Friday, December 10th, 2022

To: Editorial Board of *Cortex*

Dear Drs. Guediche and Caffarra,

# We are re-submitting our manuscript, CORTEX-D-21-00884 “Most experiments on exposure effects in speech perception do not distinguish between underlying mechanisms: A computational 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 (and encouraging) reviews. 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 was lacking and not well integrated with the rest of the manuscript. Before we turn to an overview of how we addressed (1)-(4), we would like to make a more general observation.

We appreciate that both reviewers saw that there is potentially much to be gained from developing a model like ours (which we now call **ASP for *adaptive speech perception***). We also understand that R2 found the paper lengthy, and was somewhat underwhelmed by its perceived contributions. Much of this, we think, was due to presentational issues the reviewers identified. As we summarize below, our revision clarifies that several of the contributions the reviewer attribute to the Introduction are actually novel insights we derived only after applying the ASP model. One key insight is the recognition that a framework like ASP is needed to solve the empirical indeterminacy. Simply put, there is no brief way to introduce a model that can capture the complexities of one of the most puzzling aspects of speech perception, and illustrate its workings to a broad audience. Neither, as the revised paper now clarifies, do we know a trivial path forward. This, we submit, is neither a short-coming of ASP, nor a short-coming of the paper but reflects the realities of this domain: as we now anticipate in the Introduction and clarify in the General Discussion, our case studies show *that* ASP and similar frameworks will be *required* to move the field forward. Future computational studies are needed to determine the exact types of designs necessary to empirically distinguish between the competing mechanisms. We now explicitly explain *why* this is far from trivial, and beyond the scope of the present work.

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.

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 tutorial-like approach because we want other researchers to take our document and *understand* the models.
* Following this overview, a new subsection describes the “State of the field(s)”. This section reviews 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.
* Importantly, the previous version of the manuscript did not clearly state what we consider an important contribution (reflected in feedback from R2). We are now clear that the literature we review in the Introduction leaves open *whether* the signature results of previous studies distinguish between competing hypotheses, resulting in a *possible* empirical indeterminacy. We believe that our case studies are the first to *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.

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, the revised discussion now *begins* with a summary of recommendations for future work:

* We are now clear right from the start of this section that we are not yet in a position to point to trivial design choices that are guaranteed to distinguish between all three mechanisms *without quantitative model comparisons* (something that the reviewers asked). We have been able to identify only one *potential* ‘shortcut’ that does not require quantitative model comparisons, and might be able to reject one of the hypotheses (normalization) as sufficient explanation for a specific adaptive behavior. We now discuss this approach, as part of our recommendations (p. XXX).
* However, this still leaves at least two mechanisms to be distinguished between by means of quantitative model comparisons. Additionally, model comparison will likely be required to distinguish between more specific alternatives *within* each of the three hypotheses (e.g., to test whether category expansion or category shift—both are representational changes—can explain a given adaptive behavior).

To facilitate such model comparisons, we make the same general recommendations as in the previous manuscript. We have, however, revised them to be clearer about the overall take-home point: in essence, 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*). The dependencies on stimulus properties are also the reason why it is impossible at this point to provide specific simple recommendations to researchers.

* Finally, the revised General Discussion is now more specific *how ASP or similar frameworks can* *facilitate* model comparisons, and *why this is not trivial* (Figure XXX on p. XXX)—i.e., why we don’t already provide more specific design recommendations in this paper.

1. **Shorten the manuscript and make it more accessible.** Both reviewers mentioned that the manuscript was long and challenging to read/review. 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. While we have not been able to drastically reduce the length of the manuscript, we have implemented reviewers’ helpful suggestions wherever possible. The main text of the document has been shortened from 67 to XXX double-spaced pages. This was achieved primarily by:

* Restructuring the Introduction. For example, we had moved the introduction of the experimental paradigms for the two case studies from the Introduction to the sections where they become relevant (3 and 4).
* Simplifying the change model for decision-making and shortening its presentation (Section 2). Moving non-critical technical details into footnotes or the Supplementary Information (this mostly affected Section 2, with smaller changes in Sections 3 and 4).
* As a result of its scope, almost XXX% 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 cut the length of the bibliography.

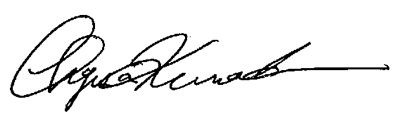
As an attempt to bridge the gap between computational and experimental research, we have deliberately kept the somewhat tutorial-like style of Section 2. We hope that the relatively verbose exposition of our framework can serve that purpose. We have also tried to further improve our figures and animations to that end. However, if the manuscript is still too long or inaccessible, we could move further details of the framework into an appendix OR collect them in a methods article (if that is suitable).

1. **Better integration of neuroimaging research. Following** R1’s suggestion, we have improved our presentation of neuroimaging research. This has primarily affected the Introduction (p. XXX) and the General Discussion (p. XXX-XXX). Additionally, we have integrated relevant neuroimaging research throughout the paper wherever relevant (e.g., at the start of Sections 3 and 4). In particular, the introduction now also clarifies that:

*“Compared to the behavioral research …, it is more common in neuroimaging work to directly contrast hypotheses about different mechanisms … However, in contrast to behavioral work, neuroimaging research tends to not distinguish between hypotheses (A) and (B), grouping both hypotheses together as functionally distinct from higher-level, decision-related mechanisms further downstream (C).*” [footnote 3]

In addition to these major revisions, we have also updated the data sampling method (details described in the Supplementary Information) so that a phonetically balanced set is used in the simulations. This numerically changes the simulation results for the normalization model. The qualitative patterns and our main argument remain unchanged. Next, we respond to the remaining comments point by point.

Sincerely,





Xin Xie T. Florian Jaeger and Chigusa Kurumada

**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. xXX) 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) and demonstrate its importance in Section 2.2.3. We have also revised Section 2.2.3 to be 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?

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. E.g., shifting the /d/ VOTs 10msecs up is not the same as shifting the /t/ stimuli 10msecs down (because /d/ and /t/ typically differ in their variance). One could aim for equivalent shifts in the subject probability of being identified as the targeted category, which would require detailed norming of many stimuli and likely entail different amounts of shifts for /d/ and /t/. Comparable approaches 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. Our computational simulations capture the qualitative approach taken 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).

Empirical work often assumes no change of performance during test and tends to use short test session with many participants to overcome this problem experimentally. 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 (LINK)? 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.

We addressed the general point of the reviewer at the beginning of the letter. No, there are no known simple conditions that distinguish between all three mechanisms. We believe that this is one of the reasons *why* the three types of theoretical perspectives have co-existed for many decades, without serious attempts to distinguish between them. Our general recommendations, however, are *precisely pointing the way to how researchers can determine what type of stimulus regime can distinguish between the competing perspectives.* This was not previously the case but is now made possible through the development of ASP. 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. Moreover, it is possible that listeners employ combinations of these mechanisms (potentially for different stimulus and task regimes). One strength of ASP is that it can model any *combination* of mechanisms. We have now made this point clear in the Introduction (p. xXX) and General Discussion (p. XXX).

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?

Normalization is indeed the one hypothesis for which we believe there is be a way for a decisive experiment, and we now mention it as part of the general discussion (p. XXX).

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—for the first time, as far as we can tell—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. Human listeners also exhibit this directional constraint. More generally, the normalization and decision-making models are particularly constrained in what they can do *conditional on specific exposure and test stimuli* (hence our recommendations).

Finally, we point to the fact that all three change models employ only 1-2 free parameters to model incredibly complex human behavior. Relative to human perception and cognition, our models are in fact highly simplified. If even such incredibly over-simplified models as ours point to empirical indeterminacy of existing results, it’s time to increase the informativity of experimental data & analyses, not to simplify the models.

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’. One would have to conduct similar simulations, as we have done here for PR and AA experiments, for the alternative paradigms the reviewer has in mind. We have done so for what we believe to be the third-most commonly used paradigm (“unsupervised distributional learning”, Clayards et al., 2008 and follow ups, or “dimension-based statistical learning”, Idemaru & Holt, 2011 and follow-ups). Unlike PR and AA experiments, experiments in this type of paradigm usually report the exposure statistics, and location of test items. This makes them more informative for all the reasons we mention in our recommendations (we continue to refer to these studies as part of our recommendations).

The specific experiment we analyzed (Clayards et al., 2008) would seem to rule C-CuRE normalization (this is why we selected it). However, if normalization also involves standardization of cues (as, e.g., proposed for Lobanov normalization of vowels), then normalization can account for the Clayards findings. We now discuss this finding as part of the general discussion. The more general problem though is that each new paradigm requires additional modeling decisions. Given the sparsity of previous work that has aimed to address any of these questions, this usually entails substantial research. For example, both the distributional learning paradigm as studied in Clayards et al. (2008) and the dimensional learning paradigm (e.g., Liu & Holt, 2015) involve *unlabeled* exposure, which requires specification of how listeners draw inferences about category labels for such stimuli. (Given the length of the manuscript and given that Section 2 states that our modeling is limited to labeled input, we decided to not add discussion of these issues).