**Overview of revisions**

The main message of the reviewers’ comments, as we understood, was two-fold. First, the paper needed to be accessible and has to offer concrete ways in which the current approach can be used to conduct an informative experiment. To increase readability, we have moved some details of the databases and the three change models to supplementary information (SI) (The information can be found in SI §2 and §3, respectively). We have decided to keep some of the expositions about the modeling approach in the body of the manuscript (Section 2: Modeling adaptive challenges in speech perception). For this manuscript to be of use to researchers in the field, the assumptions we have implemented under the three change models need to be made clear with details. To reduce the complexity, we have added some explanations and visual representations (e.g., Fig.16) to better situate the details in a general overview of the approach. We hope that the new structure will help increase the utility of the current manuscript as a technical tutorial, which is likely of great interest even to researchers who are not modelers themselves, while reducing the complexity of the argumentations.

Second, the reviewers suggested that we should make clear how our approach could handle the—arguably most likely—scenario in which *all three mechanisms are simultaneously engaged*. In the previous version, we put an emphasis on a comparison of the three mechanisms and how existing experimental results do not distinguish between them. The overarching goal of our paper, however, was precisely to devise an approach that can identify the role of each mechanism as well as their relative engagement when speech perception draws on multiple adaptive mechanisms. As we detail below, we have clarified this point and discussed how ASP and similar computational frameworks can help pave the way for achieving this goal.

Finally, we note that the manuscript includes interactive figures. In case these animations continue to cause issues in printing, we have created a printable PDF downloadable from our OSF repository (<https://osf.io/q7gjp/>).

**Responses to reviewers’ main points of concern**

Next, we summarize the revisions we made in response to (1)-(4) mentioned in the letter to the editor (and repeated below). After that, we respond to the remaining comments point by point. Given the substantial revisions, we have not tracked changes. We have, however, added section numbers below wherever appropriate to indicate where the changes have been made.

1. **Clarify goals, scope, and contributions.** This was particularly helpful feedback. We have completely revised the introduction:

* After the first introductory paragraph, the introduction now provides a high-level overview of (a) our long-term goals to understand the *mechanisms* underlying adaptive speech perception and (b) the three more immediate goals of the present study: (i) to introduce a new analytical framework (ASP) that can support our long-term goals, (ii) to demonstrate *why* computational frameworks like ASP are required to advance theories of speech perception and illustrate the use of ASP through two simulation-based case studies, and (iii) to provide initial guidance on how to conduct computationally-guided empirical studies that can distinguish competing hypotheses about the relative engagement of different mechanisms. We have also clarified our rationale behind introducing ASP in a somewhat small-stepped, tutorial-like approach.
* Following this overview, a new subsection 1.1 describes the “The state of the field(s)”. This section reviews the past work in the field and states our contributions. This includes our classification of dozens of competing hypotheses into three qualitatively different types of hypotheses about adaptive speech perception (normalization, representational changes, and changes in decision-making)—a contribution that we failed to convey in the previous version.
* As part of this review, we have added new paragraphs on how neuroimaging research has illuminated roles of the decision-making mechanisms in contrast to behavioral works in the field. This motivates a synthesis of these partially overlapping, but largely divergent, lines of work, which we hope to facilitate through the use of ASP.
* The previous version of introduction did not clearly state what we consider an important contribution of ASP. As pointed out by R2, the survey of the literature already highlights a current lack of clarity in an understanding of underlying mechanism(s) of adaptive speech perception. The core contribution of our framework and the two case studies is to, for the first time, *show* that the signature results of two influential lines of research can be accounted for by any of the hypotheses, unlike common interpretations. This was *not* previously known. We now state this contribution in the new Section 3 (overview of case studies). In the same section (and in Section 1), we clarify that ASP can also be used to understand the relative engagement of the underlying mechanisms, that we demonstrate how this can be done in the general discussion (Section 6).

1. **Clarify take-home points.** Guided by reviewers’ comments, we have completely restructured the general discussion to clarify our take-home points. After briefly summarizing the findings of our two case studies:

* A new subsection (Section 6.1) discusses why we believe that it will be critical for future research to go beyond testing the *(in)sufficiency* of each of the change mechanisms—i.e., tests that assess whether any of the change mechanisms is sufficient to explain adaptive speech perception. Section 6.1 reminds readers that it is unlikely that a phenomenon as complex as adaptive speech perception relies on a single change mechanism (in line with reviewers’ comments). We briefly summarize existing evidence that—if interpreted in this light—strongly suggests that none of the three change mechanisms *by itself* is sufficient to explain adaptive speech perception, and refer to a more detailed summary in SI, §7 (the same SI section proposes additional experiments that would provide even more decisive tests of whether any of the three mechanisms is sufficient to explain adaptive speech perception, as requested by the reviewers). Instead, we argue, future work will need to address how *combinations* of mechanisms jointly explain adaptive speech perception, and how the involvement of each mechanism depends on available cognitive resources, task demands, and individual differences. And these future efforts will likely require—or at least stand to benefit from—quantitative computational modeling. This motivates the rest of the general discussion.
* We have also largely rewritten Section 6.2, which lists our recommendations as to how the field(s) can facilitate effective quantitative model comparisons. This subsection presents revised and clarified content that was already present in the previous version.Our recommendations boil down to the *when*sand *where*sof exposure and test: i.e., after how much exposure to test (*when*), whether to test repeatedly with intervening additional exposure (*when*), andthe location of exposure and test stimuli in the acoustic space (*where*). We now also state more clearly what challenges will have to be overcome for this approach to be informative. Finally, we removed one recommendation (i.e., the need to phonetically annotate stimuli and share those annotations) that was not specific to the goal to facilitate quantitative model comparison. This point is now reduced to one sentence at the start of Section 6.2.
* We have added a new conclusion section (as requested by R1).
* In addition to these edits to general discussion, we have also added three additional simulations at the end of Case Study 2 (5.3 Summary). These simulations consider three distinct, empirically attested, ways in which a phonetic contrast in L2-accented speech deviates from what is expected in L1-accented speech (= cases of contrast *reduction*, *shift,* *collapse*). The effects of these L2-accent characteristics illustrate the importance of considering underlying phonetic cue distributions in examining adaptive speech perception. We hope that these additional demonstrations will help bring home the point about the significance of predicting when and where adaptive adjustments should occur.

**(3) Shorten the manuscript and make it more accessible.** Both reviewers mentioned that the length of the manuscript could limit its impact. We have:

* Streamlined and simplified the introduction. Most notably, we had moved the introduction of the experimental paradigms for the two case studies from the introduction to the Case Studies where they become relevant (Sections 4 and 5).
* Restructured the general discussion in a way that specific directions for future research are easier to identify.
* Simplified the change model for decision-making, shortening its presentation (Section 2). We also made minor simplifications to the normalization model. These changes have numerically changed some of the results of our case studies (as would any re-run of the models, as they are probabilistic). **The qualitative patterns and our main argument remain unchanged.**
* Moved some of the technical details into footnotes or the SI (this mostly affected Section 2, with smaller changes in Sections 4 and 5). As suggested by the editor, these edits decreased the amount of the formulas from Section 2. As an attempt to bridge the gap between computational and experimental research, we have deliberately kept the somewhat tutorial-like style of Section 2. We hope that the relatively verbose exposition of our framework can aid this goal. We have also tried to further improve our figures and animations to that end.

However, these substantial reductions traded off with clarifications inspired by reviewers’ comments. The main text is still about 75 double-spaced pages long (plus more than 25 pages of references). This is perhaps not surprising given that the paper (1) reviews work from three theorical perspectives that have largely proceeded in separation, (2) draws on two lines of experimental research, while (3) combining behavioral, neuroimaging, and computational findings.

**(4) Better integration of neuroimaging research.** FollowingR1’s suggestion, we have improved our presentation of neuroimaging research. This has primarily affected the introduction (Subsection 2.1 The state of the field(s) p.5, p.9) and the general discussion (Section 6.2.2 Recommendation 5: Integration with neural models of speech perception). Additionally, we have integrated relevant neuroimaging research throughout the paper wherever relevant (e.g., at the start of Sections 4 and 5).

Finally, we made many additional revisions to improve accessibility of both the text and the code. We revised most figures. We streamlined our code, and added additional documentation, updated the library we introduce, and have removed simplifying assumptions wherever possible. Mostly, this has affected the SI, which now contains much additional information for anyone interested in applying ASP to their work.

Detailed responses to the remaining points of the reviewers are presented separately in our response letter.

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

-when introducing the lapse rate parameter on p. 21, I was not initially sure of why this parameter would receive such prominent treatment in the paper, although the case was nicely made later on in the paper.  Given the importance of lapses was not discussed in detail earlier in the paper, it could be helpful to foreshadow the importance of this parameter earlier on.

Thank you. We now anticipate the importance of this parameter when we first introduce it (Section 2.1.3: Post-perceptual decision-making: incorporating priors, response biases, and attentional lapses) and demonstrate its importance in Section 2.2.3. We are now clearer about how the introduction of attentional lapses in decision-making can explain what we commonly observe as adaptive changes in speech perception. This is a significant point because adaptive speech adaptation is not traditionally considered as a consequence of decision-making.

-on p.36: Examining Figure 14A, my impression was that different stimuli were used in the /d/-shifted vs. /t/-shifted panels of the figure.  Would the tightest control not contain the same base stimuli shifted in either direction?

(This same figure is now Figure 18A). We first answer whether this would be a good idea, and then whether that is what is done in perceptual recalibration (PR) experiments.

Unfortunately, it is unclear what “same” stimuli would even mean. Note that “sameness” would have to be achieved *across* lexical items as *different* words have to be used for each bias condition (e.g., *lemona?e* for the d-bias and *resona?e* for the t-bias). One might thus aim to achieve “sameness” across lexical items acoustically/phonetically or perceptually. But neither approach is trivial. E.g., phonetically shifting the /d/ VOTs 10msecs up is not the same—neither perceptually nor relative to the distribution of phonetic stimuli—as shifting the /t/ stimuli 10msecs down (because /d/ and /t/ typically differ in their variance). Alternatively, one could aim for equivalent *perceptual* shifts in the subjective probability of being identified as the targeted category. This would be possible but require detailed norming of many stimuli and likely entail different amounts of shifts for /d/ and /t/.

Approaches comparable to this latter alternative exist—typically under a different name and using somewhat different paradigms (e.g., unsupervised distributional learning paradigms or dimension-based statistical learning paradigms, e.g., Clayards et al., 2008; Idemaru & Holt, 2011). In our preparation of the manuscript, we explored another case study using these alternative paradigms, confirming that the indeterminacy we describe generally extends to those paradigms (but, as the reviewers pointed out, the manuscript is long as-is).

Regardless of the *possibility* of changing PR experiments to employ an approach more akin to what the reviewer suggests, it is not what has been done so far. PR experiments do *not* typically parametrically manipulate the acoustic-phonetic properties of stimuli. To our knowledge, the most common approach to the generation of exposure stimuli is to (i) record typical /d/ and /t/ versions of each stimulus (e.g., *lemonade* and *lemonate*), (ii) blend these two stimuli together under various amplitude weightings (from 100% *lemonade-* 0% *lemonate* to 0% *lemonade*-100% *lemonate*), (iii) select based on experimenters’ intuition or a small norming study the most ambiguous blend for each stimulus and call it the “shifted”, “ambiguous”, or “atypical” stimulus version (with 100% *lemonade* remaining the “typical” or “unshifted” stimulus). There are rare exceptions to this (for a review and critique of this approach, see Theodore, 2021).

In short, PR experiments do not carefully select the tokens *within* each category, nor is there any form of counter-balancing *across* the two categories. As we state in Section 6 (as well as in the SI §8. Advanced standards of data annotation and sharing), it is extremely rare that the acoustic properties are even measured (previously, we only stated that they are not *reported*). Our computational simulations capture the qualitative properties of stimuli typically used in PR experiments (in a separate project, we have measured the acoustic properties for our and dozens of other PR experiments to confirms this).

p. 37: The focus on the simulations is on the beginning of the test phase; however, should the model not also be able to account for performance throughout the test phase?  If not, why not?  Is this reflective of some additional parameter not included in the model (e.g., a reluctance to keep changing beyond a certain point?), particularly in the face of repeated stimuli?

This is another interesting point. In essence, this is a point that has received very little attention in previous experimental work. In Liu & Jaeger (2018), we showed that repeated testing reduced the effects of exposure. This and subsequent studies by us and other researchers suggest that at least 2-3 factors contribute to this:

(1) Continued unsupervised adaptation over the unlabeled input with non-bi-modally distributed acoustic properties. Test stimuli tend to span some continuum, with each location along that continuum being repeated equally often. Even when some locations are repeated more often, it tends to be those in the center of the continuum leading to a uni-modal distribution. Either way, the distribution of test stimuli violates listeners’ expectations based on lifelong input and deviate from the exposure distributions.

(2) Meta-expectations, including expectations specific to the task structure of experiments. For instance, the expectation that a 2AFC task with two possible answers displayed on the screen likely means that each option will occur about equally often.

(3) Dis-engagement due to the repetition of highly similar sounding stimuli (i.e., increased lapse rates).

As we now clarify, these factors are not relevant for our conclusions and thus not modeled in the present study (though all can be added to ASP, and some are already implemented). Since this point is not critical for the purpose of this article, we have moved it into a larger footnote (footnote 9) at the start of the result section.

p.55 the authors state "the highest accuracy is obtained for the fastest changes, and it matches that observed for changes in decision making."  Looking at the data, I am not sure that the match is especially strong, but I may be misinterpreting the data being referenced here or the level of "match" that the authors are referring to.  Perhaps this could be clarified?

The reviewer is correct about the previous version of this case study. We have revised the presentation of this result to be clear that we focus qualitative similarities—specifically, the fact that L2-accented exposure conveys an overall benefit, compared to L1-accented exposure, and that this benefit is more pronounced for /d/—the category that differs in the L2-accent. It would be hard to interpret which quantitative pattern is the ‘right’ one since existing perception experiments (for which human performance is ‘known’) do not provide phonetically annotated stimuli.

We have, however, also changed the second case study to make it even more directly comparable to existing studies on accent adaptation. Specifically, we now model exposure to Korean-accented English based on actual production data by Schertz and colleagues. As it happens, this has made the three models even *quantitatively* identical.

All of this made us realize that we had never clearly stated that the specific quantitative predictions of the change models, of course, depend on the specific accent properties. We now anticipate this point earlier in the manuscript, and at the beginning of Case Study 2. We also present three new simulations at the end of Case Study 2 that succinctly illustrate this point (Figure 28). This sets the stage to the main take-home points that we introduce in the general discussion.

I had several issues using the pdf document, including generating a printed copy.  I suggest the journal and the authors be mindful of this if this paper is moved to production.  I was on windows 10 using the current version of Adobe Acrobat when I encountered these issues.

Thank you for making us aware of this. We apologize for the inconvenience. We have noticed that some printers struggle when printing the PDF *double-sided*. We suspect that this is due to the size of the manuscript, which is in turn due to the use of animations.

We have a back-up strategy (alternative figures) in case the animations will not be accepted by *Cortex*. For now, we have also made available in OSF a PDF for printing that we hope will avoid the problems (https://osf.io/q7gjp/)? We have also changed the default state of the figures to show the *end* state of the animations, i.e., the state of maximal differences. The captions have been adjusted accordingly. This is more informative.

**Responses to the remaining comments of Reviewer 2**

This concern points to a benefit for situations where predictions of qualitative distinctions are meaningful. In particular, it would be helpful if the authors could identify conditions that can’t be accounted for by some of the mechanisms, no matter the parameter choice. For example, are there certain types of stimuli or training regimens that would only predict an effect if representations change, but can't be explained by normalization or response bias?

We addressed this point at the beginning of the letter. A new section in the general discussion (6.1) now describes how computational limitations of each change model can be used by researchers for decisive tests of the sufficiency of each change model. Normalization is indeed the one hypothesis for which we believe there is a way for a decisive experiment, and we now mention this proposal in this new section at the start of the general discussion.

Can the authors point to any truly discriminant measures by which we can rule out a mechanism as incapable of explaining a pattern of results, rather than just offering a poorer quantitative fit? Or is the whole enterprise here a question of finding the specific region of parameter space that best accommodates whatever data can be collected? This isn't necessarily disqualifying - ideas like parameter space partitioning have proven a useful tool for comparing simulations - but it raises questions about whether the model is just overly flexible. Can we fit basically all the same qualitative patterns of data with each mechanism, if we find the right parameterizations? And if so, is it worth doing a more formal parameter space partitioning analysis to see if some of the approaches more stably predict this?

We agree that careful, qualitative model comparisons are often fruitful, and we now cite Pitt, Kim, Navarro & Myung (2006) to highlight the contribution of parameter space partitioning (Section 6.2.4). However, we would like to offer two important clarifications in regards to the specific concerns raised by R1. First, it is not the case that one can fit any qualitative patterns of data with each mechanism. As we now clarify in Section 2—a novel contribution, as far as we know—there are limits to the type of change that, for example, changes in decision-making can explain (Section 3.2.3). It is also not the case that the models have tremendous functional flexibility. On the contrary, they are quite constrained in the direction of change (in the categorization function)—it is completely determined by the input. The fact that the same is true for human is informative about human perception rather than a sign of exceeding functional flexibility of the models. As we illustrate in additional simulationsin Section 5.3, it is not the case that all patterns of data can be explained by any of the three models to the same extent.

As we now state more clearly in the general discussion (Section 6.2), the problem is not model flexibility but the fact that research on speech perception continues to interpret results at an incredibly impoverished level of analysis (e.g., changes in accuracy) instead of analyzing the *link between observable stimulus properties and observable responses*. The three change models each employ only 1-2 free parameters to model incredibly complex human behavior. Relative to human perception and cognition, our models are bound to be *highly* over-simplified. If even such incredibly over-simplified models point to empirical indeterminacy of existing results, it is time to increase the informativity of experimental data and analyses, not to simplify the models.

There's a wide array of other speech adaptation and/or talker normalization tasks out there beyond those simulated here. Are there any of these for which the model offers qualitatively discriminant predictions? 

There are a handful of findings that are difficult to reconcile with one or the other change hypothesis(we now mention these findings in Section 6.1 and discuss them in more depth in the SI, Section 7). This findings typically involves manipulations that are limited in ecological validity and susceptible to participants’ meta-reasoning about the purpose of the experiment—such as playing a sine tone before a (synthesized) isolated vowel to see whether vowel perception is affected.

Perhaps most importantly though, the revised manuscript now clearly states that our ultimate goal is a framework that allows us to effectively study how the three change mechanisms *jointly* explain adaptive speech perception (in both Sections 1 and 6). None of the existing paradigms easily affords that without further advances in design, stimulus selection, and analysis. The ASP framework offers support for all of these aspects.