Friday, December 10th, 2022

To: Editorial Board of *Cortex*

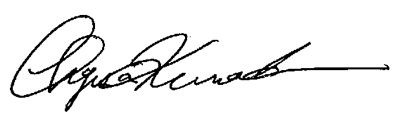
Dear Drs. Guediche and Caffarra,

# We are re-submitting our manuscript, CORTEX-D-21-00884 now titled “What we do (not) know about the mechanisms underlying adaptive speech perception: A new computational framework and review”, authored by Xin Xie, Florian Jaeger, and Chigusa Kurumada 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.

We are grateful for the constructive reviews. We appreciate that both reviewers saw that there is potentially much to be gained from developing a modeling framework like ours (which we now call **ASP for *adaptive speech perception***). As you summarized in your letter, both reviewers pointed to a need to (1) clarify the goals and scope of this manuscript, (2) clarify the take-home points, in particular whether there exist conditions for which any of the proposed mechanisms can be ruled out, and (3) shorten the manuscript and/or make it more accessible. Finally, R1 pointed out that (4) our presentation of neuro-imaging work could be better integrated with the rest of the manuscript. We have overhauled the introduction and general discussion, as well as the descriptions of the models and simulations as needed, to address all (1)-(4). In what follows, we first provide a general overview of the changes we have made and then elaborate on how we responded to (1)-(4).

Thank you for considering our resubmission for publication in *Cortex*.

Sincerely,





Xin Xie T. Florian Jaeger and Chigusa Kurumada

**1. Revision overview**

The main message of the reviewers’ comments, as we understood, was two-fold. First, the paper needed to be digestible 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 change models to Supplementary Information (§2 and §3, respectively). After deliberations, 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 clearer structure will help increase the utility of the the current manuscript as a technical tutorial while reducing the complexity of the argumentations.

Second, the reviewers suggested that we 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 experimental results do not distinguish between them. The overarching goal of this demonstration, however, was precisely to devise an approach that can identify roles of each mechanism as well as the relative engagement of more than mechanism. As we detail below, we have attempted to clarify this point, and how ASP and similar computational frameworks can help pave the way for achieving this goal.

**2. Major comments**

Next, we summarize the revisions we made in response to (1)-(4), and then respond to the remaining comments. Given the substantial revisions, we have not tracked changes. We will, however, add section numbers wherever appropriate to indicate where the changes have been made.

1. **Clarifying 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 is required to move the field forward and illustrate the use of ASP through two simulation-based case studies, and (iii) to provide initial guidance on what factors determine whether an experiment can decide between competing hypotheses about adaptive speech perception. We are now also clear that we deliberately introduce ASP in a somewhat small-stepped, tutorial-like approach because we want other researchers to take our document and understand the hypothesized mechanisms are implemented in models.
* Following this overview, a new subsection 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)—something 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.
* Importantly, the previous version of introduction did not clearly state what we consider an important contribution of ASP is. As accurately 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. This was *not* previously known. In Section 2 and 3, we state this contribution as well as the overarching goal of utilizing the ASP framework to understand the relative engagement of the underlying mechanisms.

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 5.1) discusses how the current approach can go beyond testing the *sufficiency* of each of the change mechanisms, i.e., tests that can reject the hypothesis that any of the change mechanisms is sufficient to explain adaptive speech perception. (We have also added a brand-new section in the supplementary information §7 to survey *existing* findings that—if interpreted in this light—that already shed light on this question.) Section 5.1 clarifies that such sufficiency tests can never 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. This, we believe, motivates the rest of the general discussion where we provide concrete, actionable, proposals for researchers in the field to implement informative tests of mechanisms underlying adaptive speech perception.
* We have also overhauled Section 5.2, where we list a number of *general* considerations 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 are also state more clearly what challenges will have to be overcome for this approach to be informative.
* 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 Section 4 (4.3 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 undergird 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 shifts should occur.

**(3) Shorten the manuscript and make it more accessible.** Both reviewers mentioned that the length of the manuscript could limit its uptake. While we concur with the reviewers, we also believe that the length is justified given that the manuscript bridges research from three theorical perspectives that have largely proceeded in separation and draws on two lines of experimental research, while combining behavioral, neuroimaging, and computational findings. Additionally, some of the reviewers’ questions and suggestions helped us clarify significant contributions of the approach, which inevitably increased the contents. Thus the current manuscript is not much shorter than the original submission. However, through carefully addressing the reviewers’ comments, we have made efforts to increase its accessibility and clarity. Specifically, 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 sections where they become relevant (Sections 3 and 4).
* 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 supplementary information (this mostly affected Section 2, with smaller changes in Sections 3 and 4). As suggested by the editor, these edits decreased the amount of the formulas from Section 2. If further cuts are required, we could move most of Section 2 into the supplementary information. However, both in our conversations with experimenters, and in our own experience reading computational papers, we often feel the frustration that comes with a lack of shared backgrounds. 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 serve that purpose. We have also tried to further improve our figures and animations to that end.
* As a result of the comprehensive review we provide, almost 25% of the manuscript length are references. If recommended by the reviewers, we could further cut background information on the different lines of research, which would also reduce the length of the bibliography.

**(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, 9) and the general discussion (Section 6.2.2 Recommendation 6: Integration with neural models of speech perception p. 76-78). Additionally, we have integrated relevant neuroimaging research throughout the paper wherever relevant (e.g., at the start of Sections 3 and 4).

**3.　Remaining comments**

**Responses to the remaining comments of Reviewer 1**

I would imagine that there is a good case to argue that adaptation could occur at all three levels simultaneously, at least to some degree.

We completely agree. In fact, we consider ASP’s ability to model any *combination* of mechanisms a *major* strength. As we have shown in our cases studies, the empirical coverage of individual mechanisms is more powerful than previously assumed, and exhibits dynamics that go beyond what can easily grasped by intuitions. This is even more true when the effects of combinations of these mechanisms are considered. We have revised the introduction (p. XX) and general discussion (p. XXX) to state this clearly.

-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 how the introduction of attentional lapses allows adaptive changes in speech perception—which are not traditionally considered a consequence of decision-making—to be explained by changes in 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. 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 on p. XXX, 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: e.g., the expectation that a 2AFC task with two possible answer 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, none of these factors are modeled in our study (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 on p. XXX 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. We have revised the presentation of this result to be clear that we mean qualitative similarities (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).

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 (5.1) 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 be a way for a decisive experiment, and we now mention this proposal in the 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?

No, as we now more clearly highlight 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 (p. XXX). It is also not the case that the models have tremendous functional flexibility. 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 now state more clearly in the discussion, 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 & analyses, not to simplify the models.

The new subsection 5.1 further describes how specific computational limitations of each change model can be used to test whether that change model is sufficient to explain adaptive speech perception.

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?

Yes, we think that this could be a worthwhile future endeavor. It took us about 1-2 years to distill from the literature the generalization of three types of mechanisms, develop ASP to implement these mechanisms with as few controversial assumptions as possible, and validate it in simulation studies (beyond those shown in the paper), and understand the consequences well enough to write this paper. Parameter space partitioning is now mentioned in the general discussion, as an interesting way forward.

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? 

Neither our review of the literature, nor our own research, revealed such ‘silver bullet’. There are a handful of findings that are hard to reconcile with one or the other change hypothesis *for the particular paradigm those studies employ* (we now mention these findings in Section 5.1). This typically involves manipulations that lack in ecologically validity and lend themselves to interpretations by participants in terms of meta-reasoning about the purpose of the experiment—such as playing a sine tone before a (synthesized) isolated vowel to see whether this affects vowel perception.

But, as we now state at the start of the general discussion, most of the common paradigms are (to varying extents) subject to the same issues that our two case studies highlight. Additionally, the new Section 5.1 discusses some of the additional paradigms and the conclusions one can or cannot draw from them.

Perhaps most importantly though, the revised manuscript is much clearer that the ultimate goal must be a framework that allows us to effectively study how the three change mechanisms *jointly* explain adaptive speech perception. 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.