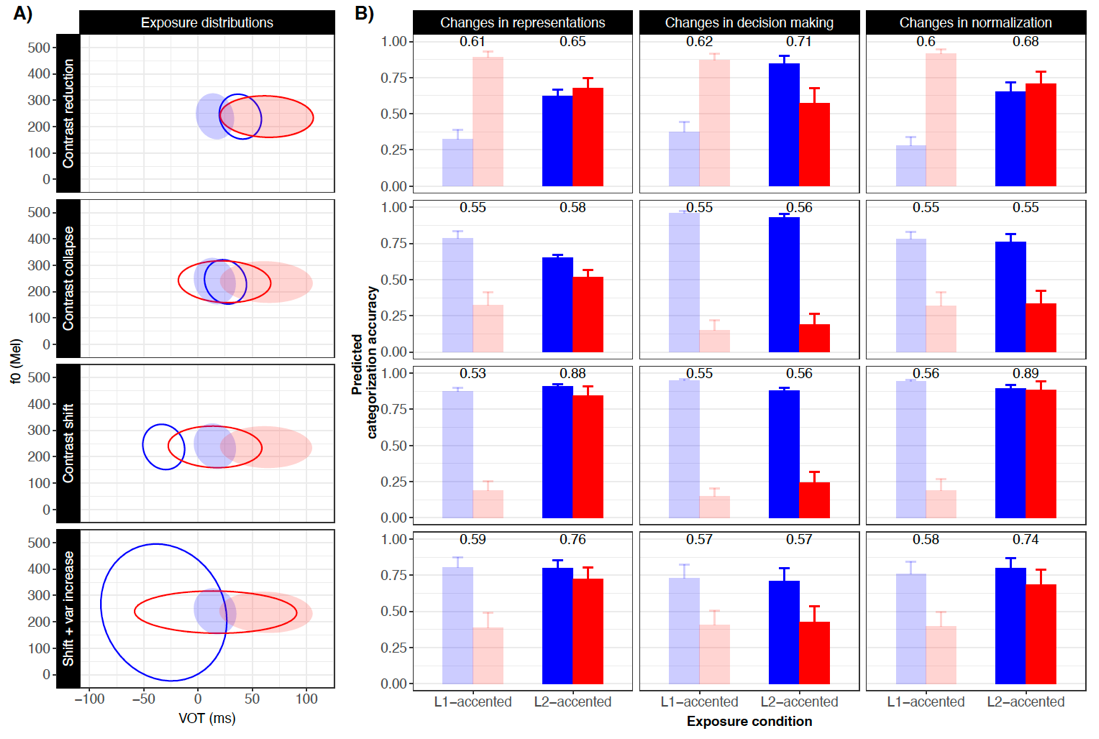
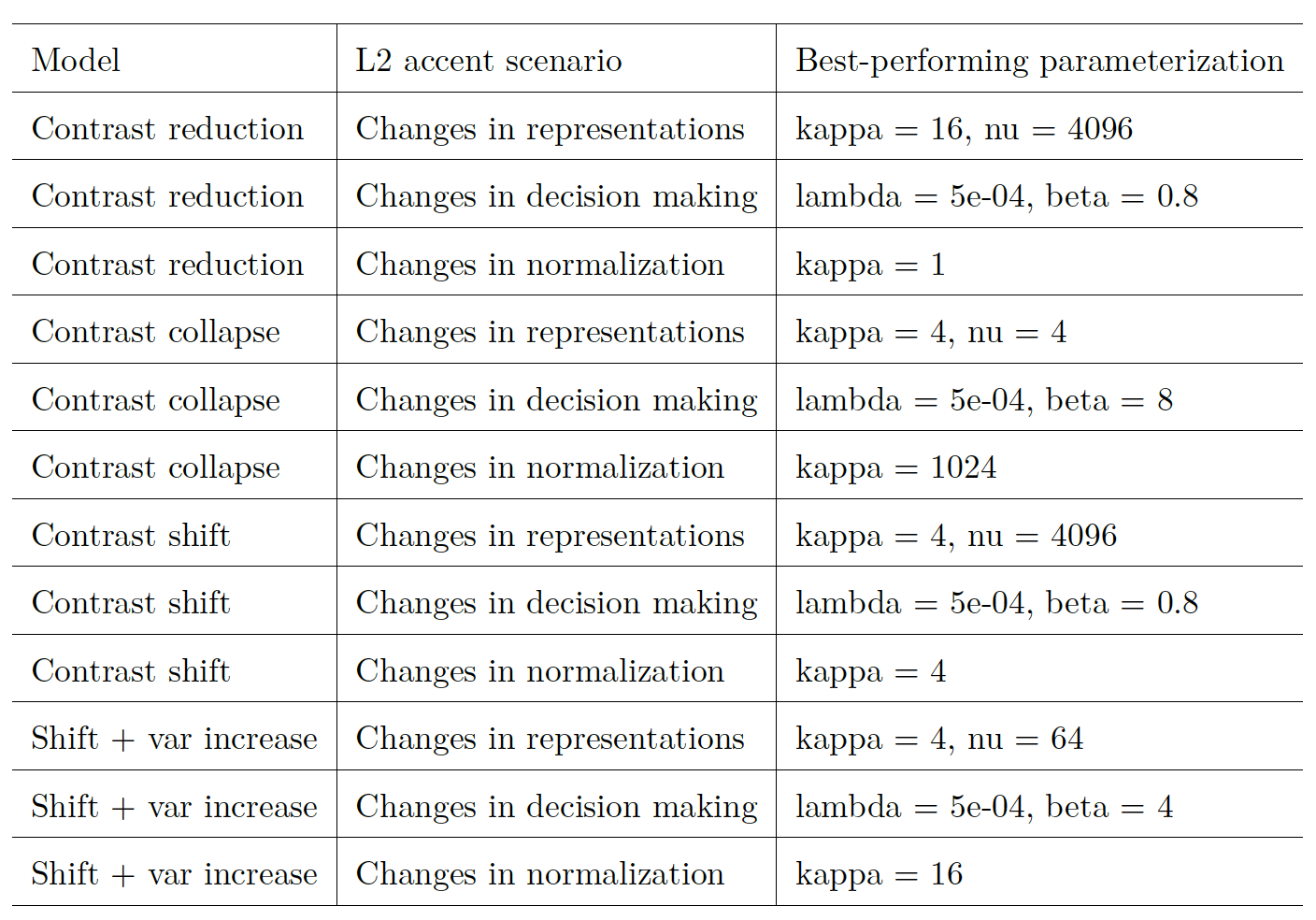
# Summary of AA section



CAPTION. .^[We extended the range of parameter values considered for each change model until we saw that more extreme values resulted in decreasing accuracy. While the exact values of the best-performing parameters depend on the level of granularity in searching the parameter space, the searched range allows us to capture the general relationship between the parameters and model behaviors.]

Paralleling Case Study 1 on perceptual recalibration, we find that any of the three change mechanisms can \*qualitatively\* explain the signature results of experiments on accent adaptation. Even non-representational mechanisms---normalization and decision-making---can explain improvements in the perception of L2 accents that differ in complex ways from L1 listeners’ prior expectations. And, contrary to common assumptions, changes in decision-making can lead to \_overall\_ improvements in recognition accuracy, rather than only zero-sum trade-offs. Overall improvements in accuracy after exposure to an unfamiliar accent are thus \_not\_ diagnostic of listeners' adaptation to L2-specific category realizations. This supports the conclusion of Case Study 1: at the level of analysis that is commonly applied in previous work (and thus here), experiments on accent adaptation do not decisively rule out any of the three mechanisms as a driver of the facilitatory effects of exposure.

Beyond the similarity in their qualitative predictions, the three change models exhibit quantitative differences. For the hypothetical L2 accent, the maximal benefits of L2-accented exposure are predicted to be largest for the representational change model (overall improvements from `r best\_performing\_case.representations.both$mAcc[1]` to `r best\_performing\_case.representations.both$mAcc[2]`), followed by the decision-making change model (from `r best\_performing\_case.bias.both$mAcc[1]` to `r best\_performing\_case.bias.both$mAcc[2]`) and smallest in the normalization change model (`r best\_performing\_case.normalization.both$mAcc[1]` to `r best\_performing\_case.normalization.both$mAcc[2]`). However, these differences do not mean that the representational change model is \_always\_ more advantageous than the other two change models. Rather, the quantitative results of Case Study 2 illustrate a critical insight of the ASP framework: the specific predicted benefit of L2-accented exposure under each mechanism depends on the statistics of the exposure and test stimuli, relative to L1 listeners’ prior expectations. For experiments that employ naturally accented exposure and test tokens, this also means that the results are predicted to depend on how exactly the acoustic-phonetic distributions of an L2 accent deviate from those of the L1 accent.

To further illustrate this point, we simulated four additional scenarios of adaptation to an L2 accent (Panel A in Figure \@ref(fig:AA-plot-results-additional-cases)). Each of these scenarios is representative of attested L2-accents, selected to illustrate some of the dynamics of the different mechanisms and their dependence on exposure and test statistics. The first row of Figure \@ref(fig:AA-plot-results-additional-cases) shows an example of \*contrast reduction\*, in which the L2 accent shows a greater category overlap than the L1 accent. Here it is simulated by shifting /d/ category towards /t/ along VOT while keeping the latter unchanged [as qualitatively attested for, e.g., vowels in Spanish-accented English, @wade2007]. The second row of Figure \@ref(fig:AA-plot-results-additional-cases) shows a more extreme example, exhibiting almost complete \*contrast collapse\* [similar to the loss of the [s]-[`r linguisticsdown::cond\_cmpl("θ")`] contrast that can occur in Mandarin-accented English, @zheng-samuel2020]. Here it is simulated by shifting both /d/ and /t/ towards a common mean. The third and four rows show examples of \*contrast shift\*, where two categories in L2-accented speech are shifted along one or more of the cue dimensions as compared to L1-accented speech. Here it is simulated by shifting both /d/ and /t/ categories towards lower VOT [qualitatively attested for, e.g., word-initial stops in French-accented English, @Sumner2011a]. The only difference between these two scenarios is that the bottom scenario includes an increase in category variance in addition to a shift in the category means.

The four different accent scenarios differ significantly in terms of the difficulty they are predicted to pose for L1 listeners prior to exposure (see L1-accented exposure condition in Panel B of Figure \@ref(fig:AA-result-comparison-best-three-models-case2). Contrast reduction (first row), for example, resembles the scenario considered in Case Study 2: without informative exposure, L1 listeners unfamiliar with the accent are predicted to exhibit substantially more difficulty with the recognition of /d/ than the recognition of /t/ . The opposite is observed for our example of contrast collapse (second row) and contrast shift (bottom two rows). This illustrates how differences in the realization of L2 vs. L1 accents can give rise to starkly different perceptual consequences for L1 listeners.

The four scenarios also differ in the maximal extent to which exposure to the L2 accent can improve recognition accuracy. We describe three aspects of this. First, confirming the point we raised above, changes of representations do not always outperform the other two mechanisms. For the case of contrast reduction (top row), changes in decision making and normalization perform as well or better than changes in category representations (for the parameter values considered here). This shows that the computational complexity afforded for the representational change model does not always result in higher recognition accuracy than the other two mechanisms.

Second, the same change model can support more or less overall accuracy improvements, depending on the specific differences between the L2 and L1 accent. Consider the case of contrast shift (third row). While changes in representations and normalization both predict large improvements in overall recognition accuracy in the L2-accented exposure, changes in decision making do not (i.e., the difference in the predicted accuracy across L1- and L2-accented exposure conditions is negligible). This outcome aligns with the zero-sum trade-offs often attributed to this change mechanism (i.e., The recognition of /d/ improves at the cost of the recognition of /t/). The heterogeneity of the results across the three scenarios makes it clear that the trade-off is not an inherent property of the decision-making mechanism. It is instead a possible outcome arising from a combination of the specific change model and properties of the L2 accent (or L1-L2 accent differences).

Third, \_none\_ of the change models predicts significant benefits from L2-accented exposure for contrast collapse (second row). This highlights an important, yet often under-appreciated, insight: sometimes null results are *\_predicted\_* for exposure-test experiment, because of the specific stimulus statistics [see @tan2021; @zheng-samuel2020]. Put differently, an absence of improvements does not \_necessarily\_ constitute evidence against accent adaptation or any of the three mechanisms.^[This also means that individual differences in accent adaptation around such a null effect are \_not\_ expected to be indicative of the individual’s ability to adapt to L2-accented speech, but instead are likely to reflect random noise [under-mining the central argument made in @zheng-samuel2020].]

While the scenarios considered here cover only a small subset of the cross-accent differences that exists in the world, they serve as an initial illustration of a general challenge to adaptive speech perception: different types of cross-talker differences (including L2 accents) constitute different adaptive challenges, and benefit from different \_rates\_ and \_types\_ of adaptation. Since the properties of unfamiliar accents are not \*a priori\* known to listeners, this means that the optimal level of flexibility is achieved if the neural and cognitive systems underlying adaptive speech perception may adjust based on the properties of the input [see the discussion of the trade-off between flexibility and stability in @kleinschmidt-jaeger2015, pp. 180-182]. In Figure \@ref(fig:AA-result-comparison-best-three-models-cases) (Panel B), this is reflected in the fact that the best-performing parameterizations vary across the different accent scenarios.

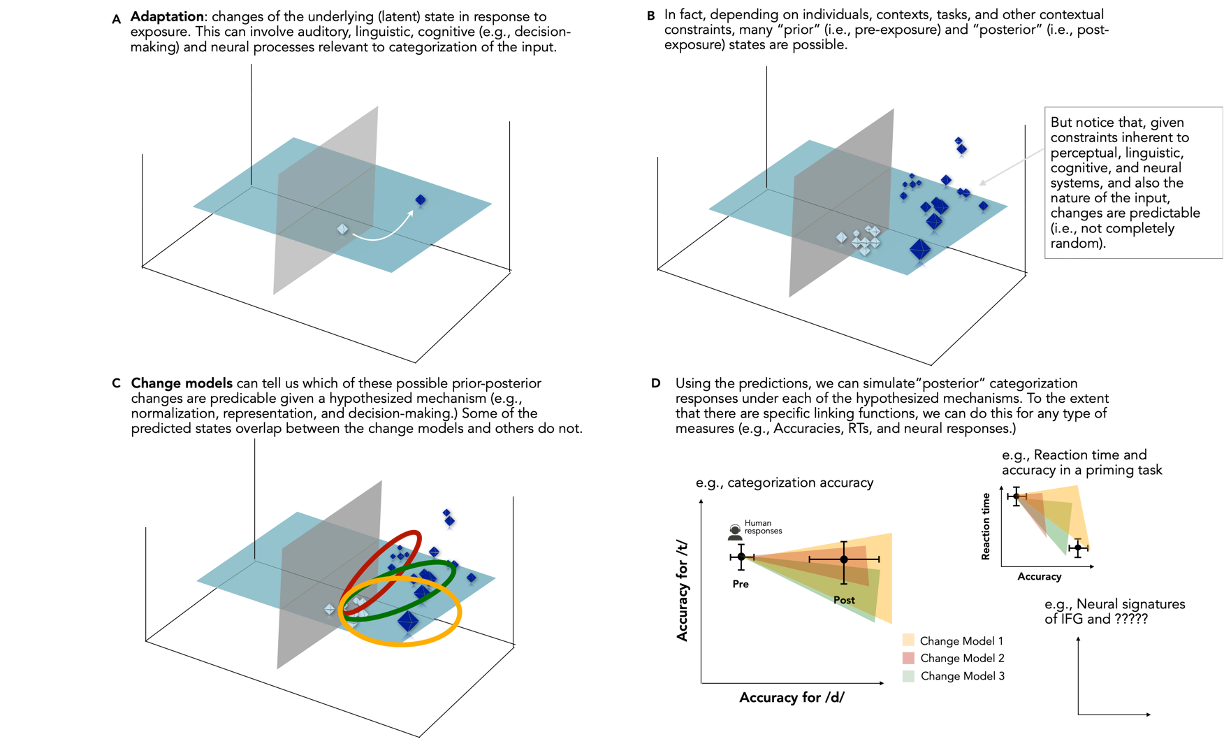
In summary, benefits of L2-accented exposure expected under each change mechanism hinges on the specifics of L2-accented categories. Beyond the accent scenarios examined here (cue-reweighting, contrast reduction, contrast collapse and contrast shift), naturally produced L2 accents exhibit substantial variations: L2-accented categories often deviate from L1-accented categories not only in terms of their means as simulated here, but also in variances and/or covariances [the category-specific correlation between cues, e.g., @smith2019; @wade2007; @vaughn2019; @xie-jaeger2020]. These arguably more complex differences can pose difficulties to L1-listeners and researchers alike. At the same time, they also offer a unique and fertile ground to test the relative engagement of each change mechanism. Because the three change mechanisms would predict distinct perceptual outcomes depending on the properties of the exposure and test stimuli, these differences can be used to distinguish between model predictions. In the general discussion below, we explore how such tests are possible.

# General discussion

The two case studies we have presented provide two important insights. First, 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). Additional case studies not reported here suggest that this finding is likely to generalize at least partly to other types of exposure-test paradigms, such as distributional learning paradigms over a single phonetic cue [@clayards2008; @kleinschmidt-jaeger2016cogsci] and dimension-based statistical learning paradigms that manipulate the relative informativity of cues [@idemaru-holt2011; @idemaru-holt2020].^[There are, however, important differences in the extent to which different research traditions have already implemented the recommendations we make below. For example, some lines of research tend to interpret findings relative to the distribution of phonetic cues in the speech input [e.g., studies on distributional learning, @bejjanki2011; @clayards2008; @kleinschmidt-jaeger2016pbr; @malone2021; studies on dimension-based statistical learning, @idemaru-holt2020; @schertz2016; for review, see @schertz-clare2020], whereas this remains the exception in other lines of research [but see, e.g., @hitczenko-feldman2016; @tan2021 for accent adaptation; @drouin2016; @theodore-monto2019 for perceptual recalibration]. See also SI, \@ref(sec:discussion-sufficiency).]

As already mentioned, this does not mean that normalization, representation, and decision-making mechanisms always predict the same response patterns—they do not. Nor does it mean that the relative involvement of the three mechanisms cannot be empirically determined through experiments. Rather, as we show in more detail below, it means that the \_design\_ and \_level of analysis\_ employed in many studies severely limit what the resulting experiments can tell us about the relative engagement of the different change mechanisms. This, we argue, calls for changes to the paradigms and analyses that are standardly employed in research on adaptive speech perception, and we make specific recommendations below.

Key to these recommendations is the second insight derived from our case studies: as we illustrated in the discussion of Case Study 2, each of the three mechanisms yields different predictions depending on the specific acoustic-phonetic properties of the exposure and test stimuli. This flexibility is not a weakness of the change models but rather a necessity: since human listeners are sensitive to the acoustic-phonetic properties of the input, any adequate model of adaptive speech perception must be so, too. Critically, however, the flexibility of each change models is constrained by its computational limitations. This is schematically illustrated in Figure \@ref(fig:changes-in-states). On the one hand, since listeners generally benefit from exposure, any plausible change model has to predict changes that move the model from its state before exposure in the direction of the ideal state (an idealized learner with perfect knowledge of all relevant statistics). This causes the models to be similar to each other (and to listeners) in their predictions (B). On the other hand, the \_specific\_ states that can be reached given both the state prior to exposure and the input received during exposure differ between change models (C). By recognizing this, we can simulate categorization responses under different hypothesized mechanisms *post* exposure to specific instances of acoustic inputs. By employing designs and analysis approaches that take advantage of this link between acoustics and participants’ responses, it is possible to move towards more informative experiments (D).



Below we discuss how future research can best achieve such experiments through a focus on computationally-informed experiment design and quantitative model comparison. First, however, we discuss \*why*\** we believe that computational simulations and model comparisons will be critical to advancing the field.

## Beyond qualitative (in)sufficiency tests

For our two case studies, we considered each of the three change models of the ASP framework \_*separately\_*. This served two purposes. The first was presentational: by presenting the predictions of each model while the other two change models were ‘switched off’, we hope to have provided readers with clearer intuitions about the inner workings of each model. The second purpose was the test of \_(in)sufficiency\_ of each change mechanism in explaining the signature results from the two paradigms: we asked whether each of the mechanisms is sufficient to predict adaptive changes of speech perception. Ultimately, however, research on adaptive speech perception will have to go beyond such sufficiency tests. With a phenomenon as complex as human speech perception, it is most plausible that multiple mechanisms, from early perception to decision-making, \_jointly\_ contribute to the observed adaptive behavior.

Indeed, existing evidence supports the idea that none of the three mechanisms is sufficient to explain the range of adaptive behaviors listeners exhibit in response to recent exposure. We summarize this evidence in more depth in the SI [\@ref(sec:discussion-sufficiency); see also @weatherholtz-jaeger2016]. For example, the finding that non-speech stimuli---such as sine tones---can affect subsequent vowel perception [e.g., @holt2001; @holt2006; @huang-holt2011] is easily explained in terms of pre-linguistic normalization but difficult to explain through changes in linguistic representations or decision-making [see also @chodroff-wilson2020]. Other findings, however, cannot easily be accommodated by an account that solely relies on pre-linguistic normalization. One example comes from research on perceptual recalibration: the effects of exposure do not only depend on the acoustic-phonetic cues of the exposure tokens but also their category labels [e.g., @norris2003, Experiment 2; for details, see SI [\@ref(sec:discussion-sufficiency)]. Further support for the hypothesis that adaptive speech perception is the result of multiple mechanisms comes from neuro-imaging studies: across paradigms, adaptive speech perception implicates brain regions ranging all the way from pre-cortical structures [e.g., in the brain stem, @REFS] to the frontal lobe [@REF].

In short, research on adaptive speech perception is at an important juncture. On the one hand, existing findings suggests that no single change mechanisms can explain the full variety of adaptive responses that humans exhibit. Put this way, 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 depend on stimulus properties, task demands, or individual differences between listeners. For example, simple, unidimensional shifts of categories (as simulated in perceptual recalibration experiments) may engage change mechanisms different from those that are used to navigate more complex shifts as seen in natural accent [for discussion, see @baese-berk2018; @samuel-kraljic2009; @zheng-samuel2020]. Another, mutually compatible hypothesis, holds that the earliest moments of exposure to an unfamiliar talker primarily reflect change mechanisms that are computationally less flexible but simpler. With increasing input, more flexible mechanisms, which ultimately can support higher recognition accuracy, would increasingly come to determine listeners’ behavior.^[In ASP, this is most naturally be accounted for in terms of slow ‘learning rates’ for the more complex change model (e.g., high $\kappa\_[0,c]$s and $\nu\_[0,c]$s for changes in representations).] Such an approach to inter-talker variability would reduce the risk of overfitting parameters to the input, a risk that applies more strongly to change models with a larger number of parameters [for discussion, see @apfelbaum-mcmurray2015; @kleinschmidt-jaeger2015; @toscano-mcmurray2010]. 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-jaeger2015, p. 179-180; @kraljic-samuel2007; @xie2021cognition].

However, the paradigms and analyses commonly employed in the study of adaptive speech perception are not well-suited to address these questions. This issue is most obvious for the behavioral paradigms we have discussed in our case studies: if a paradigm cannot distinguish between three mechanisms \_even if one assumes that only one of these mechanisms is at work\_, that paradigm cannot possibly determine the relative engagement of multiple mechanisms. However, neuro-imaging studies are not exempt from this issue. While such studies might determine the involvement of multiple brain regions, they leave open what types of computations are performed in each brain region. It is not known, for example, whether pre-cortical areas involved in auditory processing are necessarily restricted to normalization or whether these areas can take into category identity [e.g., through documented feedback projections from cortical areas, @REF]. Similarly, it is an open question whether differential activation of early cortical areas indicates changes in category representations or normalization [e.g., the involvement of Heschel’s gyrus in adaptive speech perception, @REF]. These observations about the current state of the field motivate the recommendations we present next.

## ## Methodological advances to facilitate more informative model comparisons {# sec:methodological-advances}

We offer six concrete recommendations as to how future research on speech perception can benefit from computationally-guided experiment design and data analysis methods.

### ### Recommendation 1: *Pre*diction based on sufficiently constraining theories instead of (only) informal reasoning

Our first recommendation is to develop and employ *strong* theories and models, in the sense of Platt’s strong inference approach to scientific inquiry [@platt1964]. The majority of published research on adaptive speech perception---including some of our own work---continues to employ informally formulated, under-specified hypotheses. This leaves too much room for \*ad-hoc\* or \*post-hoc\* reasoning, with all the downsides that have been discussed in the context of the replicability crisis and elsewhere [e.g., @REFS]. Even for research explicitly framed in terms of more or less clearly specified theories like “distributional learning” or “exemplar theory”, reliance on informal reasoning alone is both risky and likely to miss insights that can be gained if computational frameworks are employed.

We have experienced this in our own work---for example, in @kleinschmidt-jaeger2015 we were initially surprised that standard perceptual recalibration results can be explained through changes in beliefs about category \_*variances\_* rather than only through changes in beliefs about category means. It was the use of a computational model with spelled-out linking hypotheses about the entire chain from inputs (stimuli) to participants’ categorization responses that led to that insight---an insight that in \_hindsight\_ is obvious. Another example comes from the now classic study “The Weckud Wetch of the Wast” [@maye2008]. The study exposed listeners to speech in which the vowel categories had undergone systematic (phonological) shifts and observed adaptive changes in subsequent vowel recognition. Based on additional control conditions, Maye et al. (2008) reasoned that this finding could not be explained by “general relaxation of the criterion for what constitutes a good exemplar of the accented vowel category” but instead argued that listeners had learned ways in which vowel representations are shifted. A recent study [@hitczenko-feldman2016] revisited this result, deriving predictions from competing accounts within a Bayesian ideal adaptor framework [@kleinschmidt-jaeger2015]. Hitczenko and Feldman found that the results of Maye et al. (2008) were just as well, if not better, predicted by a learning model that expanded the categories, rather than shifting them.

We submit that computational models are not only useful, but indispensable given the complexity of adaptive speech perception. All major theories of adaptive speech perception agree that two types of information affect listeners’ interpretation of speech from an unfamiliar talker: (1) listeners’ prior expectations based on the statistics of previously experienced speech input; and (2) the statistics of the present speech input (i.e., in test) relative to those prior expectations. In exposure-test experiments like those discussed in our case studies, (1) is further divided into two components: (1a) expectations based on the long-term speech statistics experienced prior to the experiment; and (1b) the input experienced during the exposure phase, where (1b) is assumed to incrementally change (1a) that a listener brings into a subsequent perceptual response. In short, there is broad agreement that the commonalities and differences between prior experience, exposure, and test jointly affect listeners’ behavior in our experiments. This points to complex interactions that are difficult to understand without the use of computational models [for further examples and discussion, see @apfelbaum-mcmurray2015; @sohoglu-davis2016; @tan2021; @theodore-monto2019; @toscano2018; @xie2021cognition].

Just as computational researchers benefit from familiarity with experiments, experimental psychologists and neuroscientists can benefit from increased familiarity with the computational models that are directly relevant to their research. Computational models complement hypothesis testing in advancing our understanding of the cognitive functions underlying complex tasks such as adaptive speech perception [@kriegeskorte2018], and they are now more accessible than ever. For instance, formal models of normalization have been available for decades and some of them are no more complex than standard data analysis [e.g., linear regression for C-CuRE, @mcmurray-jongman2011]. Similarly, fully specified distributional learning models have been available for at least two decades [e.g., mixtures of Gaussians, @mcmurray2007; @toscano-mcmurray2010; @feldman2013; exemplar models, @johnson1997; @apfelbaum-mcmurray2015; ideal adaptors, @kleinschmidt-jaeger2015]. Both classes of models have some variants that are now freely available as R library [e.g., for normalization: \texttt{phonTools}, @barreda2015; for changes in category representations: \texttt{beliefupdatr}, @kleinschmidt-jaeger2015; and its extension \texttt{MVBeliefUpdatr}, @R-MVBeliefUpdatr]. ASP integrates models of all three change mechanisms in R (provided at <https://osf.io/q7gjp/>). Specifically, the three change models of ASP spell out competing hypotheses about how exposure (1b) is integrated with prior experience (1a). Then the categorization model (updated based on the exposure input) determines how these changes are expected to affect listeners’ responses to test items (2) (see Figure @ref(fig:overview-change)).

These resources can help researchers derive predictions prior to conducting an experiment, based on all the relevant acoustic-phonetic properties of the planned exposure and test stimuli (see Recommendations 2 and 3), and to compare the fit of different change models (or different combinations of them) to the results once data collection has been completed (Recommendations 4 and 5). The use of ASP and similar models can also reduce the file drawer problem, by helping researchers understand seemingly unexpected null results [for demonstration, see tan2021] and reduce the need for \*post-hoc\* reasoning [e.g., the postulation of different mechanisms, where simpler explanations in terms of ceiling effects, @floccia2006; or predicted null results @zheng-samuel2020 might suffice].

### ### Recommendation 2: Analyses that link the acoustic-phonetic properties of stimuli and participants’ responses

In our case studies, we followed the approaches that continue to be employed in the majority of experiments on adaptive speech perception. In Case Study 1, we analyzed changes in categorization responses at six different test items. Following the majority of research on perceptual recalibration [for exceptions, see @drouin2016; @saltzman-myers2021], we did not further relate these changes to the \_acoustic or phonetic properties\_ of the exposure or test stimuli [for related discussion, see also @clayards2018; @theodore2021]. In Case Study 2, we analyzed changes in *accuracy*. Like some previous studies [e.g., @xie2016jep; @zheng-samuel2020], we showed these changes separately for the two categories investigated in the case study. Other studies on accent adaptation further simplify the dependent measure and analyze only overall improvements in accuracy [@bradlow-bent2008; @sidaras2009; @tzeng2016] or processing speed [@clarke-garrett2004].

Although in common use, these types of analyses over aggregate data constitute a missed opportunity. They discard stimulus-level data that could otherwise provide valuable information on the nature of the mechanisms underlying adaptive speech perception. This is a general, and well-known, problem: many competing hypotheses might provide plausible explanations for coarse-grained aggregate data (like overall or category-specific accuracies). On the other hand, models that make the same predictions for aggregate data might be distinguishable when compared against more fine-grained data. For experiments on adaptive speech perception, this means that researchers stand to benefit from analyzing changes in the \_*categorization function\_* ---i.e., the mapping from acoustic-phonetic properties of the test stimuli to participants’ responses---in response to recent exposure.

We illustrate this point in Figure \@ref(fig:show-model-categorization-3D-plots-similar-accuracy), which shows the predicted categorization functions after L1- and L2-accented exposure for each of three change models for parameter settings that predict highly similar overall accuracy. While predicted categorization functions are quite similar across the three change models for the L1-accented exposure condition (second row), they vary more widely for the L2-accented exposure condition (third row). The bottom row of Figure \@ref(fig:show-model-categorization-3D-plots-similar-accuracy) further suggests that the three change models differ qualitatively in what \_type\_ of change in the categorization function they predict after L2-accented exposure. Whereas the effect of L2-accented, compared to L1-accented, exposure can be complex for the representational model (bottom left panel), it appears more constrained for changes in decision-making (bottom center panel) and normalization (bottom right panel). These differences reflect the computational limitations of each change models, and are even more pronounced when we consider the best-performing parameters for each type of change model (Figure \@ref(fig:show-model-categorization-3D-plots-best-performing)).

Arrow

Description automatically generated

Shape

Description automatically generated with medium confidence

Critically, differences like those in Figures \@ref(fig:show-model-categorization-3D-plots-similar-accuracy) and \@ref(fig:show-model-categorization-3D-plots-best-performing) will only be apparent when the data is analyzed at appropriately fine-grained detail---i.e., by assessing changes in the mapping from acoustic-phonetic properties of stimuli to participants’ responses. Fitting ASP and similar models to the results of experiments is one way to conduct such analyses. Previous work has done so for competing normalization models [see, e.g., @mcmurray-jongman2011; @apfelbaum-mcmurray2015; @persson-jaeger2022; @richter2017; @xie2021] and competing parameterization of representational change models [e.g., @kleinschmidt-jaeger2016; @kleinschmidt2020; @tan-jaeger2023]. Future work can employ ASP to compare the fit of all three types of change models---including combinations of them---to the results of experiments, allowing researchers to investigate what factors contribute to the relative engagement of different change mechanisms.

Alternatively, standard approaches to data analysis can be used to investigate changes in the mapping from acoustic-phonetic properties to participants’ responses [for an excellent review, see @schertz-clare2020]. This includes regression models as long as they employ appropriate linking functions [e.g., logistic or multinomial regression for categorization, @jaeger2008; @winter-wieling2016]. This approach is now increasingly common, taking advantage of stimulus-level variability within and across conditions to test hypotheses [e.g., @clayards2018; @idemaru-holt2020; @schertz2015]. Regression analyses can further be expanded into psychometric models [@wichmann-hill2001] by adding lapse rates and response biases [e.g., @clayards2008; @kleinschmidt2020], and can be fit in standard statistics software [e.g., \texttt{brms}, @burkner2017]. Optionally, such regression analyses can be enriched with predictive ASP simulations of the type we have employed in our case studies. Such simulations are computationally less demanding than \_fitting\_ ASP models to perception experiments, and can help guide the interpretation of results [for examples of approaches that mix predictive simulation with standard data analysis, see @bejjanki2011; @clayards2008; @hitczenko-feldman2016; @tan2021; @theodore-monto2019; @xie2021; see also discussion in @bent-baese-berk2021].

To fully take advantage of the types of analyses we have discussed here, it will, however, be important to plan experiments with these analyses in mind. This leads to our next recommendation.

### ### Recommendation 3: Dense and targeted sampling of the stimulus space

Our third recommendation is to obtain data that more strongly characterize incremental changes in categorization functions that occur with exposure. Simply put, the more data an experiment yields on how different types of exposure change the shape and location of the categorization function, the easier it is to determine the relative engagement of the different change mechanisms. Each change model is subject to different computational limitations, and these limitations affect what types of changes in categorization functions the different change models can account for (see, e.g., Figures \@ref(fig:show-model-categorization-3D-plots-similar-accuracy) and \@ref(fig:show-model-categorization-3D-plots-best-performing) in the previous section). This idea is further visually illustrated in Figure \@ref(fig:constraints-on-change-models). In this section, we discuss how researchers can design their experiments to be more informative about changes in categorization functions.

Diagram

Description automatically generated

Several design properties under researchers’ control can make an experiment more informative about the constraints on changes in listeners’ categorization functions. As described under Recommendation 1, the changes in categorization functions predicted by different change models depend on both the exposure and test stimuli. For exposure, both the \_stimulus location\_ in the acoustic-phonetic space---relative to listeners’ prior expectations---and the \_number of exposure stimuli\_ affect the predictions of change models. The location of exposure stimuli determines the \_type\_ of change that is predicted to occur (e.g., shifts vs. changes in the slope of categorization function; changes in the relative weighting of different cues). The amount of exposure determines \_how much\_ change of that type is predicted.

Experimenters can take advantage of both aspects. For example, experiments can employ more than one exposure condition. For instance, @kleinschmidt-jaeger2016cogsci exposed different groups of participants to six different exposure conditions with distinct VOT distributions for /b/ and /p/. The data were then used to contrast different parameterizations of the same change model [for additional analyses and insightful discussion, see also @kleinschmidt2020]. While some studies have employed multiple exposure conditions in perceptual recalibration and similar paradigms [e.g., @babel2019; @sumner2011], this approach remains under-utilized, and computational analyses of the results lacking.

Similarly, incremental testing *within* subjects can increase the ability to contrast the predictions of different change models. Even when all three change models predict the same *outcome* of adaptation, they can differ in the *trajectory* of changes they predict [for some pioneering behavioral work, see e.g., @bertelson2003; @bonte2017; @vroomen2007, and for computational models, see @kleinschmidt-jaeger2012].

In addition to considerations about exposure, researchers can also benefit from selecting \_test\_ stimuli to be maximally informative about changes listeners’ categorization functions. Paralleling our recommendations for exposure stimuli, this includes considerations about both the location and the number of test stimuli. For example, to detect differences between the different categorization functions, an experiment needs to employ test stimuli that are sufficiently distributed across the acoustic-phonetic space. This is worth emphasizing since it is typically \_not\_ the case in experiments on perceptual recalibration and accent adaptation. For perceptual recalibration, experimenters tend to target stimuli that are expected to be maximally ambiguous. This makes sense if the goal is to detect the \_existence\_ of a shift in the categorization function. But it is far from optimal when the goal is to fully understand changes in the shape of the categorization function (Recommendation 2).

Similarly, test phases in experiments on accent adaptation tend to densely sample the space around the accented category means---a side effect of using naturally accented stimuli. But any change model that can predict general improvements in accuracy will also tend to correctly predict the categorization for those stimuli, making them relatively uninformative about the types of changes that result from exposure. Future experiments can advance our understanding of exposure-driven changes in speech perception through more targeted sampling of the acoustic-phonetic space during test. Researchers should derive model predictions and sample test stimuli from the regions where model predictions would be maximally distinct from one another [for initial efforts along these lines, see @burchill-jaeger2022]. As shown in Figure \@ref(fig:show-model-categorization-3D-plots), some of the largest predicted differences can occur in parts of the acoustic-phonetic space that are neither close to the category means, nor on the line between them.^[The optimal choice of stimuli depends on the specific goals of the experiment. For example, stimuli locations that are optimal for estimating the categorization function after one exposure condition, likely are not optimal for estimating the categorization function in another exposure condition, and neither choice might be optimal if the goal is to detect \_differences\_ between the two conditions or \_differences\_ between the different change models. Predictive (power) simulations can help determine the optimal stimulus locations and repetitions, for any (combination) of these goals (Recommendation 5).]

While targeted sampling is most easily achieved for paradigms that employ (re)synthesized or otherwise phonetically manipulated stimuli [for examples, see @bejjanki2011; @burchill-jaeger2022; @idemaru-holt2020], it is also possible in combination with naturally accented stimuli. For example, @chodroff-wilson2020 took advantage of naturally occurring variability, and created different exposure conditions by selecting different subsets of natural stimuli. This approach strikes a particularly intriguing balance between ecological validity and experimental control that deserves further attention in future work.

### ### Recommendation 4: Advanced standards of data annotation and sharing

As we mentioned under Recommendation 1, predicting adaptive changes of recognition requires estimates of 1a) the prior state of listeners; 1b) the acoustic/phonetic properties as well as category labels of the exposure stimuli; and 2) the acoustic/phonetic properties of the test stimuli. Estimates 1a) can be obtained from sufficiently large phonetically annotated databases that capture the type of speech input a typical participant in the experiment is likely to have received throughout their life. Fortunately, large and phonetically annotated databases can offer good estimates of the speech input of listeners receive. Such databases are now available for an increasing number of phonetic contrasts and languages [e.g., @chodroff-wilson2018; @clopper-pisoni2006; @hillenbrand1995; @newman2001; @theodore2009; @xie-jaeger2020, among many others].

One caveat is that many of these databases either only contain a subset of the speech varieties that an average listener of a language has plausibly been exposed to or contain very little data for each variety. In particular, databases that contain both a large number of talkers and a large number of tokens per talker continue to be the exception. Researchers thus need to carefully consider these implications when employing databases. In some cases, researchers might find that the best way forward is to collect phonetically annotated data that meets the specific requirements for their study. @xie2021cognition, for instance, collected ~3000 production tokens of prosodic categories and used them to model listeners’ responses in a perception experiment.

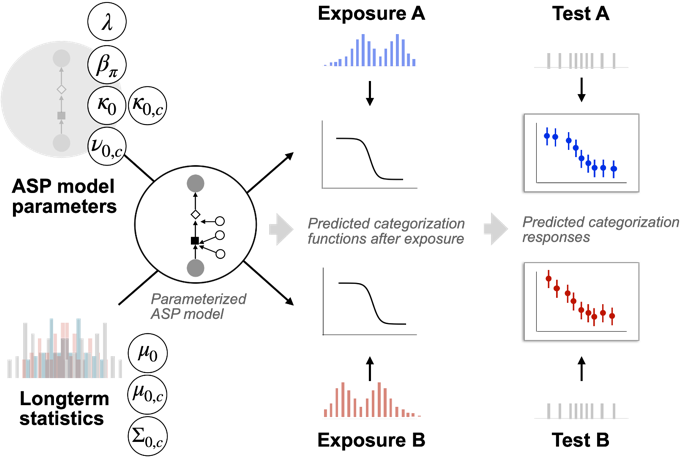
An alternative option is to eschew phonetically annotated data in favor of computational methods that work from the raw signal or some automatically obtained transformation of the raw data [e.g., Mel-Frequency Cepstral Coefficients, (MFCCs) @Mermelstein1976]. Such models have been developed for automatic speech recognition and have been employed to model human language acquisition [@dupoux2018; @feldman2013; @toscano-mcmurray2010; @vallabha2007]. More recently, they have also been applied to address questions about the effects of recent exposure [@richter2017]. The ASP framework can, at least in theory, be combined with such models.

Estimates for 1b) and 2) require researchers to phonetically annotate the stimuli in their perception experiments. Such annotations entail a sizable but manageable effort: perception experiments typically employ a small number of speech stimuli that are repeated across participants. A typical perceptual recalibration experiment would require the annotation of fewer than 100 isolated word recordings. A large study on accent adaptation like Bradlow and Bent’s (2008) Experiment 2 would require the annotation of about 1000 sentences. (As a reference point, studies on phonetic production regularly annotate datasets many times larger.) These efforts will be supported by clear standards for reproducibility and software developments that aid phonetic annotation and data sharing [e.g., @cassidy-schmidt2017; @picoral2021; @roettger2019; @winkelmann2017]. If annotations are not reported and shared—and ideally even if they are—then all audio recordings should be shared in an open and accessible way (e.g., OSF). This will require perception researchers to use human subject protocols that gives them the consent to distribute stimulus recordings for the purpose of scientific inquiry. In other words, perception and production researchers should make concerted efforts to link the listener’s knowledge of productions and inferences they make during perception.

### ### Recommendation 5: Predictive power analyses prior to conducting experiments

Like other computational frameworks, ASP can be used to estimate expected effect sizes and statistical power *before* conducting an experiment. Unlike commonly used power estimates, which assume effect sizes from other experiments [e.g., “moderate” effects, @zheng-samuel2020], predictive power-analyses of the type we envision are based on effect sizes that are derived from the acoustic properties of the very exposure and test stimuli used in the given experiment. This makes power simulations considerably more informative. Furthermore, ASP enables predictive power-simulations under *combinations* of multiple mechanisms. In the Case Studies, we considered only one mechanism at a time by “switching off" the other mechanisms. By the same logic, one can switch one or no mechanism off to simulate scenarios where more than one mechanism are engaged.

There are, however, challenges that will have to be met so that ASP can be used for predictive power simulations. Figure \@ref(fig:predictive-power-simulations) illustrates these challenges. For a specific set of exposure and test stimuli, and specific parameter settings for all change models, ASP predicts an expected categorization function. This is what we showed in, for example, in Figure @ref(fig:show-model-categorization-3D-plots) and the result figures for Case Study 1 [see also, e.g., @tan2021; @theodore-monto2019; @xie2021 for similar approaches applied to only representational change models]. This categorization function can be used to derive a predictive power estimate by repeatedly sampling responses for the planned test stimuli and the planned number of participants. This can be done for any individual exposure conditions, for combinations of exposure conditions, and/or the predicted differences between exposure conditions.



However, the parameterization of the different models is typically *not* known to researchers—if they were, there would be no need for the present article. One important direction for future research will thus be to better characterize the functional consequences of each model’s computational limitations. For example, in the SI (\@ref(sec:consequences-of-lambda), we show that changes in decision-making can only explain additive effects of exposure on the log-odds of categorization responses (for lapse rates of 0 or 1) or effects that resemble a step-function (for all other lapse rates). Further study of this and similar constraints will shed light on what range of adaptive behaviors each model can predict [for discussions of global, qualitative model comparisons, see @pitt2006; and @apfelbaum-mcmurray2015]. Another related approach is to derive predictions and calculate statistical power while marginalizing (i.e., averaging) over a plausible \_distribution\_ of ASP parameters for the models they wish to contrast. In Bayesian terminology, this is achieved by defining “priors” over the ASP parameters. Initially, when relatively little is known about the relevant distribution of parameters, very weak priors are recommended that will consider a wide range of parameters as probable. As additional studies become available, this will further inform the range of plausible priors. Even now though, the plausible range of parameters is constrained by the fact \_that\_ listeners can adapt within a certain number of observations. This constrains the joint distribution of ASP parameters since any plausible parameterization needs to give at least one of the change models sufficient flexibility.

If researchers further wish to know which exposure or test stimuli would increase the predicted difference between hypotheses, that further requires comparison of different stimulus selections. Comparable approaches exist for simpler computational problems, and have been used extensively in psychometric research [incl. online stimulus selection *during* the experiment, depending on subject-specific performance, e.g., @vul2011; @prins2013; @wichmann-hill2001]. We believe this is an area of research that holds substantial potential.

### ### Recommendation 6: Integration with neural models of speech perception

ASP as we put forward here is a model that makes behavioral predictions. Formal models that describe how computational demands of adaptive speech are met by brain activity remain lacking [see @guediche2014 for a review]. In the introduction, we outlined that models like ASP can support neuroimaging research by making more concrete (and thus more testable) the types of computations that are hypothesized to take place in different brain regions or networks. At the same time, future neuroimaging research can further constrain the hypothesis space, and inform frameworks like ASP. Our final recommendation is to further integrate ASP with neural models of speech perception.

As in behavioral studies, existing neuroimaging results so far have not always singled out neural networks underlying behavioral changes in response to recent exposure. Adaptive changes in speech perception have been found to involve several different brain regions. These include cerebellum [@guediche2015], early auditory regions [anterior planum temporale, @kilianhutten2011; @bonte2017] and regions responsible for representing phonemes and syllables [e.g., posterior superior temporal gyri/superior temporal sulcus, @bonte2017; @myers-mesite2014; @ullas2020], as well as regions for talker recognition [right temporal regions, @luthra2020a]. The left parietal lobe and the insula, which are implicated in perceptual decision-making [e.g., @dacremont2013; @keuken2014], have also been shown to exhibit distinct activation for different exposure conditions.

While certain brain regions and/or networks have been identified to respond differently to different exposure conditions (e.g., familiar vs. unfamiliar talker accents), these regions often play multiple roles in cognitive processing. The exact interpretation of the findings still depends on the researchers’ hypothesis of the underlying mechanism. For instance, the activation of left parietal lobe might reflect its general role in perceptual decision-making [e.g., @dacremont2013; @keuken2014], or it could be due to a more specific role in phonological processing [e.g., processing abstract category information, @guediche2014]. Similarly, changes in inferior frontal gyri activation in response to accented speech has been interpreted as reflecting greater computational demand [@yi2014neural] or decision-related phonetic categorization of ambiguous stimuli [@myers-mesite2014]. This ambiguity is compounded by univariate analysis methods which test an increase or a decrease of neural activity while largely leaving open the information content represented in the distinct brain regions. Clearer links between acoustic-phonetic cues and expected changes of recognition, as facilitated by ASP, can help resolve this ambiguity. The five recommendations above apply to stimuli and tasks used in future neuro-imaging studies.

Recent research has begun to apply multivariate analyses that can be effectively combined with the approach we have described above. Multivariate analyses can be used to identify the encoding of perceptual experiences across a distinct array of experimental conditions, which helps to distinguish between cognitive models that make different predictions regarding the type of information encoded by certain brain regions and networks [e.g., @Blank2016]. Multivariate analyses are also more sensitive in detecting fine-grained patterns \*within\* regions [e.g., @bonte2017; @luthra2020a] and therefore hold the promise of shedding greater light on brain information processing. For instance, while recent electrocorticography studies have provided some direct evidence for prelinguistic cue-level normalization within the middle and posterior STG areas [@sjerps2021; @tang2018], it remains an open question how normalization supports the kind of adaptive changes as observed in perceptual recalibration or accent adaptation. For instance, fMRI representational similarity analysis (RSA) allows researchers to take advantage of specific similarity matrix for a set of stimuli, rather than just the overall activation level for each condition. If test stimuli can be selected in a way so that the similarity matrix can be analyzed to differentiate between accounts, then RSA can be an effective way to generate novel insights on the computations performed by the auditory cortex. The model-guided predictions and experimental design as described in Recommendations 1-5 will provide a roadmap to achieve such stimuli selection.

Finally, the synergy between behavioral, computational, and neural studies may uncover new knowledge about the relative engagement of the three mechanisms over time. @myers-mesite2014 found that boundary shifts in a perceptual recalibration study were initially driven by changes in the brain regions responsible for decision making in the absence of changes in acoustic-phonetic representations (as indexed by STG activation). However, over the course of 20 critical exposure trials, evidence of retuned perceptual sensitivities slowly emerged. The authors concluded that this indicates a “transfer of decision-related or lexical-level information to more bottom-up or perceptual processes in the temporal lobes”. To extend this finding, future research should track brain responses over a more extended period of exposure. ASP models’ abilities to predict incremental changes as a function of exposure statistics, as well as to do so under combinatory effects of the mechanisms, make them a tool well-suited to aid the design of such new experiments.

Another promising avenue is to pair temporally-sensitive techniques with imaging methods with good spatial resolution. For instance, studies using a combination of EEG and MEG have teased apart the role of low-level signal clarity or high-level prior knowledge in regulating perceptual learning of degraded speech—despite that they recruit a common region in STG—by capitalizing the fact that they operate over different timescales [@sohoglu-davis2016; @sohoglu-davis2020]. Temporally-sensitive methods are particularly helpful in distinguishing between distinct mechanisms that may underlie adaptive speech perception. Using P200 as an index for acoustic-phonetic processing, Romero-Rivas et al., [2015](https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12438#lang12438-bib-0049) found that the extraction of spectral information and other acoustic features was more difficult for foreign-accented speech than for native speech. Critically, this difference in acoustic-phonetic mapping did not attenuate within a single exposure session, compared to faster improvements in lexical-semantic processing, as indexed by N400. With the caveat that the examined ERP components reflect the underlying distinction between acoustic-phonetic vs. post-lexical processes, these results are compatible with the possibility that neural changes associated with decision making criteria—as aided by lexical information—occur faster than those responsible for acoustic-phonetic mapping.

## Conclusion

Through formalizing and simulating three distinct mechanisms of adaptive speech perception, we have explored how these mechanisms may be empirically distinguished from one another. A general conclusion we draw from the current investigation is that, for processes as complex as human speech perception, it is hard to make strong inference without an aid of analytical tools like the one we put forward here.