to. To further aid the reader, it might be useful to use short descriptors prior to the variable names when referring to them (e.g., in figure captions). For instance, in the caption for Figure 9 (p. 26), the authors could precede ,0 and ,0 with a brief descriptor such as "strength-of-belief parameters." By making similar changes throughout the manuscript, readers will be more easily able to follow along with the authors' approach.

We appreciate this comment. **As suggested by the reviewer, we now spell out the variable names in more places throughout the manuscript.** Doing so for each mention of the variable would further lengthen the manuscript. We also considered adding a table with all variables but we’re not quite sure how that would go beyond Figures 8, 10, and 13. Would a figure that consolidates all three of these figures into one perhaps help?

Minor Comments

p. 6, lines 178-179 — It might be unfair to equate Zheng and Samuel (2020) with perceptual retuning, given that those authors do describe how "criteria relaxation" differs from recalibration. (I also wonder if, given that criteria relaxation involves a change in a listener's \*decision\* about what qualifies as an acceptable exemplar of a category, why this example doesn't refer to the third stage of the hierarchy. As discussed in Major Comment 2, some clarification as to what specifically is meant by post-perceptual decision-level changes would help me here.)

**To avoid confusion, we have removed this part.** The issue with terms like “criterion relaxation” is precisely that they are vague and atheoretical (and yet commonly used!). At least some of us, initially interpreted Zheng & Samuel’s (2020) reference to “criteria relaxation” exactly as the reviewer seems to do (as referring to decision-making). However, when we had reached out to Arty Samuel about this paper, his description of the idea he had in mind seemed to be more akin to “widening of the category variance”, though he didn’t seem to commit to any particular view. Elsewhere, we have seen “criteria relaxation” also used to refer to increases in variance (Hitzcenko & Feldman, 2016).

p. 7, lines 211-216 — I'd recommend citing Magnuson and Nusbaum (2007) for an alternative conception of normalization — specifically because their view holds (a) that normalization is not automatic and (b) that talker information is not discarded. Pisoni (1997) also offers a useful perspective on normalization.

XXX

p. 8, line 222 — Maybe the question is whether normalization can \*fully\* explain talker-specific adaptation. As discussed elsewhere in the manuscript, there's some good work suggesting that while normalization helps, it only gets the listener so far, and it can be useful to condition category identity on other acoustic cues even after applying a normalization mechanism (Crinnion et al., 2020; Kleinschmidt, 2019).

XXX

p. 9, lines 250-253 — The Myers and Mesite example is a striking one here because an additional analysis with the same dataset (Luthra et al., 2020) shows how these adaptive changes are tied to the activity of relatively early (temporoparietal) brain regions, potentially suggesting multiple mechanisms underlying talker adaptation and not just decision-level ones.

Thank you for making us aware/reminding us of this work. **We now mention Luthra et al. (2020) as part of our discussion, which returns to this point (p. XXX).**

p. 10, lines 279-281 — The authors write, "The general conclusions and recommendations we arrive at in the present study are unlikely to be affected by these choices." However, this strikes me as a rather bold assertion to make without having tested a variety of implementations, especially for normalization and decision-level mechanisms — I'd omit it.

**Good point. We have removed it.** We actually have tested a variety of implementations (considerably more than reported, including slightly different versions of change models, etc.) but we agree there’s no need for such a bold statement.

p. 14 -16, lines 353-355 — There are several other prominent examples of how the same acoustic stimulus can be perceived differently from moment to moment. For instance, the authors might consider citing Billig et al. (2013), Leonard et al. (2016), and/or Schuerman et al. (2022).

XXX

p. 22-23, lines 469-472 — N appears in lowercase in this sentence and when it appears in subscripts (see Figure 8 / Equation 3) but is in uppercase otherwise. Is there a distinction to be made between the upper and lowercase forms of N/n? If so, what's the difference? (If not, please use just one case!)

XXX

p. 24 — I do not believe the term , which appears in Equation 4, is defined in the manuscript.

Unfortunately, the symbol did not transfer to the review. It would seem that Gothic D is the only symbol that is not mentioned in the text? That is the notation used to refer to “the data” in statistics. **As we now clarify, here it refers to previously experienced inputs.**

p. 28 — This section appeals to studies of adaptation to accented L2 speech (e.g., Xie et al., 2017) to explain why results might emerge through a change in bias for the labeled category. Given that a major point of Xie et al. (2017) is that adaptation to Mandarin-accented English involves "more than a boundary shift," it might be helpful to foreshadow here that changes in response biases can capture these effects if the lapse rate is greater than 0 (Figure 15).

p. 34 — While it's certainly the case that most phonetic recalibration studies have manipulated lexical bias between participants (e.g., whether participants hear ambiguous sounds in /s/- or /∫/-biased contexts), it's noteworthy that a number of studies have successfully manipulated this factor within participants (e.g., Saltzman & Myers, 2021; Heffner et al., 2022).

Thank you. This point seems orthogonal to our discussion. Given that the manuscript is already very long, we have opted not to mention these works.

p. 35, lines 693-697 — It might be more straightforward to list all the locations implicated and then provide the citations at the end — that is, something like "which range from primary auditory cortex and superior temporal cortices to more frontal and parietal areas (Bonte et al., 2017; Kilian-Hütten et al., 2011; Luthra et al., 2020; Myers & Mesite, 2014; Ullas, 2020; for review, see Guediche et al., 2014)." I suggest this because many of the studies referenced here don't simply implicate one set of regions (i.e., just frontoparietal or just temporal). Additionally, while it is true that the Killian-Hütten et al. paper referenced here did implicate temporal regions, those authors also published a paper that same year in Neuroimage, which used different analysis techniques and implicated frontoparietal cortex.

Great point. **Adopted.** Thank you!

p. 36, lines 723-726 — In introducing Figure 18, the authors write that "the conventional way of visualizing the results of perceptual recalibration experiments wrongly suggests…". I find this a bit misleading, though, since Figure 18 shows the characteristics of the stimuli; it does not visualize the results.

**We have removed this point, as it was an aside and one of the few place we felt we could cut without much loss of information.** We note that our point here is not that *we* are visualizing results differently (we’re not; we’re intentionally following the standard of the field when we present the results). Rather, we’re making a point about how thinking about PR studies in terms of a single continuum is misleading. Regardless of whether experimenters plot their results along a single cue dimension, listeners might use all available cues, and that can really change how one ought to interpret the result (e.g., in a separate paper we are finding that evidence that would appear as rejecting changes in decision-making if one falsely assumes a single cue dimension is actually *not* evidence against decision-making if one correctly recognizes that listeners draw on multiple cues.

p. 62, lines 1203-1204 — The authors write that "existing findings [suggest] that no single change mechanisms can explain the full variety of adaptive responses that humans exhibit." I don't think this has been shown, though; the current manuscript just argues that any one of three mechanisms can explain shifts in phonetic category boundaries. I'd encourage the authors to provide some references for this claim.

This sentence refers to an entire chapter in the SI, in which we summarize that evidence. The next sentence reads “*We summarize this evidence in more depth in the SI (§7)”.* We then continue with some brief examples from that section that constitute such evidence.

**We have slightly reworded this paragraph.** We would be grateful for feedback from the reviewer as to whether this is still ambiguous, and what we could do to clarify it.

p. 64, lines 1256-1257 — I'd add the excellent work of Guest and Martin (2021) to this list.

That does look like a really nice paper. Thank you (and added)!