Adaptive Speech Perception: Empirical Indeterminacy and a Path Forward

Shawn Cummings[1] **T. Florian Jaeger**[2] **Chigusa Kurumada**[2] **Xin Xie**[1]
[1]Department of Language Science, University of California, Irvine
[2]Department of Brain and Cognitive Sciences, University of Rochester
{scummin2,xxie14}@uci.edu
{fjaeger,ckuruma2}@ur.rochester.edu

Abstract

Human listeners rapidly adapt to unfamiliar talkers, but the underlying computational mechanisms remain contested. Three candidate hypotheses—pre-linguistic normalization, changes in phonetic category representations, and changing decision biases—have largely been pursued in separation, using subfield-specific paradigms. Researchers working in these paradigms often assume that adaptivity observed in their particular paradigm can only be explained by one of the three mechanisms. We test this assumption for one of the most popular experimental paradigms (lexically-guided perceptual learning or LGPL) using a unified computational framework (ASP). We apply ASP to the largest existing LGPL data: 89,600 categorization responses from over 1000 listeners after lexically-guided exposure to 32 different stimulus sets. Despite the unprecedented scale of these data, we find that behavioral data are equally compatible with all three candidate mechanisms. We discuss how model-guided stimulus selection can increase the diagnosticity of future LGPL experiments. Our simulation code can easily be adapted to other experimental paradigms.

1 Introduction

One of the core computational challenges facing the human brain is inducing stable, generalizable structure from noisy and variable perceptual signals. Speech perception is a particularly vivid illustration of this challenge. Even a simple phonological contrast like /s/ (e.g., *sip*) vs. /ʃ/ (e.g., *ship*) is realized across a vast acoustic space: a talker's /s/ may be acoustically more similar to another talker's /ʃ/ than to their own prior productions, due to variation in vocal tract anatomy, speaking rate, dialect and social identity, etc. Yet listeners perceive speech with remarkable consistency across talkers (for review, Johnson and Sjerps, 2021; Weatherholtz and Jaeger, 2016).

Decades of experimental work have established that this robustness is not merely passive tolerance of variability, but reflects active adaptivity. Listeners recalibrate their perception based on the acoustics of recent inputs, so that the same acoustic signal might be recognized as an instance of a different linguistic category (e.g., phonemes, syllables, or words). Unlike most commonly available automatic speech recognition systems, this adaptivity can be rapid—sometimes occurring within a single utterance—and yet can persist for days. This adaptivity is now considered a fundamental property of human speech perception, documented across phonetic contrasts, languages, and listener populations (for review, Bent and Baese-Berk, 2021; Johnson and Sjerps, 2021).

The central question we seek to address here concerns the computational mechanisms responsible for this adaptivity. It remains fundamentally unclear what computations support adaptive speech perception, how their consequences interact to jointly shape perception, and what determines the relative engagement of different mechanisms across tasks, stimuli, and contexts. Here, we contribute to recent demonstrations that **less is known about the mechanisms underlying speech perception than commonly assumed**, and begin to demonstrate how **minor changes to the stimulus design of a highly popular experimental paradigm could substantially enhance its diagnosticity** about the underlying mechanisms. We pursue these goals within a recently proposed minimal computational framework for adaptive speech perception (ASP), described below.

ASP (Xie et al., 2023) incorporates competing proposals about the mechanisms underlying adaptive speech perception, allowing us to simulate and analyze changes in perception under different hypotheses about these mechanisms. Through the computational simulations we present—fit against, and informed by, human behavioral data—we hope

to re-emphasize an important insight that—while repeatedly and cogently made in previous work (Newell, 1973; Yarkoni and Westfall, 2017; Guest and Martin, 2021)—remains underappreciated among many experimenters: informally stated hypotheses can productively guide research in a field's early stages; however, there comes a time in the development of a field when progress requires sufficiently specific computational models of the remaining plausible accounts to effectively contrast them across paradigms and phenomena. This is particularly true when—as is typically the case for any non-trivial problem in the language sciences—multiple mechanisms contribute to observable behavior, so that research needs to go beyond qualitative existence proofs.

1.1 The problem: unknown uncertainties

Three broad classes of proposals have been advanced in the literature to account for adaptivity in speech perception (for review, Xie et al., 2023). The first attributes changes to **pre-linguistic signal normalization**: low-level auditory processes that track the overall statistical properties of the input—such as the mean of a cue distribution—and adjust perception accordingly, without reference to linguistic categories. The second attributes adaptation to **changes in linguistic category representations**: listeners update their implicit knowledge of the cue distributions associated with phonological categories (e.g., shifting or expanding the distribution for /s/ along its phonetic cues). The third attributes adaptation to **changes in decision-making**: listeners adjust response biases or decision criteria rather than the perceptual or representational processes that precede them. These three classes of proposals assume fundamentally different cognitive architectures, with divergent implications for theories of speech perception, language acquisition, and the malleability of neural representations of phonological categories.

Despite decades of research, principled comparison of these mechanisms are largely lacking. Theoretical proposals have often remained underspecified—formulated verbally rather than as computational models—and predictions from competing mechanisms are rarely tested against the same data (for review, Tan and Jaeger, 2025; Xie et al., 2023). As a result, different research communities continue to operate in parallel, each citing evidence from different paradigms in support of their preferred account. For instance, researchers who conceive of their experimental paradigm as tapping into normalization processes might point to findings that non-speech acoustic contexts can shift the perception of subsequent speech inputs (Huang and Holt, 2009; Laing et al., 2012)—a result that can be taken to challenge representational accounts. Researchers who attribute changes in listeners' behavior to changes in phonological category representations, on the other hand, might cite evidence that adaptation can be affected by category labels: the same input inferred to be an /s/ can have different consequences on subsequent perception than if that input is inferred to be an /f/ (Norris et al., 2003; Jesse, 2021)—a finding argued to be incompatible with normalization. Similarly, researchers favoring decision-making as the source of behavioral changes might point to neuroimaging evidence implicating prefrontal and parietal regions (Myers and Mesite, 2014)—typically taken to reflect post-perceptual processes rather than low-level signal transformation or representational change.

It is, however, inherently risky to argue based on qualitative findings and theoretical intuitions in the absence of a common theoretical vocabulary. For instance, the argument that signal normalization cannot account for the findings in Norris et al. (2003) rests on the assumption that normalization does not consider the category inferred to underlie the current input. While this assumption applies to some influential normalization accounts (e.g., C-CuRE, McMurray and Jongman 2011), alternative accounts that normalize phonetic cues based on inferred category labels have long existed (Nearey and Assmann, 2007), and have been shown to better explain speech perception (Barreda and Jaeger, 2025). This illustrates that even strong intuitions about the interpretation of existing results often rely on assumptions not shared by all researchers. An alternative approach—which we pursue here—is to develop models that deliver clear, quantitative predictions that can be evaluated using standard goodness-of-fit metrics for model comparison. This approach, at a minimium, makes all relevant assumptions explicit.

1.2 A path forward: a computational framework for adaptive speech perception (ASP)

As a step towards this goal, Xie et al. (2023) introduced a computational framework for adaptive speech perception (ASP). As illustrated in Figure 1, ASP formally specifies a categorization model

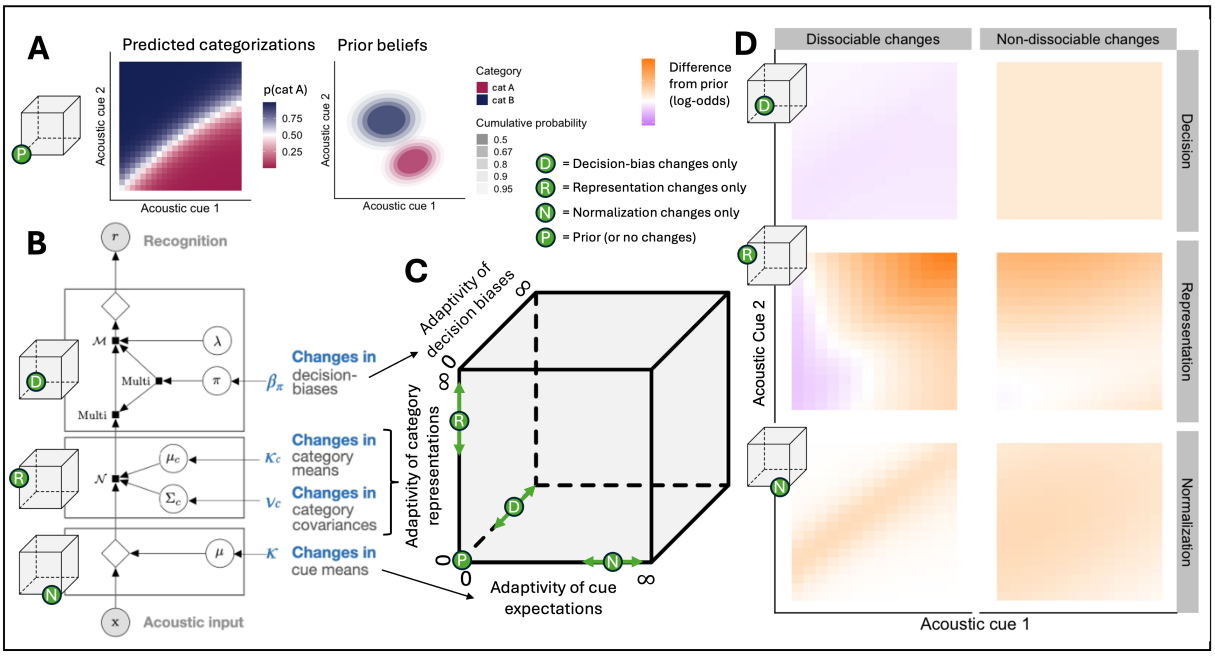

Figure 1: The ASP framework, parameterizations, and potential predictions. **Panel A:** One potential set of prior beliefs about cue-category mappings in a two-dimensional acoustic cue space, and resultant predictions from the categorization model. **Panel B:** graphic illustration of the categorization model and parameters over which the change models operate. **Panel C:** visualization of the multidimensional parameter space of ASP, with marked annotations for cases where only a single model contributes to adaptation and the case where no models exhibit any flexibility, resulting in no changes from prior beliefs. **Panel D:** Predicted differences between each model's prediction and a model without any adaptation (= predicted adaptive changes of categorization responses). Columns depict two possible scenarios in which these predictions diverge (left) or converge (right), depending on the acoustic characteristics of the exposure input, where convergence contributes to empirical indeterminacy.

that maps acoustic inputs to linguistic categories via normalization (linear transformations of the acoustic inputs), category representations (category likelihoods), and decision-making (category priors that represent decision and response biases). ASP also specifies change models for each of these three components to describe how normalization, category representations, and decision biases might be adapted in response to recent exposure.

ASP leaves open whether humans draw on only a single mechanism when adapting to unfamiliar speech. All three change models could operate at the same time. Figure 1 (panel C) illustrates this point by representing parameterization in ASP as a 3D space, where each axis determines a hypothetical listener's flexibility or willingness to change according to a different model. The 'true' characterization of a given listener could lie anywhere in this three dimensional space. Before the field can address this question productively, however, we need to address a simpler question: whether existing data can even distinguish between scenarios in which only one change model is operating. These scenarios represent how theories are often evoked in the field: in isolation and at the exclusion of one another. These scenarios also provide the strongest potential for measurable differences in predictions. If existing data cannot even distinguish between the extreme scenarios in which only one model is active, those data certainly cannot be used to determine what mixture of mechanisms listeners might draw on.

Given the separate computational architectures of each model, one might assume them to beget very different predictions (Figure 1D, left column). However, initial simulations that applied ASP to synthetic data produced a striking result: the signature findings from two highly popular and frequently cited experimental paradigms in the field (lexically-guided perceptual learning, henceforth LGPL, and nonnative accent adaptation) are qualitatively compatible with all three mechanisms (Xie et al., 2023) . This is illustrated in Figure 1D (right column).

This raises two questions, the answers to which will determine what a feasible path towards more diagnostic experimental paradigms will look like. First, is the empirical indeterminacy observed by

Xie and colleagues an artifact of the specific, simplified stimulus configurations used in the initial simulations? Perhaps the mechanisms generate divergent, distinguishable predictions for actual experiments with natural speech stimuli—especially once one considers more than a single experiment at a time. If so, popular paradigms might *in principle* risk being uninformative about the three mechanisms, while in practice providing sufficient signal about the (mixture of) mechanisms that underlie listeners' responses. Second, is the empirical indeterminacy an inherent artifact of the low-dimensional behavioral measures assumed in Xie and colleagues' simulations—specifically, the two-alternative forced choice (2AFC) task? 2AFC tasks continue to dominate the field, in part because they are often simple and efficient for both participants and experimenters. They have, however, long been critiqued for being less informative than other tasks (see also recent debate about 2AFC vs. visual analogue scale, Apfelbaum et al., 2022). It is thus possible that no amount of tweaking the stimulus selection will make 2AFC experiments diagnostic of mechanisms underlying adaptive speech perception.

Here, we focus on the first question and demonstrate that findings from one of the most popular paradigms in the field (LGPL, Norris et al., 2003; Kraljic and Samuel, 2005) do not, in fact, provide diagnostic evidence for one mechanism over the others—contrary to many researchers' intuitions. This matters beyond the immediate paradigm: LGPL findings have been used as a foundation for theoretical claims in adjacent subfields. For instance it has been claimed that older adults have decreased flexibility in the adjustment of phoneme categories, based on reduced magnitudes of boundary shifts in LGPL studies (Scharenborg and Janse 2013). Such claims presuppose that LGPL results can be attributed to a specific mechanism. We stress that our findings do not indicate that LGPL is inherently uninformative about mechanisms; as we illustrate in the Discussion, relatively modest changes to stimulus selection can substantially improve its diagnosticity. Moreover, the indeterminacy we identify is unlikely to be unique to LGPL: any paradigm whose theoretical interpretations rest on informally stated hypotheses rather than formalized computational models risks the same problem. In the Discussion, we turn to the second question and illustrate how ASP-guided stimulus selection practices hold promise to remedy the indeterminacy.

2 Data

We use the largest available database of LGPL experiments (Cummings, 2025): a total of 89,600 post-exposure categorization responses from 1,280 listeners across 16 LGPL experiments—each with two exposure conditions—that only differed in their exposure and test stimuli.

In LGPL experiments, listeners are exposed to speech in which a perceptually ambiguous sound (e.g., midway between /s/ and /ʃ/) is embedded in words that disambiguate its intended category (e.g., *per/?/onal* biases interpretation toward /s/). Their subsequent categorization of that sound on a test continuum shifts in the direction consistent with the lexical label, taken as a measure of adaptation to the talker's pronunciation. The design, procedure, and stimuli of the 16 experiments in Cummings (2025) closely follow conventions of the field, including the use of a lexical decision task during exposure and a 2AFC task during test ("Did the talker say *asi* or *ashi*?"). A schematic for this paradigm can be found in Appendix A1.

2.1 Stimuli

The **exposure** stimuli included auditory recordings of 100 English words and 100 nonwords (e.g., *baliber*) produced by 16 talkers (8 male and 8 female) of American English. 20% of the words contained an acoustically ambiguous sound, created by digitally mixing energy (via waveform averaging) from natural /ʃ/ and /s/ productions of each of the same 16 talkers. The **test** stimuli consisted of a 7-step continuum ranging from /aʃi/ to /asi/. Both the original ('typical') and the manipulated ('ambiguous') stimuli were annotated for two acoustic cues to fricative identity (the first and third peak in the amplitude over the frequency spectrum, P1 and P3; for more details, see Appendix A2).

2.2 Human categorization responses following exposure

40 native speakers of American English participated in each of the two conditions of the 16 experiments, for a total of 1,280 participants. Following exposure, each listener went through 10 cycles of 7-step test continua, with item presentation randomized within each cycle. This yielded a total of 89,600 post-exposure categorization responses (16 talkers * 2 bias conditions * 40 listeners * 7 steps along the test continuum * 10 cycles).

Appendix A3 reports how we analyzed the data from Cummings (2025). We found significant between-condition differences in categorizations of the same test items for 13 of 16 talkers, replicating the classic boundary shifts in LGPL. At the same time, the unprecedented scale also revealed considerable variability in the magnitude of effect sizes across experiments: how much listeners changed their categorization responses following exposure varied substantially between the 16 LGPL experiments.

Additionally reported in Cummings (2025) are a set of complementary experiments which include only the 2AFC test phase without biasing exposure (henceforth *test-only data* to distinguish it from the LGPL data). We use these independent test-only data to validate the categorization model described in Section 3.1.1.

3 Empirical indeterminacy, even at scale

We use the data from Cummings (2025) to test whether the exposure-elicited changes predicted by each of the three ASP mechanisms can be separated from each other, by comparing the fit of model predictions to human categorization responses.

3.1 Methods

3.1.1 Parameterizing the categorization model

Even before experimentally induced biases, listeners have expectations built-up over their lifetime for how talker-contingent (e.g. vocal tract length) and phonetic/co-articulatory factors affect fricative acoustics. To capture this, we used C-CuRE normalization (McMurray and Jongman, 2011) to remove talker's means from the acoustic cues to fricative identity (P1 and P3). Crucially, this step is distinct from the normalization change model evaluated below, which captures trial-by-trial adjustments to expected cue values over the course of the experiment.[1]

To estimate listeners' prior beliefs about category representations, we estimated $\mu_{c,0}$ and $\Sigma_{c,0}$ for $c \in \{/s/, /ʃ/\}$ over the normalized cues of the 40 clear, unmanipulated tokens of /s/ and /ʃ/ from each of the 16 talkers used in the LGPL studies (n = 1,280 pairs of P1-P3 values). Finally, we fit perceptual noise (Σ_ϵ) and lapse rates (λ) to the test-only data, as these parameters govern the categorization model independent of any exposure-induced changes. This reduces the number of parameters we had to fit to the LGPL data.

3.1.2 Parameterizing the change models

We assume that the parameterization of change models is identical across all 32 exposure conditions (16 talkers * 2 bias conditions). The parameters of the change models (κ_0 for normalization, $\kappa_{c,0}$ and $\nu_{c,0}$ for changes in category representations, or β_π for changes in decision-biases) were therefore fit to post-exposure categorization responses across all exposure conditions together.

We used five-fold cross-validated maximum likelihood to find the best-fitting parameters for each change model. Specifically, the data were folded between participants, so that the per-observation log-likelihoods we report below quantify the change model's ability to successfully generalize across participants. Instead of directly fitting change models against listeners' responses, we fit them against an estimate of listeners' response preferences *immediately following exposure* (for additional details, see Appendix A3). This decision was made because LGPL effects are known to diminish with repeated testing over uniform test continua (e.g., Liu and Jaeger, 2018; Cummings et al., 2026). This 'return to normal' is not something the change models we test are designed to capture.

Best-fitting parameters were similar across folds. Best fitting κ_0 across-folds averaged 66.3 (SD = 2.11). Best fitting $\kappa_{c,0}$ and $\nu_{c,0}$ averaged 48.9 (SD = 2.57) and 48.6 (SD = 1.48) respectively, and the best fitting β_π for the decision model averaged 0.12 (SD = 0.004).

3.2 Results

Figure 2 visualizes model fits to human behavior. Averaging across all 32 conditions, the three mechanisms perform similarly in capturing listeners' categorization responses following exposure.

To assess model distinguishability, we conducted pairwise t-tests comparing each pair of models' per-observation log-likelihoods across cross-validation folds, separately for each condition. Of 96 pairwise comparisons (32 conditions * 3 model pairs), only five reached significance after Bonferroni correction. Among these, no model was consistently favored over any other.

Figure 3 highlights four representative cases that span the range of outcomes. For M5./ʃ/, all

[1] We validated the decision to use normalized cues against the test-only data from Cummings (2025). Normalized cues significantly improved predictive accuracy for the test-only data, compared to unnormalized cues (Appendix A5).

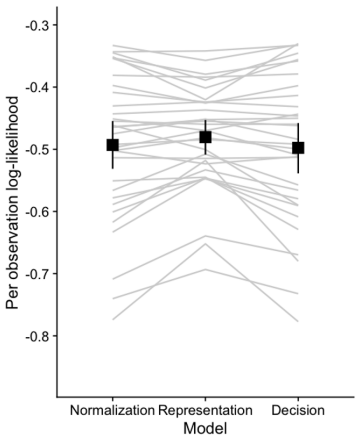

Figure 2: ASP change model fits across each of 32 exposure conditions, expressed in per-observation log-likelihood. Grey lines connect average fits per individual exposure condition over held-out post-exposure categorization responses across 5 folds of cross-validation. Points represent grand means over condition averages, with a 95% bootstrapped confidence interval.

three models provided statistically indistinguishable fits. 27 of 32 exposure conditions mirrored this trend. For exposure condition M6./ʃ/, the representation model systematically underpredicted the magnitude of the boundary shift, compared to the other two models. In contrast, changes in category representation provided the best numeric fits for exposure condition F5./s/ (though no differences reached significance following Bonferroni correction). Finally, for M2./ʃ/, changes in decision-biases provided a markedly worse fit than the other two models, which were themselves indistinguishable.

4 General Discussion

Recent findings suggest that some of the most frequently used experimental paradigms in research on speech perception provide comparatively little information about the adaptive mechanisms that they are meant to investigate (Xie et al., 2023). These findings were based on synthetic data from simulated experiments with simplified stimulus configurations. This leaves open whether the diagnosed empirical indeterminacy of existing paradigms is an artifact of the simplifying assumptions made by Xie and colleagues. The present work addressed this question directly, applying ASP to the largest existing LGPL dataset: 89,600 categorization responses from over 1,000 listeners across 32 exposure conditions. Our findings provide strong support for the conjecture provided

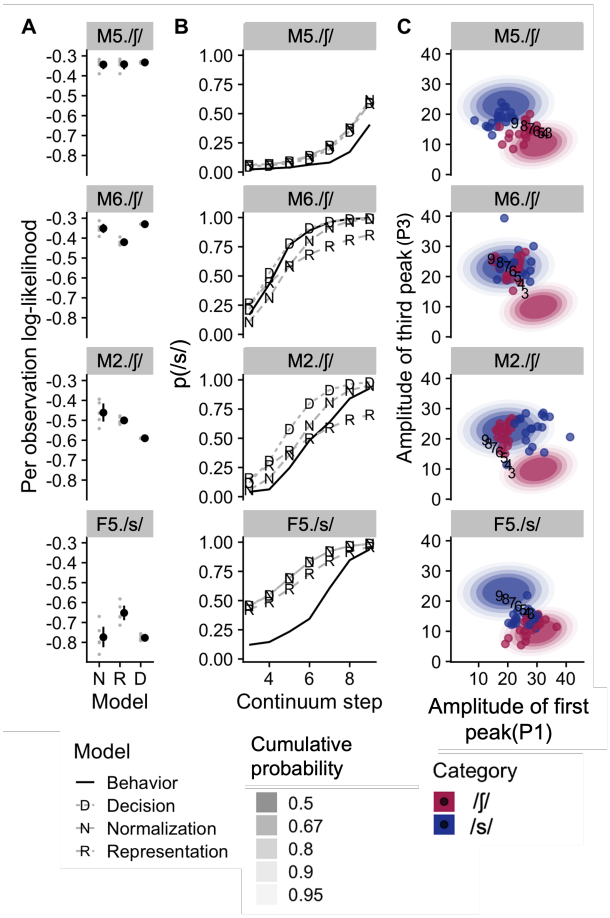

Figure 3: Divergent predictions of four representative exposure conditions (for all conditions, see Appendix A4 **Panel A:** Grey dots represent likelihoods within each fold of cross-validation; black pointranges show average likelihood and a 95% bootstrapped confidence interval over these folds. **Panel B**: Predicted perception of test continua by each model (averaged across 5 folds of cross validation, marked by the initial of the model employed) compared to human responses (solid line). **Panel C:** Input distributions by condition. Larger colored areas represent density distributions of listeners' assumed prior beliefs, colored by category. Points represent exposure input, colored by disambiguating lexical context. Black numbers represent the test tokens whose perception is shown in Panel B.

by Xie and colleagues: even for this large data set combining 16 different experiments, none of the three change models consistently outperformed the others. This should serve as a clear warning signal for future work: merely scaling up data collection within conventional experimental designs is unlikely to resolve the empirical indeterminacy between mechanisms within those paradigms.

At the same time, the scale and diversity of the Cummings (2025) data revealed important heterogeneity across exposure conditions. This provides

a glance at how much the degree of adaptivity observed in LGPL experiments might vary even if the exact same design, procedure, and approach to stimulus creation are used. For many of the exposure conditions, the predictions of the three change models were virtually indistinguishable; for some conditions, however, a subset of the three change models provided a superior fit. Importantly, when individual conditions did favor a particular mechanism, the winner was not always the representational change model—contrary to what is commonly assumed in the LGPL literature. In some cases, changes in representation yielded the worst predictions overall among the three change models (e.g., M6./ʃ/ in Figure 3). These patterns underscore a point that is easy to overlook when results are aggregated or when only a single talker is investigated (as in typical for the vast majority of LGPL experiments): the diagnosticity of an LGPL experiment depends heavily on the acoustic properties of the specific stimuli used—properties that are typically not controlled or even reported.

4.1 Tracing the indeterminacy to stimulus design

If elucidating the mechanisms of adaptive speech perception is the goal, how might one proceed? One possibility, raised in the introduction, is that 2AFC tasks are simply not sensitive enough to detect subtle behavioral differences caused by distinct adaptation mechanisms. If this is true, the field will face a formidable challenge: the 2AFC is not merely a convenience but a foundation on which theories across speech perception, phonetic acquisition, category learning have been built. Abandoning it would sever connections to decades of cumulative evidence across these subfields.

The divergence maps presented in Figure 4 point to a more optimistic alternative. To create these maps, we computed the pairwise difference between change models predictions over the acoustic space—the predicted probability of an /s/-response immediately after exposure, depending on the acoustic cues (P1 and P3). This process was repeated for all pairs of change models and for each of the 16 talkers. This reveals where (i.e., in which part(s) of the acoustic space) the three change models make distinct predictions.

The divergence maps of the two exposure conditions in Figure 4 offer two important insights. First, for both exposure conditions, diagnosticity could be improved by selecting different test stimuli. The test continua used in Cummings (2025) (open circles) occupy regions of the acoustic space where power to detect a between-condition difference is high (indicated by large tile size), but where the three change models make similar predictions (indicated by light coloring). In other words, the

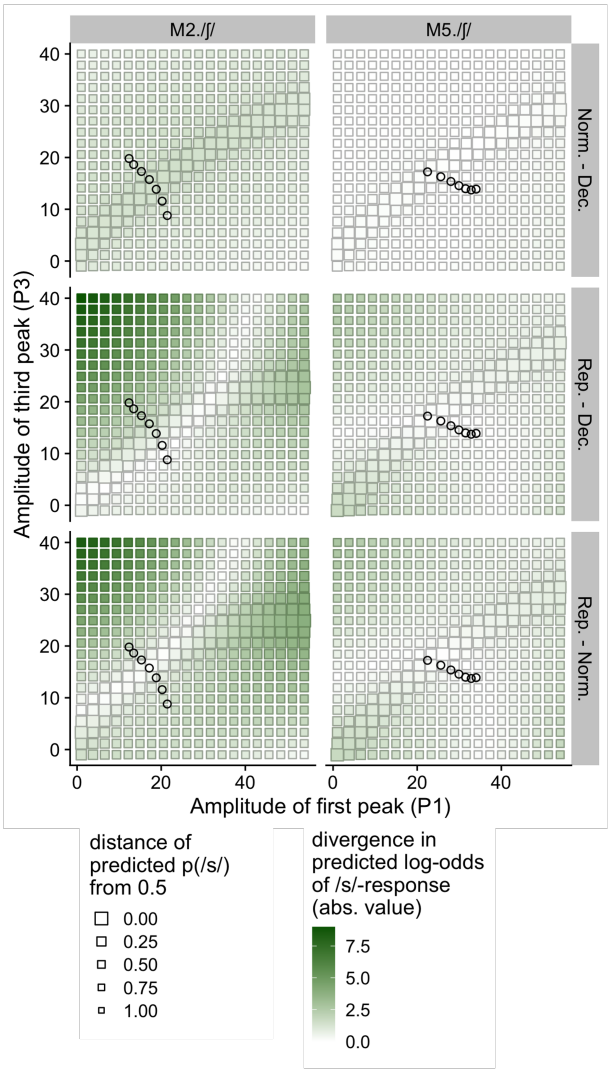

Figure 4: Predicted diagnosticity of different stimuli in the Cummings (2025) data. Shown is the predicted divergence in change model predictions across the acoustic space after two different exposure conditions (M2./ʃ/ and M5./ʃ/, columns). Each model was simulated 100x, with parameter values drawn from a normal distribution centered on the best fitting parameters to the human LGPL data and with variance drawn from the average estimated parameter uncertainty across the cross-validation folds. Diagnosticity is higher for more intense colors (larger differences between model predictions) and larger tiles (power to *detect* the difference in predictions is maximized at $E[p(/s/\text{-}response)]$). Open circles show the continuum steps presented in Cummings (2025). Acoustics for the two exposure condition are shown in Figure 3C.

conventional test tokens are well-placed for detecting *that* adaptation occurred, but poorly placed for determining *which* mechanism underlies it. A notable exception to this are the test tokens for M2./ʃ/, which cross the space that has the highest diagnosticity for the comparison of normalization vs. changes in decision-making (top left panel). Even for this exposure condition, however, diagnosticity for the other two pairwise comparisons would have been improved by selecting test stimuli elsewhere in the acoustic space (at regions with darker colors). For example, normalization and representational changes, which were indistinguishable in Figure 3A, could become separable if the experimenter were to use test tokens placed in the high-divergence regions (e.g., the large green squares in the bottom left panel of Figure 4). For M5./ʃ/, the same logic applies, though the expected gains are more modest.

Second, the two exposure conditions differ markedly in how much diagnosticity any test token placement can achieve. Across the entire acoustic space, M2./ʃ/ yields larger divergences between models than M5./ʃ/. This means that the choice of exposure input—not just test tokens—also constrains diagnosticity.

Taken together, these observations suggest that the empirical indeterminacy documented above reflects the stimulus design, rather than being a principled limitation of 2AFC tasks. The models' predictions converge not because the behavioral measure lacks resolution, but because competing mechanisms happen to make similar predictions for the particular acoustic tokens used during the exposure and test phases of the experiment. These observations point towards two concrete changes in experimental practice.

4.2 Actionable recommendations

First, **selection of test token should be guided by model-predicted divergence rather than perceptual ambiguity alone**. The standard practice for LGPL experiments of constructing a continuum between two category endpoints aims at ensuring that experimenters can detect the *presence* of adaptive changes (differences between the two exposure conditions of an LGPL experiment). However, this does not guarantee that the experiment will be diagnostic of the *mechanisms* underlying those changes. We propose that test tokens be selected according to three criteria: (i) high predicted divergence between models, (ii) avoidance of floor or ceiling response rates (where power to detect effects in 2AFC is lowest), and (iii) sufficient acoustic spread to mitigate concerns about perceptual anchoring or repetition fatigue. This change requires no modification to the task, the exposure phase, or the sample size.

Second, echoing a recommendation made in Xie et al. (2023), **researchers should examine and report the acoustic properties of their stimuli, and use model-guided simulation to evaluate diagnosticity before data collection**. It is not standard in LGPL or accent adaptation research to report acoustic cue values for exposure or test stimuli. Yet our results demonstrate that conditions vary drastically in their ability to distinguish between the mechanisms, even when identical methods are used. As illustrated by the M5./ʃ/ condition above, some exposure inputs inherently limit diagnosticity regardless of test token placement. Researchers cannot know which situation they are in without examining the acoustics and computing divergence maps, or at a minimum, simulating their experiment within ASP or a comparable computational framework. We submit that this kind of pre-experimental diagnosticity check should become routine.

Throughout this paper, we have exclusively compared models wherein adaptation is attributable to a single mechanism. Having established that the current data are not reliably diagnostic even between these extreme scenarios, they certainly cannot identify where a listener falls within the larger space of possible mechanism mixtures (Figure 1C). Nevertheless, the tools developed here are prerequisites for addressing identification weighted mixtures of mechanisms. Determining that each mechanism makes divergent predictions in isolation is a necessary first step before assessing what combination might best account for a given phenomenon.

In conclusion, the 89,600 categorization responses analyzed here carry strikingly little information about which adaptive mechanisms underlie the observed behavior, despite the fact that these mechanisms differ starkly in the computations they involve. The divergence maps provided in Figure 4 suggest that this indeterminacy is not a limitation of the behavioral task itself. It is, at least in part, a consequence of stimulus selection practices that can be improved with relatively modest, model-guided changes. We hope that the framework and tools presented here provide a foundation for this next phase of research—one focused not on whether

adaptation occurs, but on how.

Acknowledgments

Research reported in this work was supported by NIH-NICHD grant R01HD111936.

References

Keith S. Apfelbaum, Ethan Kutlu, Bob McMurray, and Efthymia C. Kapnoula. 2022. Don't force it!: Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, 152:3728–3745.

Santiago Barreda and T. Florian Jaeger. 2025. Reintroducing and testing the probabilistic sliding template model of vowel perception. *Linguistics Vanguard*.

Tessa Bent and Melissa M Baese-Berk. 2021. Perceptual learning of accented speech. *The Handbook of Speech Perception*, pages 428–464.

Paul-Christian Bürkner. 2017. Advanced bayesian multilevel modeling with the R package brms. *Preprint*, arXiv:1705.11123.

Shawn Cummings. 2025. *Linking lexically guided perceptual learning to statistical patterns in speech input*. Ph.D. thesis, University of Connecticut.

Shawn N. Cummings, Emma C. Hodges, and Rachel M. Theodore. 2026. Cumulative input sensitivity predicts both attenuation and stability of lexically guided perceptual learning. *Psychonomic Bulletin & Review*, 33:116.

Olivia Guest and Andrea E. Martin. 2021. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16:789–802.

Jingyuan Huang and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, 125:3983–3994.

Alexandra Jesse. 2021. Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47:184–194.

Keith Johnson and Matthias J Sjerps. 2021. *Speaker normalization in speech perception*, chapter 6. John Wiley & Sons, Ltd.

Allard Jongman, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Tanya Kraljic and Arthur G Samuel. 2005. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51:141–178.

Erika JC Laing, Ran Liu, Andrew J Lotto, and Lori L Holt. 2012. Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in psychology*, 3:203.

Linda Liu and T Florian Jaeger. 2018. Inferring causes during speech perception. *Cognition*, 174:55–70.

Bob McMurray and Allard Jongman. 2011. What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118:219–246.

Emily B Myers and Laura M Mesite. 2014. Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, 76:80–93.

Terrance M. Nearey and Peter F. Assmann. 2007. Probabilistic 'sliding-template' models for indirect vowel normalization. In Maria-Josep Solé, Patrice Speeter Beddor, and Manjari Ohala, editors, *Experimental Approaches to Phonology*, pages 246–269. Oxford University Press, Oxford.

Allen Newell. 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In William G. Chase, editor, *Visual Information Processing*, pages 283–308. Academic Press, New York.

Dennis Norris, James M McQueen, and Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology*, 47:204–238.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Odette Scharenborg and Esther Janse. 2013. Comparing lexically guided perceptual learning in younger and older listeners. *Attention, Perception, & Psychophysics*, 75(3):525–536.

Maryann Tan and T. Florian Jaeger. 2025. Learning to understand an unfamiliar talker: Testing distributional learning as a model of rapid adaptive speech perception. *Cognition*, 265:106195.

Kodi Weatherholtz and T Florian Jaeger. 2016. Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics*.

Xin Xie, T. Florian Jaeger, and Chigusa Kurumada. 2023. What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review. *Cortex*, 166:377–424.

Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction ver explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12:1100–1122.

A Appendix

A1: Lexically guided perceptual learning and implementation under ASP

See Figure 5, below.

A2: Acoustic measurements

Many acoustic cues have been posited to underlie the distinction between /s/ and /ʃ/, with spectral center of gravity being a popular choice (e.g., Jongman et al. 2000). /s/ is characterized by high energy at high frequencies (above 6.5 kHz), while /ʃ/ tends to have more energy concentrated below 3.5 kHz (aligned with vowel formants). To characterize our tokens, we extracted the amplitude of the spectral peak in each of these ranges for each of our tokens. These cues necessarily correlate highly with spectral center of gravity (as tokens with more energy at a higher spectral peak will have a higher spectral average). See Figure 6, below.

All exposure and test tokens were annotated based on these cues, which were used as input to the ASP model (shown in Figure 7, below).

A3: Findings of Cummings (2025)

Figure 8 summarizes the results of Cummings (2025), separately for each of the 16 talkers. The LGPL effects shown in this figure were obtained from a Bayesian mixed-effects logistic regression to estimate listeners' responses *immediately following exposure*. We used package `brms` (Bürkner, 2017) in R (R Core Team, 2021) to predict the probability of /s/-responses from as a function of continuum step and test block, as well as random by-participants intercept and slopes for all effects. We treated both continuum step and test block as monotonic, potentially non-linear, effects using function `mo()` of the `brms` package. The effects we report are those estimated for the *0th* test trial, i.e., participants' behavior before the first test trial.

For the fitting of change models, we used these estimates of $p(/s/\text{-}response|continuum\ step, test\ block, participant)$ before any effects of repeated testing. Specifically, we obtained one estimate per listener and continuum step (instead of the 10 observations per listener and continuum step). This approach has the additional upside that it avoids inflating the evidence available to change models since the 10 responses per listener and continuum steps are *not* independent observations. It should be noted, however, that our approach is conservative and likely somewhat under-estimates the available amount of evidence. An alternative, better, approach would be to develop hierarchical implementations of change models that can adequately handle repeated-measures data from participants, while also extending the models to account for the effects of repeated testing. This is a non-trivial effort to be pursued in future work.

A4: Individual conditions as simulated by ASP

Extensions of Figure 3 to include all 32 input conditions are shown in Figures 9 and 10, below.

A5: Normalized and unnormalized token distributions

See Figure 11, below.

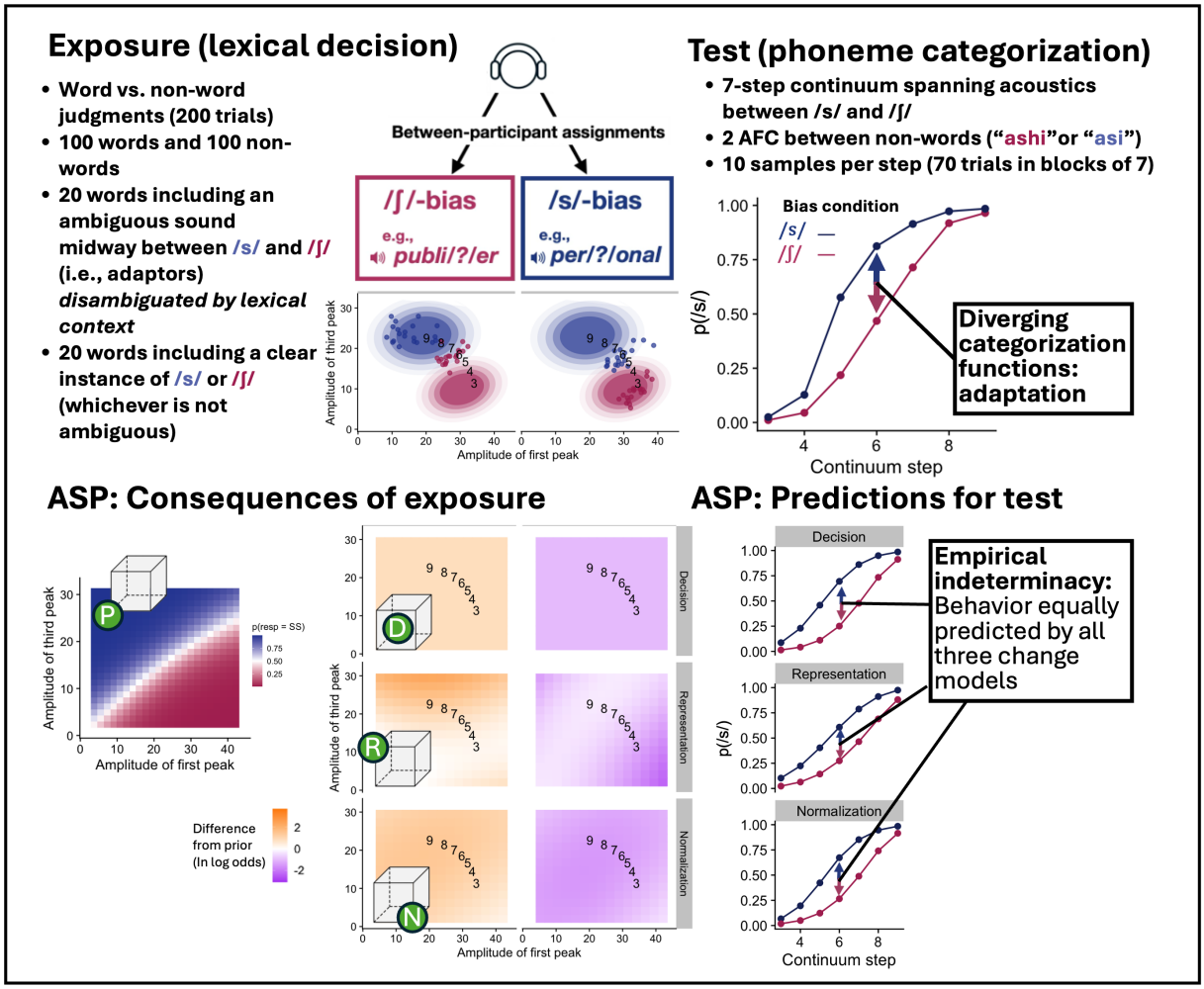

Figure 5: Schematic of the lexically guided perceptual learning (LGPL) paradigm, and simulation under ASP. In the scatter plot under 'Exposure', colored dots represent the acoustics of fricatives in exposure, while black numbers designate the acoustics of the continuum presented in test. Filled areas show listeners' assumed long-term expectations about cue-category mappings (see subsection 3.1). The leftmost panel under 'ASP' shows predictions of the categorization model, and the right panels show differences in expected categorizations after integrating exposure input (assuming only a single change mechanism is active).

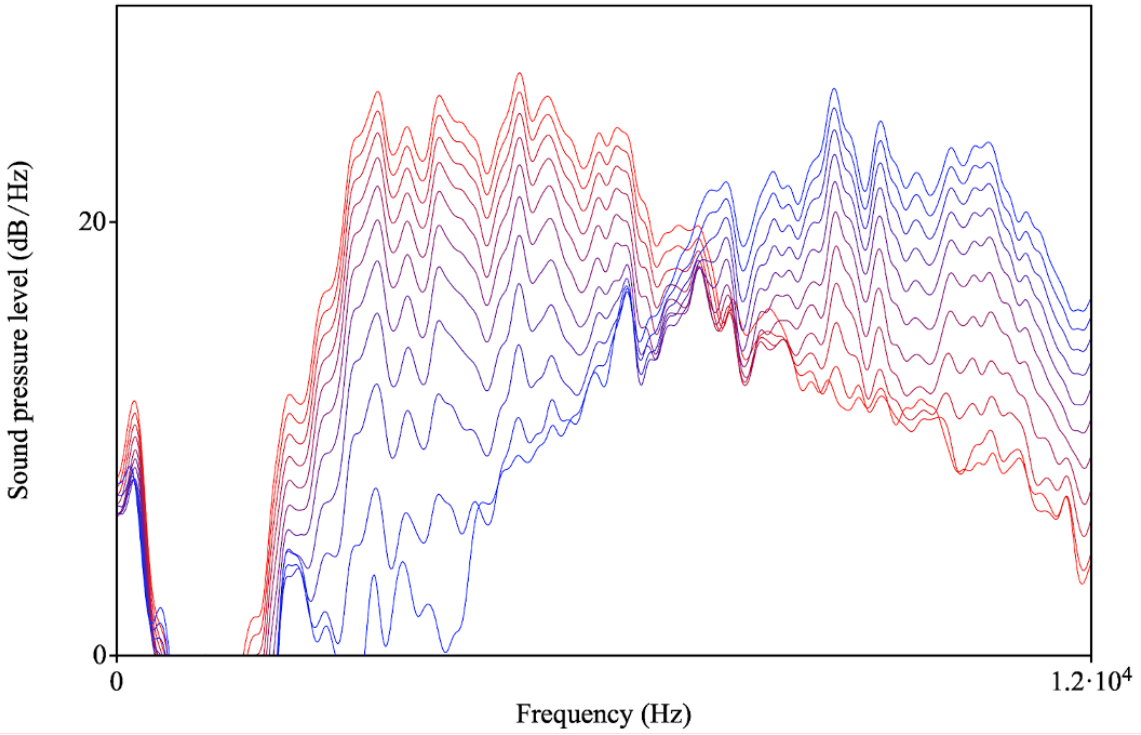

Figure 6: Long-term average spectra for 11 steps spanning a continuum from /s/ to /ʃ/ (as recorded in “asi” and “ashi” by Talker F1). Blue and red endpoints represent clear /s/ and /ʃ/ segments. As tokens become more /ʃ/-like, energy in the 1.9–4.5 kHz range diminishes and energy in the 8.5–11 kHz range increases.

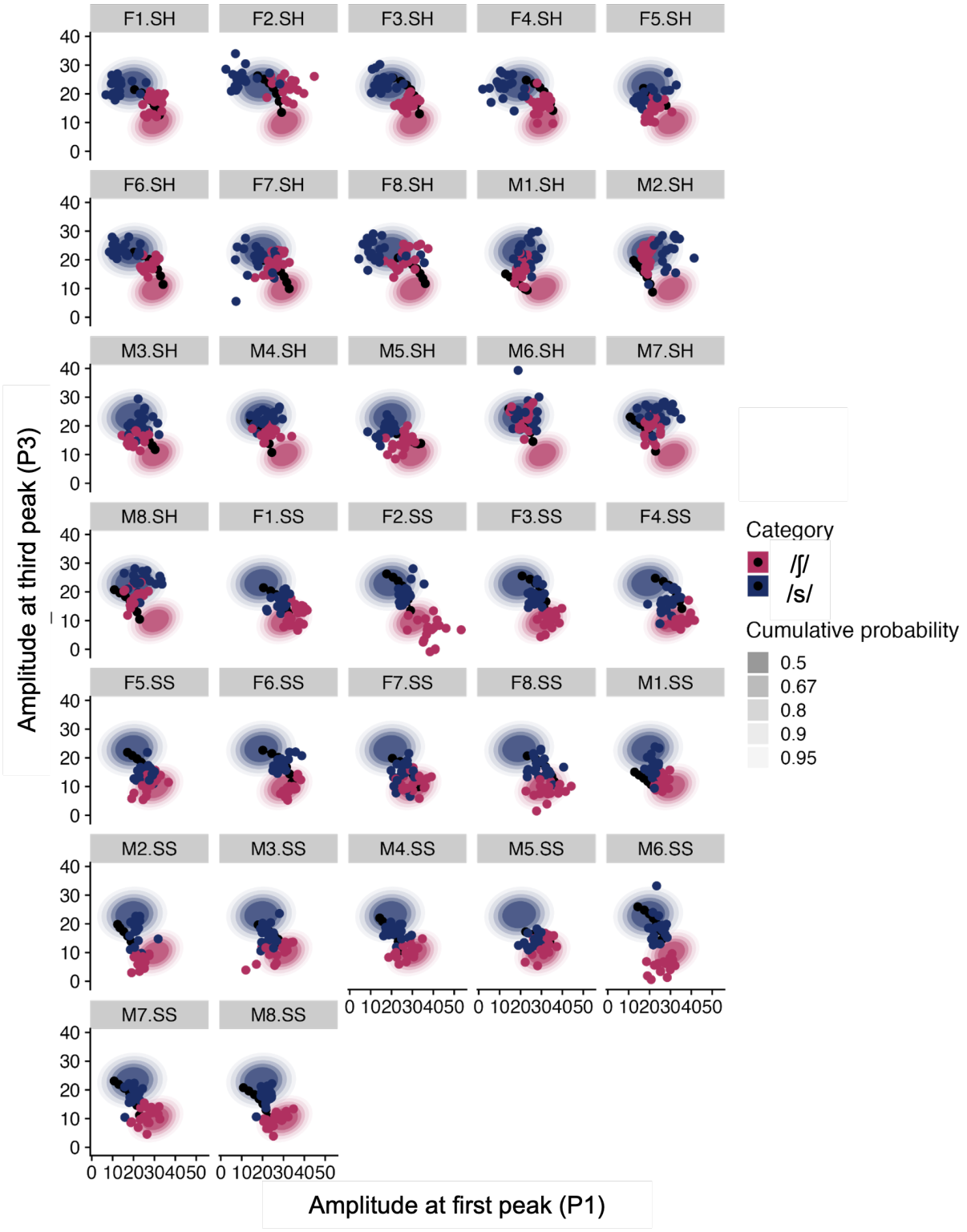

Figure 7: Extension of Panel A of Figure 2 in main text, showing all conditions. Input distributions by condition. Larger colored areas represent density distributions of listeners' assumed prior beliefs, colored by category. Colored dots represent exposure input, colored by disambiguating lexical context. Black points represent tokens presented in test.

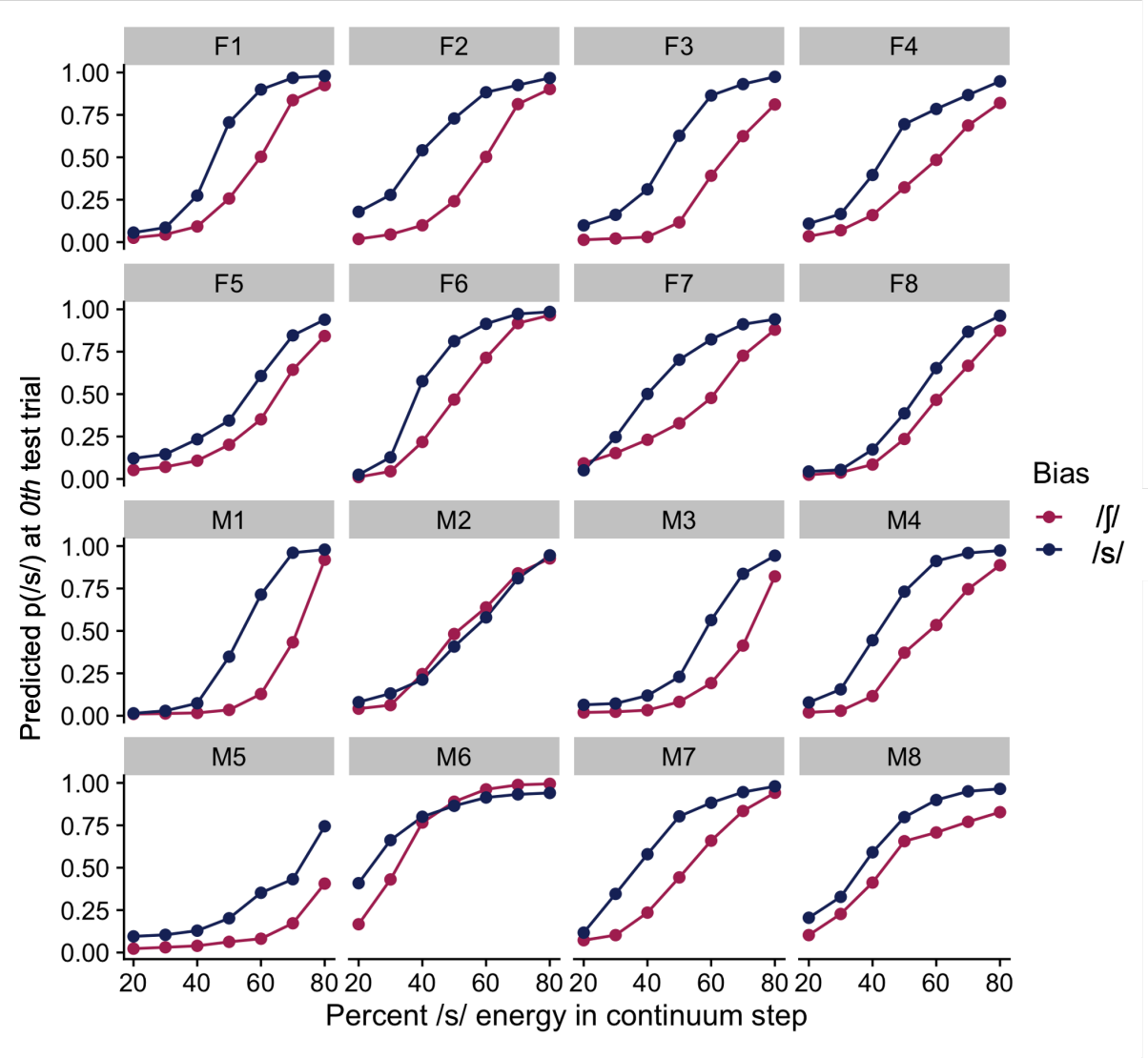

Figure 8: Results across conditions from Cummings (2025), showing an estimate of listeners' categorization responses before repeated testing could affect them.

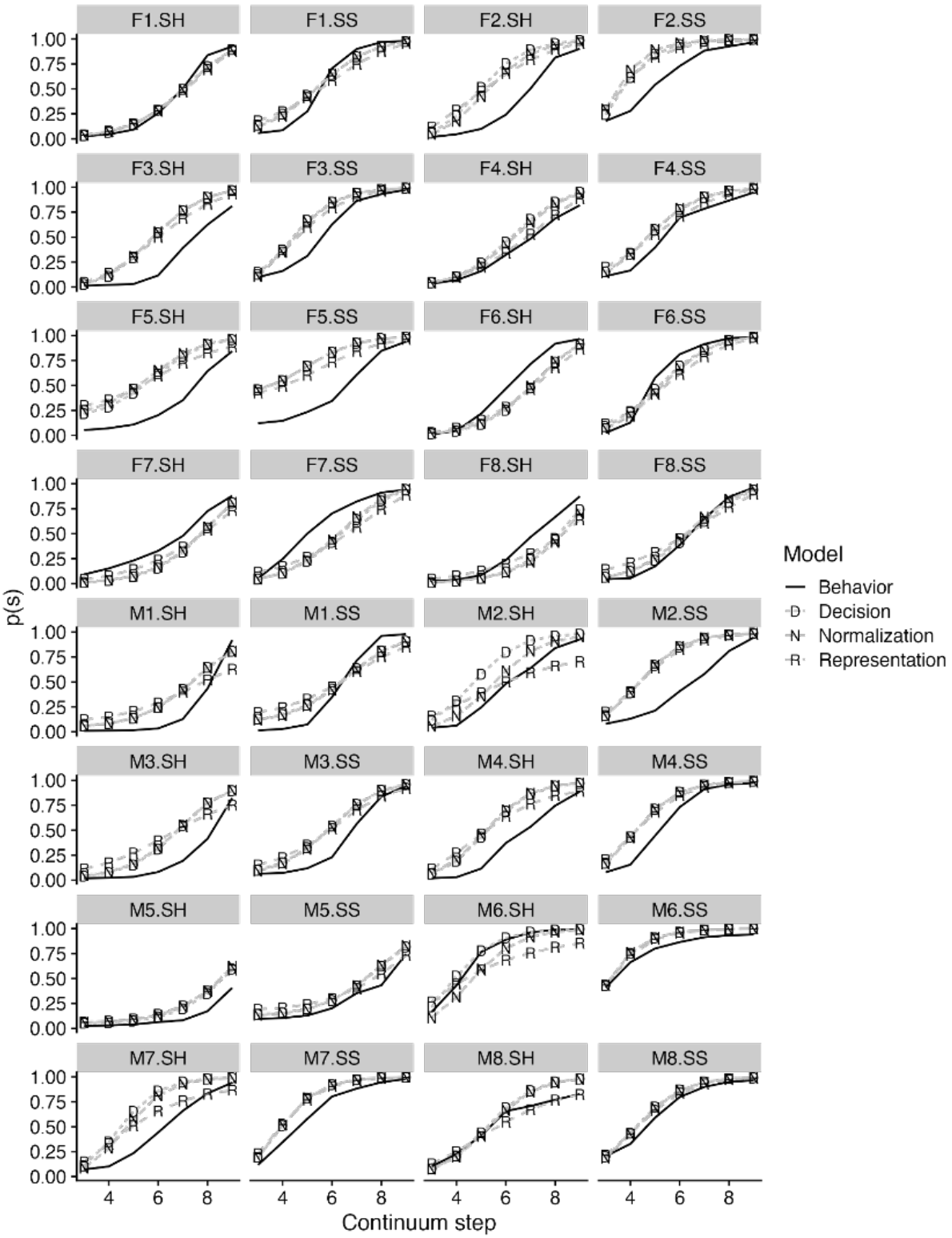

Figure 9: Extension of Panel B of Figure 3 in main text, showing all conditions. Predicted perception of test continua by each model (averaged across 5 folds of cross validation, marked by the initial of the model employed) compared to human responses (solid line).

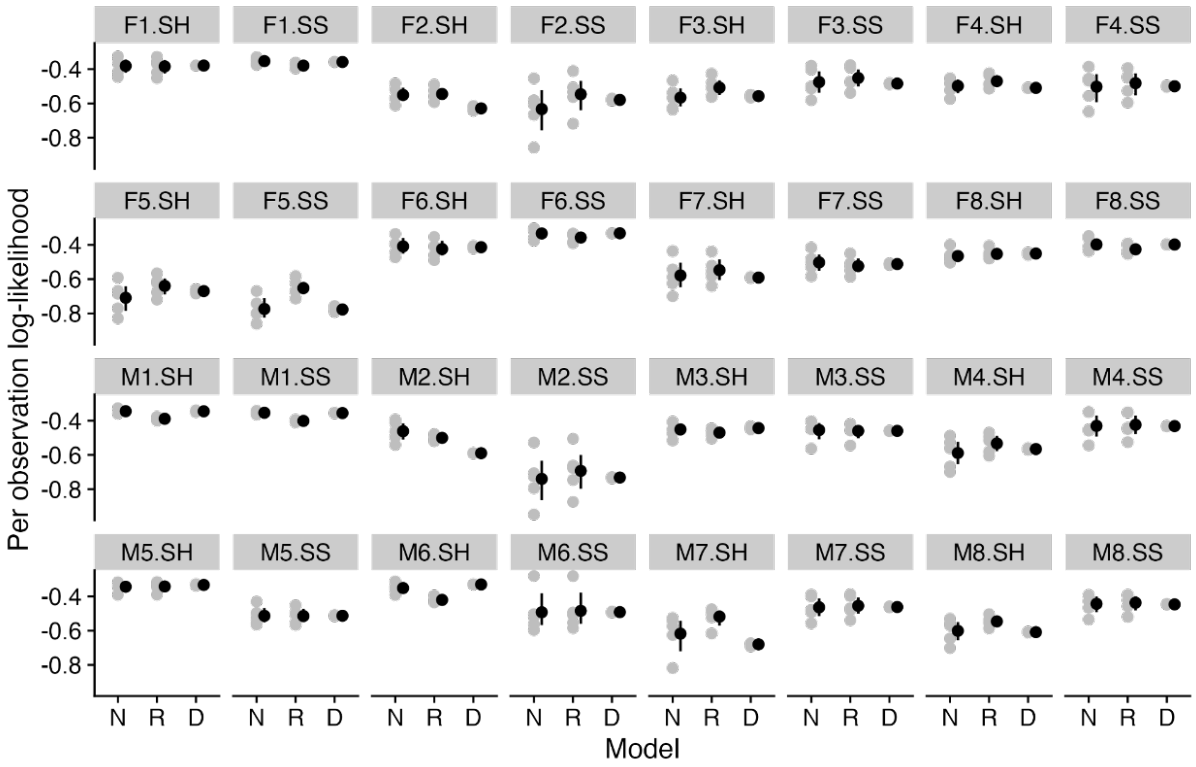

Figure 10: Extension of Panel A of Figure 3 in main text, showing all conditions. Grey dots represent likelihoods within each cross-validation fold; black pointranges show average likelihood and error as a 95% bootstrapped confidence interval.

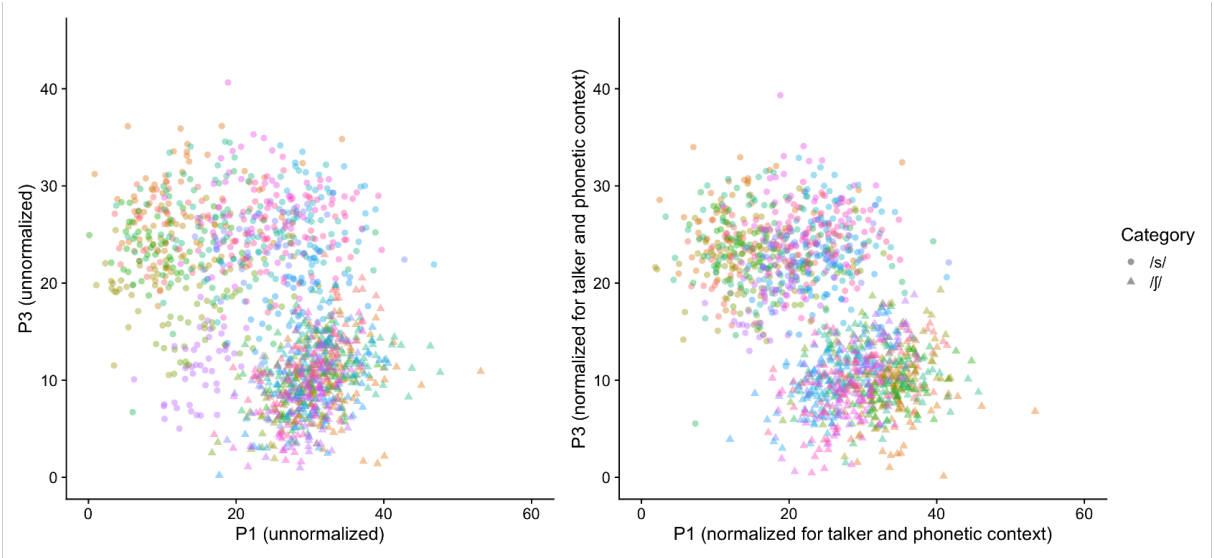

Figure 11: Normalized (right) and unnormalized (left) token distributions. Triangular tokens designate /s/, while circles designate /ʃ/. Color indicates talker. All simluations reported in main text were based on normalized acoustics.