April 20th, 2023

To: Editorial Board of *Cortex*

Dear Drs. Guediche and Caffarra,

We very much appreciate your and the reviewers' comments on our manuscript, CORTEX-D-21-00884 "What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review." The comments were extremely helpful as we finalized our manuscript for submission.

We are pleased to resubmit this manuscript for consideration for the special issue *Mapping sound to meaning under challenging conditions: converging findings and open questions across methods*. The manuscript is original, not previously published, and not under concurrent consideration elsewhere. As Dr. Guediche mentioned in the editorial letter, the remaining suggestions were mostly requests for elaboration and clarification, all of which we have now addressed.

Sincerely,

Shape

Description automatically generated with medium confidence



Xin Xie T. Florian Jaeger and Chigusa Kurumada

**Overview of revisions**

We have revised the manuscript in the following three ways.

First, R1 commented on the length of the manuscript. We agree that the manuscript is long and that this could limit its impact somewhat. At the same time, as R1 also mentioned, we did not see any major opportunities to substantially shorten the manuscript. We have by now had about 20 people read the manuscript at this point. Our general impression was that anything that could be cut from one reader's perspective would be misunderstood by another reader with a different background. So, we refrained from making any *major* cuts. We have, however, removed a long footnote in Section 2, and one paragraph in the Methods of Section 4 that were unnecessary tangents. Wherever possible, we also line-edited to further reduce redundancy.

Second, R3 requested several clarifications related to our modeling approach. We have addressed them, as detailed below. In doing so, we also hope to have strengthened the connection between our approach and the broader literature on speech perception and adaptation.

Finally, we have made additional efforts to improve the readability and accessibility of the manuscript. For example, Section 4.2 now includes a new figure of a graphical model summarizing all three change models and their parameters. This will serve as a glossary and a convenient reference point readers can consult as they read through the two case studies. (We thank R3 for the suggestion to include such a summary figure.) We have also removed the animations from the journal article, as requested by both reviewers. The relevant figures now include still images showing the end states of the changes as well as links to videos showing the animations we originally intended to present.

**Responses to the remaining comments of Reviewer 1** (reviewers’ comments 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 Footnote 1 on p. 7 that anticipate further clarification we added to a new Section 2.3 (p. 32-33). We further elaborate on the notion of parsimony in the SI (7.1).

In short, our arguments 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, generally, normalization requires the computation of a set of statistics (e.g., the mean, or the mean and standard deviation) for **each** **cue** such as VOT or f0, which can apply to multiple categories and contrasts. Changes in representations, on the other hand, require the same set of statistics **but separately for each category**(e.g., the mean VOT for the /d/ category and the mean VOT for the /t/ category). Finally, changes of decision-biases require **no such statistics** and track decision outcomes (e.g., the category was /d/ or /t/). To the extent that these fundamental assumptions hold, modeling changes of representations will necessarily require more parameters than the other two (and hence it is more complex and less parsimonious).

Of course, one could design a model of changes in category representations that is equally (or even more) parsimonious than a model of normalization. One way to do so, for instance, is to store only the mean of each cue for each category and do so only over a moving time window of 5 seconds. Likewise, one could implement a highly complex model of changes of normalization (e.g., storing the first 100 statistical moments of each cue, and doing so separately for each speaker). But this is why we emphasize the importance of *comparing like with like*—For each model of normalization, one can always construct a parallel and corresponding 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 (which is now Figure 36 due to an addition of a new figure in Section 4.2). In the interest of brevity, we have not added further detail about the updating formula for the parameters. We hope that the existing text and figures convey the gist of how this change model reacts to new input.

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

[…] 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 thoroughly engaging with our proposal and for the constructive feedback. We are glad that the manuscript got another set of eyes that hadn't seen the previous version. This clearly showed the need for further clarification and revision of 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).

We fully agree that the relative involvement of the mechanisms can (and likely) be different across different cues and contrasts. Indeed, this idea is central to recent work by the first and last authors, and was previously discussed in work by the second author (Kleinschmidt & Jaeger, 2015). In both the previous submission and the revised version, we discuss this in Section 6.1 (emphasis added here):

*“Existing findings thus strongly suggest that* ***no single change mechanisms can explain the full variety of adaptive responses that humans exhibit****. It seems obvious that the field will have to move beyond (in)sufficiency tests, towards experiments that determine how multiple change mechanisms jointly achieve adaptive speech perception. This will likely require research on how the relative engagement of different change mechanisms depends on stimulus properties,* ***cue and contrast*** [previously category] ***types****, task demands, or individual differences between listeners.”*

and a few lines down:

*“Similarly, it is possible* ***that the relative engagement of different change mechanisms depends on the type of phonetic contrast, or even the type of cue. This would be expected, for example, because different types of cues exhibit different degrees of within- and between-talker variability*** *(see discussions in Kleinschmidt and Jaeger, 2015, p. 179-180; Kraljic and Samuel (2007); Xie, Buxó-Lugo, and Kurumada (2021)).”*

We emphasize here that the empirical finding described by the reviewer—that adaptation to some “contrasts” appears to be talker-specific while adaptation to other “contrasts” does not—**can be just as easily accommodated by normalization accounts or changes in decision-making**. That is, the specific question raised by the reviewer does notundermine our central argument. (We note here that the same applies to many other potential issues we do not have space to discuss in detail.) We have added a new section to the SI (7.3) that summarizes and elaborates on this point. To briefly summarize our argument here:

There is no reason, for example, why the talker-specificity of changes in decision-making could not be contrast-specific in the same way that the reviewer seems to assume for changes in category representations. That the latter seems more intuitive to the reviewer (and to us prior to the work we did for this paper) might simply be due to (1) the fact that talker-specificity was originally discussed only for normalization and changes in representations, and (2) the particular finding by Kraljic & Samuel comes from an experiment on perceptual recalibration—i.e., a paradigm that to this day is *assumed* to elicit representational changes.

Similarly, if the finding by Kraljic and Samuel is caused by the *cues* involved in the two contrasts, rather than the contrasts themselves, then normalization provides just as good—well, arguably a more elegant—explanation of the finding than changes in category representations. For what it’s worth, this seems to us to be the explanation that Kraljic had in mind (see also the discussion in Kleinschmidt and Jaeger, 2015, that spectral cues vary more across talkers but are stable within talkers—making them the perfect target for *talker*-specific adaptation, where durational cues like VOT primarily vary with speech rate and thus even *within* talker). These two perspectives—cue vs. contrast-dependence of talker-specificity—make different predictions but are not distinguished by Kraljic & Samuel’s findings.

In short, the reviewer raises a critical point, and we have tried to take to heart R3’s caution against drawing sweeping conclusions from the current case studies. In particular, we did not mean to say that all existing results are compatible with all three mechanisms that we considered in ASP. Rather, we meant to argue that no existing result has been adequately evaluated under more than one mechanism. This is a subtle but important distinction. We have edited the manuscript throughout to clarify this.

[…] 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 moved this specific point out of the statement above to address it separately here.]

We agree that our decision to study the /d/-/t/ contrast may appear somewhat surprising to some. Before we submitted the original manuscript, we deliberated on related topics amongst ourselves. We opted to stick with the /d/-/t/ contrast because of how comparatively well understood its phonetics are (both in L1 and L2 accents), and because of the quality of available databases that provide information about the relevant phonetic distributions (e.g., Chodroff & Wilson (2018) and Schertz et al., (2015)). We also note that we deliberately avoid reference to talker-specificity or talker-(in)dependence in our case studies. Rather, we talk about the adaptive processes that occur with exposure to an unfamiliar talker, leaving open whether these are talker-specific or not. **This is now stated in a new footnote on p. 38.** We hope that this—together with the additional changes described above—is sufficient to address the reviewer’s concerns.

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.

First, with respect to the point regarding why we formalized the decision-making change model as is, we note that the **qualitative limitations that we identify for changes in decision-making should hold generally**, regardless of the specific implementation of our model. In principle, changes in decision-making can only lift/lower the overall categorization function (with some caveats when lapse rates are non-zero but even in those cases, the changes are still very limited in nature). This is visually illustrated by the four panels in Figure 7 (p.21).

We also note that our first submission used a different (less cognitive plausible) change model. That model 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 was on the time course (across trials) of adaptation, not the types of changes that the model could 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 and 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** (Section 2.2.2, p.24).We have also adjusted our general discussion (Section 6.2.5). 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. **To ensure that we faithfully represent the arguments and conclusions of SD 16, in the process of the current revision, we have corresponded with Dr. Sohoglu and confirmed this directly with him**.

The only description of an actual model that we found in SD16 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, Buxó-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."

Another very helpful observation. **We have added the sentence as suggested by the reviewer.** Thank you. **We now also clarify that *all* of the three change models are sensitive to prediction errors** (footnote 7, p.29). Kleinschmidt and Jaeger (2015), for example, presents 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 error 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* important in understanding processing and learning across the cognitive sciences is pretty uncontroversial.

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** (the beginning of Section 2.1, p. 12). 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 created a new figure that summarized the entire model with all the variables (Section 4.2, Figure 19).

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!). We 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. We hope the isolation of the decision-making process as a separate mechanism of adaptation will help resolve this conceptual confusion in the field.

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.

We have now added the two references to footnote 2 on p.8. We agree that under these normalization accounts, talker information is not discarded or removed and may be stored for in other types of representations (e.g., those used for voice recognition or social inference).

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).

Thank you. Another example that shows that normalization helps perception is Xie, Buxó-Lugo, & Kurumada (2021). We now cite these papers in Section 2.1.1 (p.14). The question of whether normalization is *sufficient* to explain human performance is, however, only addressed by Xie et al (2021), and even there rather indirectly. Since we discuss that question in the general discussion, we have not edited Section 2 (p.8) further.

We also note that while normalization models’ performance fall short of human performance in Crinnion et al., 2020 (as well as McMurray & Jongman, 2011), this does not speak directly to the issue whether conditioning category identity on other cues could further help.

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 raising this important point. **We now discuss this in the discussion** (Section 6.2.5, p. 78). We also added some nuance, and cite Luthra et al. (2020) on p.9 of the introduction.

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 have, in fact, 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).

Thank you for suggesting these papers. We have now added Leonard et al. (2016) and Schuerman et al. (2022) to this discussion (Section 2.1.1, p.16).

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!)

Thank you for catching this. We now consistently use *N*, following Murphy (2012).

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

Unfortunately, the symbol after “term” did not transfer to the review but it would seem that the reviewer was referring to Gothic D? That is the notation used to refer to “the data” in statistics. **We now clarify that 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).

Thank you for bringing up this point. We have edited this section to make the link to “boundary shift” clearer. We do not, however, at that point link to Xie et al (2017). The reviewer is correct that paper argues that adaptation to Mandarin-accented English involves “more than a boundary shift”. However, the argument presented in the paper does not refer to changes in the categorization function but rather to the observation that exposure also affects prototypicality judgments and phonological priming (i.e., in retrospect, “more than changes in categorization behavior” would perhaps have been a more descriptive title of Xie et al, 2017).

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 (and manipulating shifts within participants comes with additional questions & challenges that are still an area of ongoing investigations). 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! (and the Killian-Hütten et al. paper we cited/cite is the one that was published in *Neuroimage*).

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 places 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 perceptual recalibration 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.

This is indeed a paper relevant to the point that we are making here. Thank you (and added in Section 6.2.1)!