Modeling generalization in perceptual learning of speech

Yiming Lu
Basque Center on Cognition, Brain and Language
y.lu@bcbl.eu

Xinyu Leslie Liao
University of Toronto
xinyuleslie.liao@mail.utoronto.ca

Alejandro Tabas
Basque Center on Cognition, Brain and Language
a.tabas@bcbl.eu

Xin Xie
University of California, Irvine
xxie14@uci.edu

Abstract

A hallmark of learning is generalization to novel instances. In speech, exposure to atypical pronunciation drives perceptual adjustment that can generalize to unheard tokens. Prior work has attributed constraints on generalization primarily to acoustic similarity between exposure and test contexts. We propose that generalization can also be understood as an inference problem: listeners must determine whether, and how strongly, a learned phonetic mapping should apply in a new context. We test this proposal using data from a recent experiment in which listeners were exposed to shifted vowel pronunciations and then tested on minimal pairs varying in lexical frequency. Learning effects appeared strongest when the exposure direction aligned with a high-frequency alternative in mixed-frequency pairs, and were absent for low-frequency pairs. The observed pattern could reflect token-level acoustic similarity, reliance on prior expectations, or frequency-dependent constraints in applying the learned mapping. We formalized these alternatives within a Bayesian belief-updating framework: a talker-specific model assuming full transfer, a mixture-of-expectations model that interpolates between the updated representation and the listener's prior, and a hierarchical Bayesian model that deploys the updated representation with uncertainty. The talker-specific model captured most generalization patterns through its sensitivity to token-level acoustic properties, but overpredicted learning for low-frequency pairs. The hierarchical model best recovered the theoretically central exposure-control contrast pattern, suggesting that lexical frequency may constrain how learned representations are applied. Our results provide a computationally explicit framework for studying how contextual factors shape generalization in speech perception.

Keywords: Perceptual learning, computational modeling, generalization, lexical frequency, acoustic similarity

1 Introduction

Listeners rapidly adjust their speech perception in response to novel pronunciation patterns, demonstrating remarkable adaptivity (Norris et al., 2003; Idemaru and Holt, 2011; Clarke and Garrett, 2004; Bradlow and Bent, 2008; Maye et al., 2008). One prominent example of this adaptivity is manifested via lexically-guided perceptual learning (Norris et al., 2003): brief exposure to an ambiguous sound in a lexically biasing context—such as a vowel midway between /ɪ/ and /ɛ/ embedded in the word "l[ɛ]mon", where /ɛ/ is the only viable interpretation—shifts how listeners subsequently categorize tokens along the relevant /ɪ/-/ɛ/continuum such that more tokens are perceived as /ɛ/ following exposure. This learning effect is robust across phonemic contrasts and languages (Kraljic and Samuel, 2006; Chládková et al., 2017; Mitterer et al., 2011).

A central question is how such learning generalizes beyond the immediate exposure context. Prior work has extensively investigated generalization across talkers (Bradlow and Bent, 2008; Alexander and Nygaard, 2019; Baese-Berk et al., 2013; Xie and Myers, 2017; Xie et al., 2021; Aoki and Zellou, 2025a,b), phonetic contexts and word positions (Jesse and McQueen, 2011; Mitterer and Reinisch, 2017; Bowers et al., 2016), asking whether adjustments learned during exposure transfer to test. In this literature, similarity between exposure and test has been hypothesized to be a primary explanatory construct: generalization succeeds when the test context is sufficiently similar to exposure, and fails otherwise (Kraljic and Samuel, 2006; Eisner and Mcqueen, 2005; Reinisch and Holt, 2014; Xie and Myers, 2017; Xie et al., 2021). However, similarity has been conceptualized at different levels. At the broadest level, generalization has been shown to depend on global talker similarity — whether the test voice matches the exposure voice (Eisner et al.,

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 529–541

2013; Reinisch and Holt, 2014). At a finer grain, listeners may evaluate the acoustic similarity between specific test tokens and the distributions encountered during exposure, such that tokens closer in the acoustic space to the exposure input are more readily accepted into the updated categories (Chládková et al., 2017; Jin et al., 2025). Yet similarity defined at these levels do not always capture the empirical pattern (Lai and Tamminga, 2024), and existing accounts also leave open the question of how similar is similar enough to drive generalization. In the current paper, we investigate the possibility that beyond acoustic similarity between exposure and test tokens, generalization is further constrained by a listener's confidence in deploying the learned mapping in a given test context, which depends on additional contextual factors.

A recent experiment (Liao and Kang, 2026) (henceforth LK26) provides a particularly informative case. In LK26, listeners were exposed to an ambiguous vowel between /ɪ/ and /ɛ/ embedded in high-frequency, non-minimal-pair words (e.g., *kitchen*), with the ambiguity biased toward either /ɪ/ (lowering group) or /ɛ/ (raising group), along with a control group hearing canonical pronunciations of both vowels. At test, participants categorized tokens from a five-step vowel continuum embedded in /ɪ/-/ɛ/ minimal pairs. Critically, the test pairs varied in the lexical frequency of the /ɪ/-consistent and /ɛ/-consistent words, yielding four frequency configurations: high-high (HH; e.g., *sit-set*), high-low (HL; e.g., *kick-keck*), low-high (LH; e.g., *din-den*), and low-low (LL; e.g., *trimmer-tremor*). The results revealed that exposure-induced categorization shifts were not uniformly expressed across test conditions (Figure 1). In the mixed-frequency conditions (HL and LH), shifts relative to the control condition were observed only when the direction of exposure bias aligned with the high-frequency member of the pair: in HL, the lowering group differed from control but the raising group did not; in LH, the reverse held. In the LL condition, neither biasing group differed reliably from control, indicating an absence of detectable learning effects when both lexical alternatives were low-frequency. In HH, the lowering effect was more clearly supported than the raising effect, despite identical exposure structure.

These findings raise an important question: do the differences across frequency conditions reflect differences in the learning process itself, or differences in how a learned phonetic mapping is applied

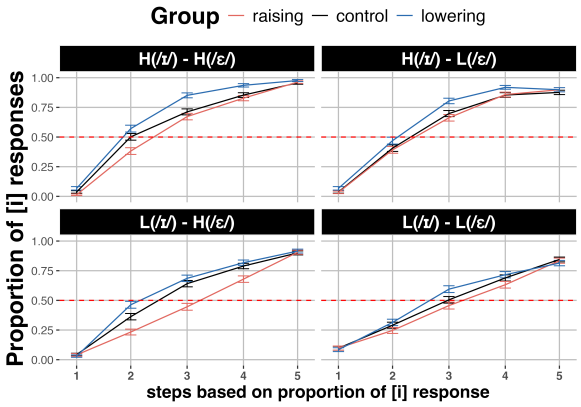

Figure 1: Results of the LK26 experiment. Points indicate the mean proportion of /ɪ/-response. Error bars indicate the standard error of the mean across subjects. On the x-axis, step 1 represents a more /ɛ/-like stimulus, and step 5 represents a more /ɪ/-like stimulus. The distributions of exposure and test items are shown in Appendix Figure 3 and 4.

at test? Several observations narrow the space of explanations. First, the same talker produced all exposure and test items, so differences across frequency conditions cannot be attributed to talker identity. Second, a frequency-based similarity account—predicting generalization only when the test frequency bias matches exposure—captures the HL, LH, and LL patterns but cannot explain why the lowering group generalized in HH, where neither alternative had a frequency advantage. Third, if lexical frequency operated as a general response bias favoring high-frequency words, exposure effects should appear regardless of whether the high-frequency member aligns with the direction of the shift. Yet the data show that only the shift aligned with the high-frequency alternative was expressed. Together, these observations suggest that lexical frequency did not operate as a simple similarity dimension or a global response bias.

Because the exposure items were held constant for each group, it is reasonable to believe that exposure-elicited learning did not vary across frequency conditions. What varied was how learning was expressed as a function of word frequency. This points to a mechanism that modulates the *expression* of learning, rather than learning itself.

In this paper, we ask whether, beyond acoustic similarity, lexical frequency may serve as the contextual factor that modulates the expression of learning. A working hypothesis is that high-frequency words, supported by more robust mental representations (Pierrehumbert, 2016; Connine

et al., 1993; Todd et al., 2019), may provide a sufficiently stable lexical context for the listener to confidently deploy a recently learned phonetic adjustment; low-frequency words, with less well-specified representations, may not. We propose that generalization of perceptual learning can be understood as an inference about the *applicability* of a learned phonetic mapping to a novel test context. In what follows, we formalize the process of *generalization under uncertainty* in three computationally distinct models and evaluate the three models against the LK26 data. Our goal is to determine what kind of mechanism gives rise to the observed empirical patterns—specifically, the selective expression of learning as a function of the frequency structure of test items. In doing so, we seek to contribute a computationally explicit framework to the broader literature on how learned phonetic adjustments generalize beyond the contexts in which they were acquired.

2 Possible strategies of generalization

How should we conceptualize the problem that listeners are trying to solve? We contend here that, similar to perceptual learning, generalization behavior can be understood in a general framework of perceptual inferences under uncertainty. Upon exposure to novel pronunciation, listeners update their internal generative models. During generalization, listeners must infer which generative model is most relevant to the current context, as well as the degree to which that model should guide perception (Xie and Myers, 2017; Xie et al., 2021; Tan and Jaeger, 2025).

One possibility is that listeners treat generalization trials as a re-encounter with a previously experienced situation, directly applying the updated model to novel tokens. Evidence for such full transfer comes from studies showing that learned boundary shifts generalize across word positions without attenuation (Jesse and McQueen, 2011). This scenario can be modeled through direct application of a Bayesian ideal adaptor updated during exposure (Kleinschmidt and Jaeger, 2015), widely employed to account for adaptation across various paradigms, including selective adaptation (Kleinschmidt and Jaeger, 2015), accent adaptation (Tan et al., 2021), cue reweighting (Lu and Xie, 2024), and lexically-guided perceptual learning (see a comprehensive review in Xie et al. (2023)).

However, generalization is not always complete. Listeners sometimes show little transfer despite successful learning—for instance, when the learned pattern is produced by a speaker for whom it is unexpected (Eisner et al., 2013). Two computational mechanisms could give rise to such constrained generalization. First, listeners may have strong prior expectations that compete with the updated model, effectively falling back on longer-term experience. Kleinschmidt and Jaeger (2015) allude to this mechanism: categorization can be determined by combining predictions of multiple generative models weighted by their inferred relevance, so that limited generalization arises from a stronger influence of the prior. We formalize this as a mixture-of-expectations model, in which responses reflect a weighted combination of the exposure-updated representation and the listener's prior knowledge.

Second, listeners may still adopt the updated model but deploy it with greater uncertainty. This intuition can be formalized within a hierarchical Bayesian framework, which has been widely used to explain phenomena such as human concept learning (Kemp et al., 2007), linguistic coordination and convention (Hawkins et al., 2023) and syntactic priming (Xu and Futrell, 2024). Related ideas have also been proposed in speech perception, although they have either not been directly implemented (Pajak et al., 2016; Kleinschmidt and Jaeger, 2015), or have been implemented in contexts that differ substantially from the present case (Pajak et al., 2013; Hosier and Bicknell, 2018; Nielsen and Wilson, 2008). Following Xu and Futrell (2024), we model listeners as first updating a higher-level talker-specific representation through exposure, then generating a lower-level model for novel test contexts by drawing on this higher-level representation. A hyperparameter controls how tightly the lower-level model draws on the updated representation: larger values yield stronger pooling (more confident deployment), while smaller values permit greater deviation (less confident deployment). In this way, limited generalization arises from higher uncertainty about the applicability of the updated model, rather than from reliance on a competing prior.

3 Modeling

We now provide the mathematical formulation for the three candidate mechanisms introduced above. All three share a common learning component—a Bayesian belief-updating model that estimates

updated phonetic category representations from exposure—and differ only in how the learned representation is expressed at test. In other words, the models implement alternative assumptions about how listeners decide whether the exposure-updated representation should be applied when categorizing test tokens in a new lexical context. In both constrained models (mixture-of-expectations and hierarchical Bayesian), the transfer parameter is allowed to vary by frequency, capturing the hypothesis that high-frequency test contexts support more confident deployment than low-frequency ones. Next, we describe the steps taken to obtain model prediction, beginning with the estimation of listeners' prior knowledge, before turning to data pre-processing, confound control, model details, and optimization procedures.

3.1 Estimating prior knowledge

We adopted the assumption that our participants' baseline knowledge reflects cumulative experience with speech in their native environment. To approximate this experience, we pooled together three Canadian English corpora, reported in a cross-dialect analysis of English /ɪ/–/ɛ/ allophony (Smith et al., 2025). We removed outliers on both the F1 and F2 dimensions, excluding tokens more than 3 standard deviations from the mean. Within each talker, we sampled an equal number of tokens from each category to ensure balanced representation. We excluded talkers who contributed fewer than the median of the distribution (104 tokens), and randomly sampled 104+ tokens from each category for each of the 96 remaining talkers, resulting in a total of 10,008 observations. The resulting corpus is shown in Appendix Fig. 5.

3.2 Pre-processing

We converted formant frequencies to the Mel scale and applied a talker-normalization scheme (McMurray and Jongman, 2011). F1 and F2 were centered relative to each talker's mean cue value. We performed this procedure on all the tokens, including the experiment tokens and those from the corpus used to estimate prior knowledge.

3.3 Category prior and version bias

The test continuum was previously calibrated so that the midpoint corresponded to equal /ɪ/ and /ɛ/ response rates, but listeners in this sample showed more /ɪ/ responses than expected in three of four conditions (HH, HL, LH in Fig. 1). This could reflect an asymmetric category prior or a bias from residual acoustic properties of the source words used to synthesize the test tokens. To estimate these confounds, we fit a generalized linear mixed-effects model to the first 20 test trials from each control-group participant, restricting the window to minimize influence of any additional learning during the test phase (Tan and Jaeger, 2025). Fixed effects included (i) scaled log-odds predicted by an ideal observer fitted from the Canadian English corpora, (ii) scaled log-odds predicted by relative frequency differences in SUBTLEX Zipf scores (Brysbaert and New, 2009), and (iii) version (sum-coded; /ɪ/ = 1, /ɛ/ = -1). Random effects included by-subject intercepts and slopes, and by-word-pair intercepts.

Listeners were biased toward /ɪ/, with an estimated baseline probability of 0.65 of /ɪ/-response according to the intercept of the regression model. Version also significantly predicted responses ($\beta = 0.54$, $z = 5.01$, $p < 0.001$), indicating that tokens morphed from /ɪ/-words were more likely to be categorized as [ɪ], and vice versa. The same regression coefficients were used in all models below to offset the effects of these variables when computing categorization responses.

3.4 Learning model

The learning model is a Bayesian ideal adaptor, as proposed in Kleinschmidt and Jaeger (2015). This belief-updating model assumes that listeners represent vowel categories as multivariate Gaussian distributions over F1 and F2, parameterized by mean μ_c and covariance Σ_c. Listeners' prior knowledge for each category c is estimated from the speech corpus and parameterized by $\mu_{c,0}$ and $\Sigma_{c,0}$. Two hyperparameters, κ_0 and ν_0, encode listeners' confidence in the prior mean and covariance, respectively. κ_0 and ν_0 can be interpreted as pseudo-counts: smaller values imply weaker prior commitment and therefore faster adaptation to incoming evidence.

Upon exposure to a talker, the prior is updated using the observed input for each category. Let N_c denote the number of observed tokens in category c. The posterior mean and covariance are given by:

$$\mu_{c,N} = \frac{K_{c,0}\,\mu_{c,0} + N_c\,\bar{x}_c}{K_{c,0} + N_c} \tag{1}$$

$$\Sigma_{c,N} = \frac{\nu_{c,0}}{\nu_{c,0}+N_c}\,\Sigma_{c,0} + \frac{N_c}{\nu_{c,0}+N_c}\Bigg(\Sigma_{x,c} + \frac{K_{c,0}}{K_{c,0}+N_c}\,(\bar{x}_c-\mu_{c,0})(\bar{x}_c-\mu_{c,0})^\top\Bigg) \tag{2}$$

Larger N_c provides stronger evidence from the exposure talker and yields posterior category representations that more strongly reflect the recent input. Note that the update of covariance (Eq. 2) is also influenced by the updating of the prior mean.

3.5 Generalization models

Building on the learning model, we implemented three strategies for how listeners might deploy exposure-updated knowledge at test.

3.5.1 Talker-specific models

Since the experiment employed the same talker during training and test, listeners may apply the exposure-updated representation directly to test tokens. This talker-specific model is a straightforward application of the ideal adaptor: the posterior distributions estimated during exposure serve as the generative model for categorizing all test items. Under this model, learning is expressed uniformly regardless of the frequency structure of the test context. It therefore serves as the null hypothesis for expression gating: lexical frequency does not modulate the expression of learning, and high- and low-frequency words rely on the same updated model.

3.5.2 Mixture-of-expectations (MoE) model

At test, listeners may have uncertainty about the applicability of this updated generative model and partially fall back on the prior knowledge, interpolating between prior knowledge and the exposure-updated representation. In this model, categorization of a test token x_g reflects a weighted combination of the posterior predictive distribution (after exposure) and the prior predictive distribution (before exposure), governed by $w \in [0,1]$. Smaller w indicates greater reliance on the prior $(\mu_{c,0}, \Sigma_{c,0})$ and larger w greater reliance on the updated model $(\mu_{c,N}, \Sigma_{c,N})$. Listeners may apply the updated knowledge differently depending on word frequency. We therefore allow w to vary by frequency condition. We used the best-fitting talker-specific model as the updated model. Each term in the mixture is the posterior predictive distribution obtained by marginalizing over the category parameters, yielding multivariate t distributions:

$$\begin{aligned} p(x_g \mid c) &= \int p(x_g \mid \theta_c)\, p(\theta_c)\, d\theta_c \\ &= w\,\mathcal{T}\Bigg(\mathbf{x}_g \mid \boldsymbol{\mu}_N, \frac{\kappa_N+1}{\kappa_N(\nu_N-D+1)}\mathbf{S}_N,\, \nu_N-D+1\Bigg) \\ &\quad + (1-w)\,\mathcal{T}\Bigg(\mathbf{x}_g \mid \boldsymbol{\mu}_0, \frac{\kappa_0+1}{\kappa_0(\nu_0-D+1)}\mathbf{S}_0,\, \nu_0-D+1\Bigg) \end{aligned} \tag{3}$$

, where D is the number of acoustic dimensions (D=2). We did not mix the two Gaussian distributions directly. Instead, we computed category likelihoods under the corresponding prior/posterior predictive functions and then re-normalized the resulting probabilities after combination. w is separately fitted for high- and low- frequency words.

3.5.3 Hierarchical Bayesian model

Rather than falling back on the prior, listeners may still adopt the updated model but deploy it with greater uncertainty. We capture this by introducing a lower-level representation during test (parameterized by μ_g, Σ_g) that is allowed to deviate from the exposure-updated higher-level model (μ_N, Σ_N). The degree of deviation is governed by the hyperparameters κ_1 and ν_1. This formulation is identical to the baseline ideal adaptor, except that no additional exposure is provided. Operationally, the best-fitting posterior distribution in the talker-specific model is taken as the higher-level model. Since listeners knew that the same talker is speaking during test, the posterior mean of μ_g and Σ_g should be centered on μ_N and Σ_N, respectively. Given that $E[\Sigma_g] = \Sigma_N$ and $E[\mu_g] = \mu_N$, the test mean (μ_g) and covariance (Σ_g) are sampled from:

$$\mu_g \sim \mathcal{N}(\mu_N, \frac{1}{\kappa_N}\Sigma_g), \quad \kappa_1 = \lambda \tag{4}$$

$$\Sigma_g \sim \mathcal{IW}(S_g, \nu_1),\ S_g = (\nu_1 - D - 1)\,\Sigma_N,\ \nu_1 = \lambda + D + 1 \tag{5}$$

,where D is the number of acoustic dimensions (D=2).

For simplicity, we use a single pooling parameter λ to represent confidence in the updated model. Note that, to ensure the covariance update is always well-defined, we set $\nu_1 = \lambda + D + 1$. The resulting posterior predictive is a multivariate student-t distribution given as

$$\mathcal{T}\left(\boldsymbol{\mu}_N, \frac{\lambda+1}{\lambda+2}\Sigma_N, \lambda+2\right) \tag{6}$$

Crucially, λ governs the confidence with which the learned representation is deployed at test, not the degree of learning itself. Larger λ yields stronger pooling, reflecting more confident deployment of the exposure-updated model; smaller λ permits greater deviation, reflecting reduced confidence in applying the updated representation to the current test context. λ is fit for high and low frequency words separately.

3.6 Optimization and model selection

Model parameters were estimated by minimizing binary cross-entropy (BCE) loss. To ensure that model fitting was sensitive to the qualitative pattern across all frequency conditions, we computed BCE on the aggregated response proportions for each group × frequency configuration × step combination and then averaged across cells. For the talker-specific model and the hierarchical model, we set the upper and lower limit of κ, ν, and λ to range between 1 and 10,000. For MoE, w ranges between 0 and 1. Both MoE and HBM were built on the best-performing talker-specific model; their transfer parameters were the only additional free parameters.

To assess how well the models recovered the signature learning patterns, we also evaluated their predictions of log-odds differences between the biasing conditions (lowering/raising) and the control condition. For each frequency configuration f, we computed model–human mismatch in the exposure–control contrasts. Let $g \in \{\text{lowering}, \text{raising}, \text{control}\}$ index exposure group, s index continuum step, and q index word pair. We first averaged response proportions within each group across participants i (Eq. 7), then computed the log-odds contrast between each exposure group and the control group (Eq. 8). Finally, we computed RMSE between model-predicted and human log-odds contrasts (Eq. 9).

$$\bar{p}_{g,f,s,q} = \frac{1}{N_{g,f}} \sum_{i=1}^{N_{g,f}} p_{i,g,f,s,q} \tag{7}$$

$$\Delta_{d,f,s,q} = \text{logit}(\bar{p}_{d,f,s,q}) - \text{logit}(\bar{p}_{\text{control},f,s,q}) \tag{8}$$

where $d \in \{\text{lowering}, \text{raising}\}$.

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{d,f,s,q} \left(\Delta^{\text{human}}_{d,f,s,q} - \Delta^{\text{model}}_{d,f,s,q}\right)^2} \tag{9}$$

where M is the total number of exposure–control contrast × frequency-configuration × step × word-pair comparisons.

4 Simulation results

We present simulation results for the best-performing models. We evaluated model performance along two complementary dimensions. First, we report BCE in the raw probability space, which reflects global correspondence between model-predicted response probabilities and observed response proportions. Second, because the central empirical question concerns how exposure effects are expressed relative to the control condition, we computed a contrast-based RMSE over log-odds differences between each biasing condition and the corresponding control condition. Thus, BCE asks which model best captures the overall categorization pattern, whereas RMSE asks which model best captures the theoretically critical exposure-control contrast pattern. As shown in Table 1 the MoE model achieved the lowest BCE, indicating the best global probability-space fit. However, the hierarchical Bayesian model achieved the lowest contrast-based RMSE, indicating the closest recovery of the key generalization pattern. Because our primary question concerns selective expression of learning across frequency contexts, we focus below on the contrast-based comparison (Figure 2) while also reporting the global BCE metric. In the Appendix, Figure 6 and Figure 7 present model predictions in the raw probability space and predicted log-odds differences at each continuum step.

4.1 Talker-specific model

The best-performing talker-specific model ($\kappa = 27$ and $\nu = 10.1$) reproduces the main qualitative patterns in the human data (Figure 2): it predicts recalibration for /ɪ/-lowering, but not /ɛ/-raising in the HH condition, /ɪ/-lowering only in the HL condition and /ɛ/-raising only in the LH condition. These successes arise from the model's built-in

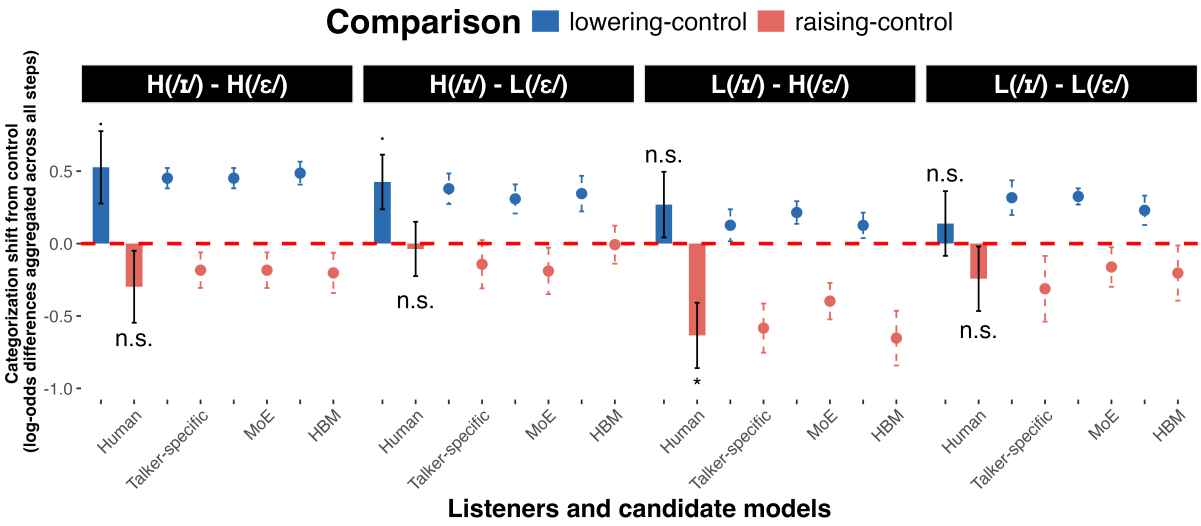

Figure 2: Log-odds differences between lowering/control and raising/control, averaged across continuum steps in each frequency condition. The red horizontal line indicates no difference between a biasing group and the control group. Positive values indicate more /ɪ/ responses than in the control group, whereas negative values indicate fewer /ɪ/ responses than in the control group. Bars indicate listeners' generalization effects. For listeners, statistical comparisons were based on post-hoc pairwise comparison derived from the linear mixed-effects model in the original experiment. P-values were corrected for multiple comparisons. Dots show the generalization effects of the best-performing models. Significance levels are coded as follows: . p < .10, * p < .05, ** p < .01, *** p < .001.

sensitivity to the acoustic properties of each test token—that is, from token-level acoustic similarity between exposure and test distributions. Importantly, the model fails to capture the LL condition: it predicts substantial learning effects in both biasing directions, whereas the corresponding effects in human data were smaller and not statistically reliable. This suggests that acoustic similarity alone accounts for most of the generalization pattern, but cannot explain why learning is not expressed when both lexical alternatives are low-frequency.

4.2 Mixture-of-expectations (MoE)

The best-performing MoE parameters were $w_h = 1$ and $w_l = 0.84$, indicating that the model relies entirely on the updated representation for high-frequency words, but places slightly more weight on the prior for low-frequency words. The MoE achieves the lowest BCE among the three models (Table 1), indicating the best global fit in probability space. However, as shown in Figure 2, the MoE model does not improve beyond the talker-specific model in capturing the theoretically critical patterns. This can also be confirmed by the largest RMSE among the three models. The model predicts a pronounced lowering effect in the LH condition, which is not observed in listeners' responses. It underestimates the degree of /ɛ/-raising in the LH condition. And crucially, it predicts robust /ɪ/-lowering in the LL condition, just as the talker-specific model does. Taken together, the MoE model predictions of learning effects do not improve beyond the learning model. The lack of improvement, particularly in LL, is likely because the empirical prior and version bias already bias the prior model toward /ɪ/, whereas listeners in LL do not exhibit an /ɪ/ bias. Consequently, lowering w_l does not improve model fit. This in turn helps explain why the best-performing model assigns a relatively higher value to w_l that preserves learning effects.

4.3 Hierarchical Bayesian model

The best performing hierarchical model parameters were $\lambda_h = 25.7$ and $\lambda_l = 17.1$, indicating higher confidence in the updated model for high-frequency words, and lower confidence for low-frequency words. As in Table 1, in global BCE, the hierarchical model performs worst among the three, indicating the worst performance among the models. The poorer predictive performance likely arises because we used a single parameter, λ, to govern uncertainty in both the posterior mean and covariance. In practice, this gives the model fewer degrees of freedom. It is possible that listeners have different degrees of confidence in the updated

estimates of the mean and covariance, and that allowing these to vary separately would yield better predictions.

Despite its higher BCE, the hierarchical model provides the best recovery of the condition-specific learning effects, as reflected in the lowest RMSE (Table 1). As shown in Figure 2, it comes closest to replicating the qualitative pattern in LL: the predicted /ɪ/-lowering effect is smaller than in the other two models, moving in the direction of the observed null effect. It also provides a better fit to the null raising effect in HL. Unlike the MoE, the hierarchical model attenuates learning not by mixing in a biased prior but by injecting additional uncertainty into the updated representation itself. This mechanism is not undermined by the /ɪ/ bias in the prior, which could be why it succeeds where the MoE does not.

We note that the hierarchical model's recovery of the LL pattern is approximate rather than exact—it reduces the predicted effect but does not fully eliminate it. The RMSE values are broadly consistent with the qualitative patterns visible in Figure 2: the hierarchical model shows the best recovery of condition-specific learning effects, while the talker-specific model and MoE perform comparably. Our goal here is not to identify the definitive account but to evaluate candidate mechanisms. The key finding is that among the three models, only the hierarchical model moves predictions in the right direction for the critical LL condition, and it does so through reduced expression confidence for low-frequency words—consistent with the expression-gating hypothesis.

Table 1: Best-fitting parameters and model performance. Learning parameters govern belief updating during exposure; transfer parameters govern expression of learning at test. MoE and HBM inherit the best-fitting learning parameters from the talker-specific model. BCE = mean binary cross-entropy (lower = better global fit). RMSE = root mean square error measuring the discrepancy between model-predicted and observed generalization effects (log-odds differences relative to control, averaged across continuum steps; lower is better)

Model	Learning	Transfer	BCE	RMSE
Talker-specific	$\kappa = 27$ $\nu = 10.1$	N/A	0.492	0.141
MoE	$\kappa = 27$ $\nu = 10.1$	$w_h = 1.00$ $w_l = 0.84$	**0.491**	0.146
HBM	$\kappa = 27$ $\nu = 10.1$	$\lambda_h = 25.7$ $\lambda_l = 17.1$	0.526	**0.121**

5 Discussions

This paper introduces a modeling framework that formalizes listeners' generalization behavior as perceptual inference under uncertainty. We applied it to LK26, a recent experiment in which listeners were exposed to shifted vowel pronunciations embedded in high-frequency words and then tested on minimal pairs varying in frequency structure (HH, HL, LH, LL). The key empirical finding is that identical learning—acquired from the same exposure items—was expressed selectively at test, depending on the frequency configuration of the test pairs.

Our modeling goal is to determine whether the talker-specific model's built-in sensitivity to acoustic similarity suffices, or whether an additional expression-gating mechanism modulated by lexical frequency is needed. We find that the talker-specific model already captures the generalization patterns in HH, HL, and LH, but overpredicts learning in LL. The hierarchical Bayesian model, which allows reduced confidence in deploying the learned representation for low-frequency words, best recovers the specific pattern across condition, as measured by the contrast-based RMSE.

Critically, the partial-transfer models, including the mixture-of-expectations and hierarchical models, do not introduce lexical frequency as an additional bias in categorization itself. Rather, building on the output of the learning model, they construct different generative models at test that vary in how tightly they are coupled to the exposure-updated representation. In this sense, these models formalize variability in the expression of learning rather than differences in the learning process itself.

Regarding model fitting and comparison, it is important to note that the models were not tailored to the observed empirical pattern. Although the MoE and HBM allowed transfer to vary by lexical frequency, the direction and magnitude of this effect were not imposed a priori. The fitted parameters could therefore have converged to no difference between high- and low-frequency contexts, or even to stronger transfer for low-frequency words. Moreover, because the talker-specific model does not include frequency-based gating, it could in principle have provided the best fit. Thus, the HBM's advantage on the contrast-based RMSE should be interpreted as an empirical outcome of

model fitting, rather than as a pattern built into the model by design. Moreover, the MoE also included frequency-specific parameters but did not best recover the contrast pattern, indicating that allowing parameters to vary by frequency was not sufficient by itself.

Our findings align with the predictions of usage-based models of phonology, which hold that listeners encode and retain fine-grained acoustic detail in memory, and that the robustness of these representations varies with frequency (Goldinger, 1996, 1998; Pierrehumbert, 2001; Wedel, 2006). Agent-based simulations of sound change have shown that—at least in some scenarios—high-frequency words tend to undergo change earlier or more rapidly than low-frequency words (Pierrehumbert, 2001; Todd et al., 2019). Our results suggest that similar stabilizing forces operate on a much shorter timescale within individual listeners: expression of short-term perceptual learning can be gated by lexical frequency, with lower-frequency words less affected by recently learned phonetic adjustments. If this perceptual conservatism is repeated over time, it may contribute to the frequency-conditioned trajectories observed in sound change. This is plausible given the structure of the lexicon: the majority of words in English are extremely rare (Mandelbrot, 1961; Piantadosi, 2014). It is sensible for listeners to treat pronunciations heard during brief exposure as relatively less relevant for predicting how such words should sound.

A related question is whether the lexical frequency of exposure items also shapes learning. LK26 did not test this, because all exposure items were high-frequency words. Recent work by Da Silva Vieira (2026) examined exposure lexical frequency and found no significant learning with low-frequency exposure items, raising the possibility that word frequency modulates the uptake of acoustic input (though see Koo et al. (2023) for contrasting evidence). However, differences in task, stimuli, and test design make direct comparison difficult. Future work should systematically cross exposure and test frequency.

Beyond the present within-talker case study, this framework can inform debates about cross-talker generalization, where competing accounts attribute transfer either to exposure to multiple talkers (Bradlow and Bent, 2008; Aoki and Zellou, 2025b) or to acoustic similarity (Xie and Myers, 2017; Xie and Jaeger, 2020; Jin et al., 2025) alone. The ideal adaptor captures both training–test similarity and cumulative exposure effects, while constrained models like MoE and HBM can isolate mechanisms beyond quantity or similarity alone.

Limitations

Several limitations should be noted. First, LK26 is, to our knowledge, the only experiment of its kind to compare generalization responses across words differing in lexical frequency. However, the different response pattern observed for LL trials may partly reflect properties of the stimulus manipulation rather than lexical frequency *per se*. In particular, the psychophysical increments between stimulus steps might not be identical across continua, as the test steps were generated through curve fitting based on categorization responses in the norming study. Moreover, both the test stimuli and the original recordings were produced by a single Canadian English speaker. It is therefore possible that the observed pattern reflects a combination of speaker-specific pronunciation, stimulus construction, and idiosyncrasies of the recruited participant sample. Further experiments are needed to determine whether the current findings can indeed be replicated.

Second, our model comparison should be interpreted with caution. We evaluated the best-fitting version of each candidate model using two complementary metrics: BCE as a measure of global fit in probability space, and contrast-based RMSE as a measure of how well each model recovered the signature exposure-control patterns. Although these comparisons are informative, differences in fit should not be taken as definitive evidence that the candidate mechanisms are uniquely identifiable from the present data. A stronger test would require simulation-based model recovery analyses, cross-validation, and related robustness checks (Wilson and Collins, 2019). Thus, the present comparison should be viewed as a theoretically motivated proof of concept rather than a definitive adjudication among mechanisms.

Acknowledgment

We thank Effie Kapnoula, Arthur Samuel, Clara Martin, Shawn Cummings, and others for their helpful comments on this work.

References

Jessica E. D. Alexander and Lynne C. Nygaard. 2019. Specificity and generalization in perceptual adapta-

tion to accented speech. *The Journal of the Acoustical Society of America*, 145(6):3382–3398.

Nicholas Aoki and Georgia Zellou. 2025a. When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception. *Glossa Psycholinguistics*, 4(1).

Nicholas B Aoki and Georgia Zellou. 2025b. Apparent talker variability and speaking style similarity can enhance comprehension of novel l2-accented talkers. *Language and Speech*, page 00238309251390505.

Melissa M. Baese-Berk, Ann R. Bradlow, and Beverly A. Wright. 2013. Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3):EL174–EL180.

Jeffrey S. Bowers, Nina Kazanina, and Nora Andermane. 2016. Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language*, 87:71–83.

Ann R. Bradlow and Tessa Bent. 2008. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

Kateřina Chládková, Václav Jonáš Podlipský, and Anastasia Chionidou. 2017. Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance*, 43(2):414–427.

Constance M. Clarke and Merrill F. Garrett. 2004. Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6):3647–3658.

Cynthia M Connine, Debra Titone, and Jian Wang. 1993. Auditory word recognition: extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1):81.

Marcelo Augusto Da Silva Vieira. 2026. Adaptation to communication constraints: prosodic flexibility in parkinson's disease and word frequency effect on auditory learning in young adults.

Frank Eisner and James M. Mcqueen. 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67(2):224–238.

Frank Eisner, Alissa Melinger, and Andrea Weber. 2013. Constraints on the Transfer of Perceptual Learning in Accented Speech. *Frontiers in Psychology*, 4:148.

Stephen D Goldinger. 1996. Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of experimental psychology: Learning, memory, and cognition*, 22(5):1166.

Stephen D Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2023. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, 130(4):977–1016.

Jordan Hosier and Klinton Bicknell. 2018. Generalization in Accent Adaptation as Hierarchical Bayesian Inference.

Kaori Idemaru and Lori L. Holt. 2011. Word Recognition Reflects Dimension-based Statistical Learning. *Journal of Experimental Psychology. Human Perception and Performance*, 37(6):1939–1956.

Alexandra Jesse and James M. McQueen. 2011. Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, 18(5):943–950.

Zhengyang Jin, Yuhao Zhu, and T Florian Jaeger. 2025. Latent speech representations learned through self-supervised learning predict listeners' generalization of adaptation across talkers. In *Proceedings of the annual meeting of the cognitive science society*, volume 47.

Charles Kemp, Andrew Perfors, and Joshua B. Tenenbaum. 2007. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321.

Dave F. Kleinschmidt and T. Florian Jaeger. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148–203.

Hahn Koo, Reiko Kataoka, Julia Thomas Swan, and Christina Y Tzeng. 2023. Effects of lexical frequency on the post-exposure magnitude of recalibration in lexically guided perceptual learning: An explorative analysis. *JASA Express Letters*, 3(8).

Tanya Kraljic and Arthur G. Samuel. 2006. Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, 13(2):262–268.

Wei Lai and Meredith Tamminga. 2024. Phonetics–phonology mapping in the generalization of perceptual learning. *Journal of Phonetics*, 103:101295.

Xinyu Leslie Liao and Yoonjung Kang. 2026. Directional bias, lexical competition, and item frequency effects in sound change: experimental evidence from perceptual learning. Submitted to Language.

Yiming Lu and Xin Xie. 2024. Modeling cue reweighting in dimension-based statistical learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Benoit Mandelbrot. 1961. On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 12:190–219.

Jessica Maye, Richard N. Aslin, and Michael K. Tanenhaus. 2008. The Weckud Wetch of the Wast: Lexical Adaptation to a Novel Accent. *Cognitive Science*, 32(3):543–562.

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(2):219–246.

Holger Mitterer, Yiya Chen, and Xiaolin Zhou. 2011. Phonological Abstraction in Processing Lexical-Tone Variation: Evidence From a Learning Paradigm. *Cognitive Science*, 35(1):184–197.

Holger Mitterer and Eva Reinisch. 2017. Surface forms trump underlying representations in functional generalisations in speech perception: the case of German devoiced stops. *Language, Cognition and Neuroscience*, 32(9):1133–1147.

Kuniko Nielsen and Colin Wilson. 2008. A hierarchical bayesian model of multi-level phonetic imitation. In *Proceedings of the 27th west coast conference on formal linguistics*, pages 335–343. Cascadilla Proceedings Project Los Angeles.

D Norris, James McQueen, and Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238.

Bozena Pajak, Klinton Bicknell, and Roger Levy. 2013. A model of generalization in distributional learning of phonetic categories.

Bozena Pajak, Alex B. Fine, Dave F. Kleinschmidt, and T. Florian Jaeger. 2016. Learning Additional Languages as Hierarchical Probabilistic Inference: Insights From First Language Processing. *Language Learning*, 66(4):900–944.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Janet B. Pierrehumbert. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee and Paul J. Hopper, editors, *Typological Studies in Language*, volume 45, pages 137–158. John Benjamins Publishing Company, Amsterdam.

Janet B Pierrehumbert. 2016. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2(1):33–52.

Eva Reinisch and Lori L. Holt. 2014. Lexically Guided Phonetic Retuning of Foreign-Accented Speech and Its Generalization. *Journal of experimental psychology. Human perception and performance*, 40(2):539–555.

Irene B R Smith, Morgan Sonderegger, and The SPADE Consortium. 2025. Patterns of pre-nasal allophony across dialects of English: A multi-corpus study of the /ɪ/-/ɛ/ contrast. OSF preprint.

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.

Maryann Tan, Xin Xie, and T Florian Jaeger. 2021. Using rational models to interpret the results of experiments on accent adaptation. *Frontiers in Psychology*, 12:676271.

Simon Todd, Janet B. Pierrehumbert, and Jennifer Hay. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185:1–20.

Andrew B Wedel. 2006. Exemplar models, evolution and language change. *The Linguistic Review*, 23(3).

Robert C Wilson and Anne GE Collins. 2019. Ten simple rules for the computational modeling of behavioral data. *elife*, 8:e49547.

Xin Xie and T. Florian Jaeger. 2020. Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5):3322–3347.

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

Xin Xie, Linda Liu, and T Florian Jaeger. 2021. Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, 150(11):e22.

Xin Xie and Emily B. Myers. 2017. Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, 97:30–46.

Weijie Xu and Richard Futrell. 2024. A hierarchical Bayesian model for syntactic priming. *arXiv preprint*. ArXiv:2405.15964 [cs].

A Additional figures

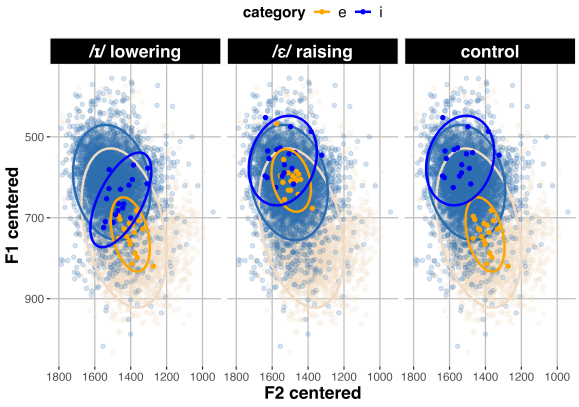

Figure 3: Distribution of exposure trials (shown in solid dots), against the Canadian corpus (in faded dots)

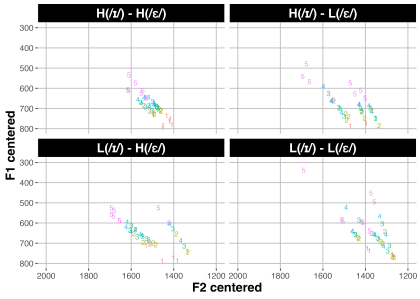

Figure 4: Test items across the four frequency combinations. Values 1–5 indicate the proportion of /ɪ/ responses estimated in a pilot norming study, corresponding to 0.01, 0.25, 0.50, 0.75, and 0.99, respectively.

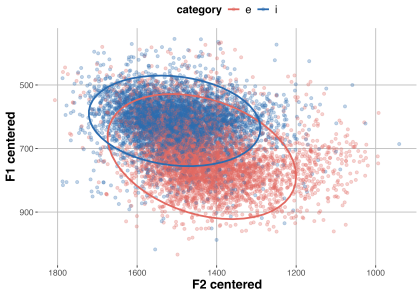

Figure 5: The Canadian corpus. Each dot represents an item containing /ɪ/ or /ɛ/. F1 and F2 were centered relative to each talker's cue mean, and the grand mean across all talkers was then added back. Ellipses indicate the 95% confidence region of the Gaussian distributions fitted across all talkers..

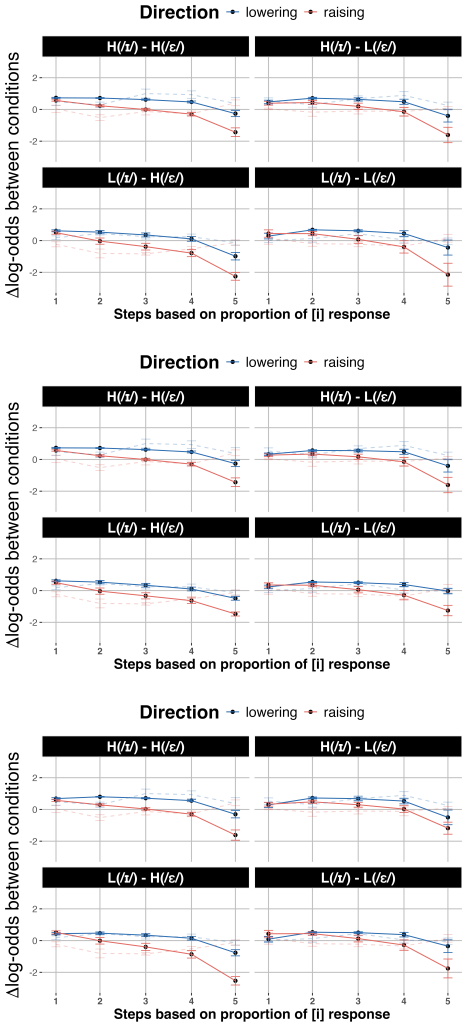

Figure 7: Predictions of the best-performing talker-specific model (top), mixture of expectation model (middle), and hierarchical Bayesian model (bottom) in the log-odds space. To highlight learning, the x-axis shows continuum steps and the y-axis shows log-odds differences between /ɪ/-lowering and control and between /ɛ/-raising and control. Solid lines show model predictions, and faded lines reproduce the human data for comparison. Error bars indicate ±1 standard error of the mean.

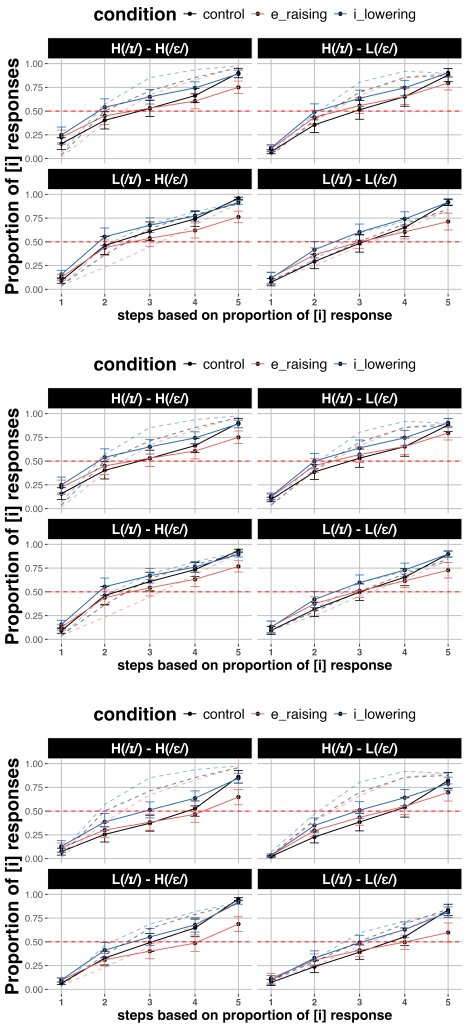

Figure 6: Predictions of the best-performing talker-specific model (left), mixture-of-expectations model (middle), and hierarchical Bayesian model (right) in the raw probability space. Solid lines show model predictions, while faded dashed lines show mean proportion of /ɪ/ responses. To improve visual clarity, only the model error bars (± 1 standard error of the mean) are shown here.