Bayesian Semiparametric Longitudinal Drift-Diffusion Mixed Models for Tone Learning in Adults

Giorgio Paulon[1] (giorgio.paulon@utexas.edu)
Fernando Llanos[2,3] (f.llanos@pitt.edu)
Bharath Chandrasekaran[3](b.chandra@pitt.edu)
Abhra Sarkar[1] (abhra.sarkar@utexas.edu)

[1]Department of Statistics and Data Sciences,
University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA

[2]Department of Linguistics,
University of Texas at Austin,
305 East 23rd Street B5100, Austin, TX 78712, USA

[3]Department of Communication Science and Disorders,
University of Pittsburgh,
4028 Forbes Tower, Pittsburgh, PA 15260, USA

Abstract

Understanding how adult humans learn non-native speech categories such as tone information has shed novel insights into the mechanisms underlying experience-dependent brain plasticity. Scientists have traditionally examined these questions using longitudinal learning experiments under a multi-category decision making paradigm. Drift-diffusion processes are popular in such contexts for their ability to mimic underlying neural mechanisms. Motivated by these problems, we develop a novel Bayesian semiparametric inverse Gaussian drift-diffusion mixed model for multi-alternative decision making in longitudinal settings. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the method's empirical performances through synthetic experiments. Applied to our motivating longitudinal tone learning study, the method provides novel insights into how the biologically interpretable model parameters evolve with learning, differ between input-response tone combinations, and differ between well and poorly performing adults.

Key Words: Auditory category/tone learning, Auditory neuroscience, B-splines, Drift-diffusion models, (Factorial) hidden Markov models, Functional models, Inverse Gaussian distributions, Local clustering, Longitudinal mixed models, Perceptual decision making, Speech learning, Wiener processes

Short/Running Title: Longitudinal Drift-Diffusion Mixed Models

Corresponding Author: Abhra Sarkar (abhra.sarkar@utexas.edu)

1 Introduction

Understanding the cognitive and biological mechanisms underlying our ability to learn new speech categories in adulthood constitute important questions in auditory neuroscience. Recent studies have demonstrated that adults are capable of learning features of a second language to a high degree of efficiency, demonstrating that age need not always constrain language learning abilities. The inherent dynamic complexities underlying learning in adulthood are not yet well understood but are being studied through extensive ongoing research.

The research reported here is motivated particularly by experiments on the acquisition of Mandarin tones by native speakers of English. Native speech categories are acquired during the first year of life, within a so-called phonetic sensitivity period. There is a greater neural commitment to native-language speech sounds, and this commitment may preclude the learning of novel speech categories in adulthood (Johnson and Newport, 1989; Iverson *et al.*, 2003). In Mandarin Chinese, there are four tone categories that systematically change word meaning, similar to consonants and vowels in English. These tones are, however, linguistically irrelevant in English. English native speakers thus struggle to distinguish the four tones and generalize their differences (Wang *et al.*, 1999; Chandrasekaran *et al.*, 2010; Maddox and Chandrasekaran, 2014). In laboratory settings, combining exposure to perceptually variable tones with trial-by-trial corrective feedback can improve tone categorization skills within a few hundred trials. Reaching a native like proficiency, however, may take several sessions of training (Xie *et al.*, 2017; Reetzke *et al.*, 2018). The perceptual and sensory representation of Mandarin tones gets fundamentally refined over the course of this learning period (Feng *et al.*, 2019). Understanding this longitudinal evolution is critical to assess the cognitive dynamics of speech category learning. The statistical challenge is to make this assessment indirectly from behavioral data on tone categorization responses and response times.

To this end, we identify the Mandarin tone categorization problem with the broader class of problems of multi-category decision making under perceptual stimuli (Smith and Ratcliff, 2004; Heekeren *et al.*, 2004; Gold and Shadlen, 2007; Schall, 2001; Purcell, 2013; Glimcher and Fehr, 2013). In such contexts, drift-diffusion processes are popular models for behavioral accuracies and response times as they mimic the accumulation of sensory evidence in favor of different decision alternatives in the human brain (Ratcliff, 1978; Ratcliff *et al.*, 2016). The existing literature on drift-diffusion models is substantive (Smith and Vickers, 1988; Ratcliff and Rouder, 1998; Ratcliff and McKoon, 2008). These classical methods, as well as their recent adaptations using reinforcement learning based ideas (Fontanesi *et al.*, 2019; Pedersen *et al.*, 2017; Peters and D'Esposito, 2020), are, however, heavily focused on the two category case with a single latent diffusion process and two boundaries, one for each of the two decision alternatives. This is despite the fact that humans often are required

to learn more than two categories at once. For example, English has 14 vowels and 24 consonant phonemes; Mandarin has four tone categories, etc. The joint likelihood of accuracies and response times under models with a single diffusion process is mathematically complex and computationally expensive (Navarro and Fuss, 2009; Tuerlinckx, 2004; Tuerlinckx *et al.*, 2001). Inference in such models is thus often based on approximations of the likelihood (Vandekerckhove and Tuerlinckx, 2007), or on the conditional likelihood of the response times, conditioned on the decisions (Vandekerckhove *et al.*, 2008). Multi-category drift-diffusion models with separate latent processes, one for each decision category and simultaneously at play, have been developed to address some of the limitations (Usher and McClelland, 2001; Brown and Heathcote, 2008; Leite and Ratcliff, 2010; Dufau *et al.*, 2012; Kim *et al.*, 2017), but the relevant literature remains sparse and focused only on simple static designs.

Learning to distinguish Mandarin tones or, more generally, to make categorization decisions is, however, a dynamic process, driven by continuous and nuanced perceptual adjustments in our brain and behavior over time. The existing simple static models are thus severely limited in their ability to capture the true inherent complexities, including assessing the biologically relevant changes that take place over the learning period. Principled statistical approaches to multi-category dynamic drift-diffusion mixed effects models, that appropriately accommodate fixed effects of experimental factors as well as random effects due to subjects, are therefore highly needed but present daunting methodological and computational challenges.

In this article, we address these challenges by developing a novel biologically interpretable flexible Bayesian semiparametric inverse Gaussian drift-diffusion mixed model for studying multi-alternative perceptual decision making processes in longitudinal settings.

Our construction proceeds by characterizing the accumulation of evidence for different input-response tone combinations by associated independent Wiener diffusion processes, resulting in an inverse Gaussian distribution based joint probability model for the final response tone and the associated response time. To adapt this to a longitudinal mixed model setting, we then assume the model parameters to comprise input-response tone specific fixed effects and subject specific random effects, modeling them both by mixtures of locally supported B-spline bases (de Boor, 1978; Eilers and Marx, 1996) spanning the length of the longitudinal experiment. Both these effects are thus allowed to evolve flexibly as smooth functions over the training period (Ramsay and Silverman, 2007; Morris, 2015; Wang *et al.*, 2016) as the participants get more experience and training in their assigned decision tasks.

Dependence in the fixed effects model spline coefficients across adjacent temporal regions is induced via hidden Markov models (HMMs) (McDonald and Zucchini, 1997; Rabiner, 1989; Frühwirth-Schnatter, 2006; Cappé *et al.*, 2005), one for each input-response tone combination but all sharing a common state space, as well as a novel

smoothness inducing Markovian prior on the core spline coefficients. The HMMs, adapted in such novel ways, induce a local clustering of the fixed effects spline coefficients associated with different input-response tone combinations, in effect, allowing us to assess local similarities and differences between the corresponding parameter trajectories in different learning phases.

This ability to infer local similarities and differences in the cognitive dynamics is theoretically and practically relevant for tone learning applications. The underlying mechanisms are expected to be very similar when the participants are first introduced to the tones; differences may appear as they get better at identifying the tones as some tones may be easier to identify than others in this stage; these differences may start to disappear again in later stages of the experiment as the participants become highly proficient in identifying all the different tones. As for individual heterogeneity, neural measures of sensory encoding information collected prior to the learning task show no clear individual differences, even though the process of learning itself results in good and poor learners (Reetzke *et al.*, 2018).

The literature on longitudinal data analysis models is enormous. See, for example, books by Diggle *et al.* (2002); Singer *et al.* (2003); Fitzmaurice *et al.* (2008) and the references therein. Bayesian methods for longitudinal data have also been extensively developed (Daniels and Pourahmadi, 2002; Chib and Hamilton, 2002; Li *et al.*, 2010; Müller *et al.*, 2013; Quintana *et al.*, 2016, etc.). The problem of modeling locally clustered effects has, however, not garnered much attention. We can only mention Petrone *et al.* (2009); Nguyen and Gelfand (2011, 2014), all of which were designed primarily for normally distributed functional data with continuous covariates. It is not clear how these approaches can be adapted to our problem.

Overall, our proposed method takes the existing state-of-the-art many significant steps forward, including (a) introducing a novel biologically interpretable class of multi-category inverse Gaussian drift-diffusion models for decision making, (b) accommodating fixed effects of perceptual stimuli and random effects due to subject specific heterogeneity in such models in a statistically principled manner, (c) adapting these models to longitudinal study designs, studying the temporal evolution of the underlying process parameters as the subjects get trained and experienced in their assigned decision tasks, (d) allowing the process parameters to be locally clustered, enabling the assessment of their similarities and differences in various learning stages.

Applied to our motivating tone learning data set, the proposed method provides many novel insights into the cognitive dynamics, allowing us to answer important scientific questions completely outside the scope of the previously existing literature. These include a detailed understanding of how biologically significant model parameters, that systematically relate to the underlying neural processes, evolve and interplay to enable gradual longitudinal learning in the participants, how similar or different these parameters are across different input and output tone combinations in different

learning phases, how these processes differ between a good and a bad learner, etc.

The rest of this article is organized as follows. Section 2 provides additional background on tone learning and drift-diffusion models. Section 3 details our novel locally varying longitudinal drift-diffusion mixed model. Section 4 outlines computational challenges and solution strategies. Section 5 presents the results of the proposed method applied to tone learning data. Section 6 contains concluding remarks. Substantive additional details, including a Markov chain Monte Carlo (MCMC) based posterior inference algorithm and results of simulation experiments, are presented in the supplementary materials.

2 Behavioral Data and Scientific Background

The behavioral data set that motivated our research comes from an intensive multi-day longitudinal speech category training study reported previously in Reetzke *et al.* (2018). In this study, $n = 20$ native English-speaking adults were trained to categorize Mandarin Chinese syllables into lexical tone categories as a function of their pitch contour. Mandarin Chinese has four syllabic pitch contours or tones that are used to convey different lexical meanings. For example, in Mandarin Chinese, the syllable 'ma' can be interpreted as 'mother', 'hemp', 'horse', or 'scold' depending on whether is pronounced with a high-level (T1), low-rising (T2), low-dipping (T3), or high-falling (T4) tone, respectively. The stimuli consisted of these tones pronounced by four native Mandarin speakers. The trials were administered in homogeneous blocks. Each block comprised 40 categorization trials for 40 different speech exemplars, corresponding to different combinations of speakers, syllables, and input tones. Participants were trained across several days, with five blocks on each day. On each categorization trial, participants indicated the tone category they heard via a button press on a computer keyboard. Following the button press, the participants were given corrective feedback ('Correct/Incorrect') on a computer screen which was previously shown to be more effective in enhancing learning compared to full feedback (for example, 'Incorrect, that was a category 2') (Chandrasekaran *et al.*, 2014). Individual categorization performance was monitored across training sessions until each participant achieved and maintained accuracy levels comparable to that of native speakers of Mandarin.

The data consist of the tone responses and the associated response times for different input tones for the 20 participants. We focus here on the first two days of training (10 blocks in total) as they exhibited the steepest improvement in learning as well as the most striking individual differences relative to any other collection of blocks (Figure 1). In that sense, they provide an optimal longitudinal frame to assess the effects of learning on decision making variables.

Tone learning can be viewed from a broader perspective of multi-category decision making tasks, and hence can be studied using computational models developed for

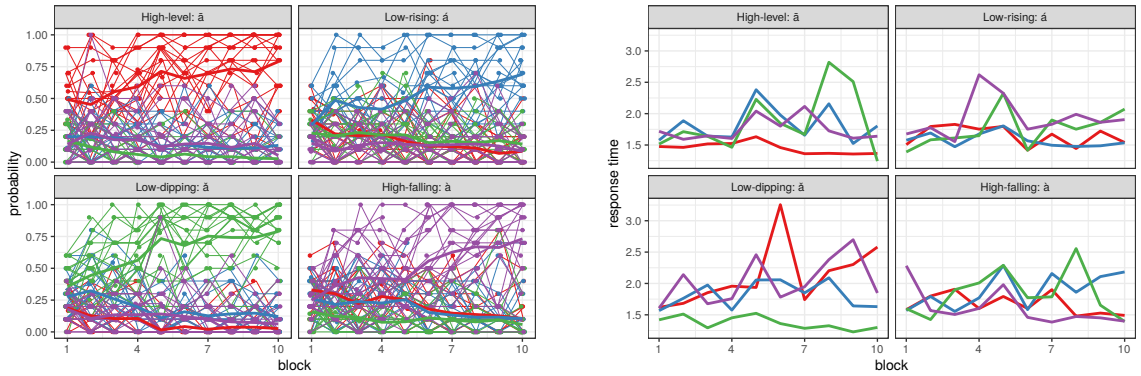

Figure 1: Left panel: Proportions of times an input tone was classified into different tone categories by different subjects. The thick line represents the average performance across subjects. Right panel: Associated response times averaged across subjects for clarity. In both panels, high-level tone responses are shown in red; low-rising in blue; low-dipping in green; and high-falling in purple.

such tasks. We present here a brief nontechnical overview of how these models relate to the underlying neurobiology. Mathematical details and developments are deferred to Section 3.

In a typical multi-category decision task, the brain accumulates sensory evidence in order to make a categorical decision. This accumulation process is reflected in increasing firing rate at local neural populations associated with alternative decisions. A decision is taken when neural activity in one of these populations crosses a particular threshold level. The decision category that is finally chosen is the one whose decision threshold is crossed first (Gold and Shadlen, 2007; Brody and Hanks, 2016).

Changes in evidence accumulation rates and decision thresholds can be induced by task difficulty, neurostimulation, and/or individual differences in cognitive function (Cavanagh *et al.*, 2011; Ding and Gold, 2013). Decision-making is also regulated by demands on both speed and accuracy as a function of the task (Bogacz *et al.*, 2010; Milosavljevic *et al.*, 2010). The overall learning accuracies ('Correct/Incorrect' response proportions) in our data set were previously analyzed in Paulon *et al.* (2019) using a binary logistic longitudinal mixed model. In a different context, Craigmile *et al.* (2010) had developed a model for response times. Separate models for accuracies and response times cannot, however, provide a meaningful interpretation of the speed-accuracy trade-off.

An excellent basis for jointly modeling accuracies and response times is obtained by imitating the underlying neural evidence accumulation mechanisms via latent drift-diffusion processes racing toward their respective boundaries, the process reaching its boundary first producing the final observed decision and the time taken to reach this boundary giving the associated response time (Figure 2) (Usher and McClelland, 2001). The drift and the boundary parameters jointly explain the dynamics of choice, including the speed-accuracy trade-off. Broadly speaking, decision thresholds remain-

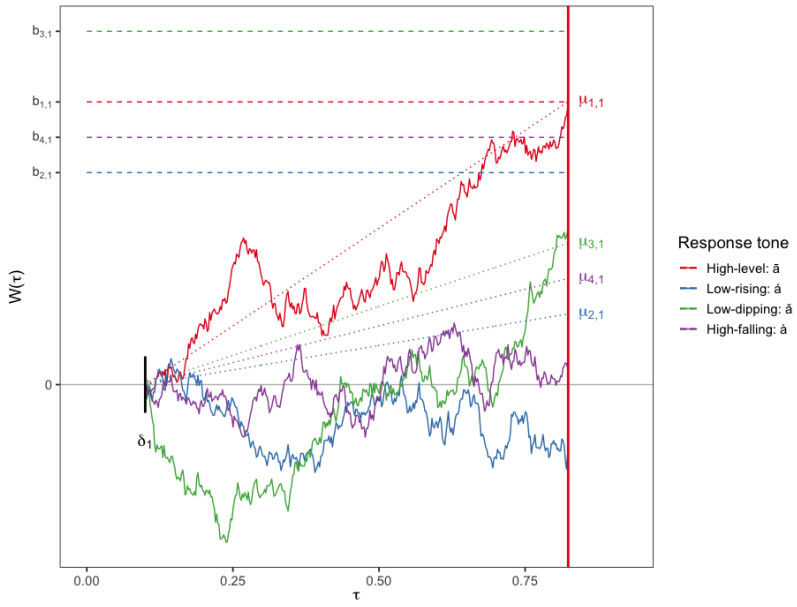

Figure 2: Drift-diffusion model for perceptual decision making. After an initial δ_s amount of time required to encode an input signal s, the evidence in favor of a response category d accumulates according to a Wiener diffusion process with drift $\mu_{d,s}$. The decision d is eventually taken if the underlying process is the first to reach its decision boundary $b_{d,s}$. Here we illustrate a tone learning trial with input tone T1 ($s = 1$) that was eventually correctly identified. Section 2 provides additional neurobiological background. Section 3 provides additional mathematical details.

ing fixed, higher drift rates lead to faster and more accurate responses; for fixed drift rates, higher decision thresholds, on the other hand, increase response times as well as inaccuracies.

In our motivating tone learning experiment, we are interested in understanding the evolution and interplay of the drift and the boundary parameters behind the improved tone identification performances over training. Importantly, as was also discussed in the introduction, we are not just interested in estimating the overall trajectories of these parameters but also how they might differ between different input-response tone combinations locally in different longitudinal stages of the experiment. Additional interest lies in assessing subject level heterogeneity in these parameter trajectories, including particularly how they differ between good versus bad learners.

3 Longitudinal Drift-Diffusion Mixed Models

The basic Wiener diffusion process can be specified as $W(\tau) = \mu\tau + \sigma B(\tau)$, where $B(\tau)$ is the standard Brownian motion, μ is the drift rate, and σ is the diffusion coefficient

(Cox and Miller, 1965; Ross *et al.*, 1996). The process has independent normally distributed increments, that is, $\Delta W(\tau) = \{W(\tau + \Delta\tau) - W(\tau)\} \sim \text{Normal}(\mu\Delta\tau, \sigma^2\Delta\tau)$, independently from $W(\tau)$. The first passage time of crossing a threshold b, $\tau = \inf\{\tau' : W(0) = 0, W(\tau') \geq b\}$, is then distributed according to an inverse Gaussian distribution (Whitmore and Seshadri, 1987; Chhikara, 1988; Lu, 1995) with density

$$f(\tau \mid \mu, \sigma^2, b) = \tfrac{b}{\sqrt{2\pi\sigma^2}}\tau^{-3/2}\exp\left\{-\tfrac{(b-\mu\tau)^2}{2\sigma^2\tau}\right\}, \quad b > 0, \quad \mu > 0, \quad \sigma^2 > 0.$$

With $\boldsymbol{\theta} = (\mu, \sigma, b)^{\mathrm{T}}$, we have $\mathbb{E}(\tau \mid \boldsymbol{\theta}) = b/\mu$ and $\text{var}(\tau \mid \boldsymbol{\theta}) = b\sigma^2/\mu^3$.

Given perceptual stimuli and a set of decision choices, the neurons in the brain accumulate evidence in favor of the different alternatives. Modeling this behavior using Wiener processes with unit variances, assuming that a response is given when the decision threshold for one of the options is crossed, a probability model for the time τ_d to reach the threshold for the d^{th} decision category under the influence of the s^{th} stimulus is obtained as

$$f(\tau_d \mid \delta_s, \mu_{d,s}, 1, b_{d,s}) = \tfrac{b_{d,s}}{\sqrt{2\pi}}(\tau_d - \delta_s)^{-3/2}\exp\left[-\tfrac{\{b_{d,s}-\mu_{d,s}(\tau_d-\delta_s)\}^2}{2(\tau_d-\delta_s)}\right], \tag{1}$$

where $\mu_{d,s}$ denotes the rate of accumulation of evidence, $b_{d,s}$ the decision boundaries, and δ_s an offset representing the collective time required to encode the s^{th} signal before evidence accumulation begins, the time to press a computer key to record a response after a decision is reached, etc. (Figure 2). We now let $\boldsymbol{\theta}_{d,s} = (\delta_s, \mu_{d,s}, b_{d,s})^{\mathrm{T}}$. Since a decision d is reached at response time τ if the corresponding threshold is crossed first, that is when $\{\tau = \tau_d\} \cap_{d' \neq d} \{\tau_{d'} > \tau_d\}$, we have $d = \arg\min \tau_{d'}$. Assuming simultaneous accumulation of evidence for all decision categories, modeled by independent Wiener processes, and termination when the threshold for the observed decision category d is reached, the joint distribution of (d, τ) is thus given by

$$f(d, \tau \mid s, \boldsymbol{\theta}) = g(\tau \mid \boldsymbol{\theta}_{d,s}) \textstyle\prod_{d' \neq d}\{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}. \tag{2}$$

where, to distinguish from the generic notation f, we now use $g(\cdot \mid \boldsymbol{\theta})$ and $G(\cdot \mid \boldsymbol{\theta})$ to denote, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of an inverse Gaussian distribution, as defined in (1). We refer to model (2) as the inverse Gaussian drift-diffusion model.

The marginal distribution of the response times τ under the influence of stimulus s is then obtained as

$$f(\tau \mid s, \boldsymbol{\theta}) = \textstyle\sum_d g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d}\{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}. \tag{3}$$

The marginal probability of taking decision d under the influence of stimulus s is likewise obtained as

$$f(d \mid s, \boldsymbol{\theta}) = \int_{\delta_s}^{\infty} g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}\, d\tau. \tag{4}$$

Interestingly, model (4) is similar to traditional multinomial probit/logit regression models (Borooah, 2002; Agresti, 2018) except that the latent variables are now inverse Gaussian distributed as opposed to being normal or extreme-value distributed, and the observed category is associated with the minimum of the latent variables in contrast to being identified with the maximum of the latent variables.

In an interesting recent work, Kunkel *et al.* (2019) have also used an inverse Gaussian distribution based hierarchical Bayesian model for decision making, albeit in a simpler binary category case, focusing primarily on individual level models with no mechanism to assess population level effects or their dynamic complexities.

For our motivating longitudinal tone learning experiment described in Section 2, for $i \in \{1, \ldots, n = 20\}, \ell \in \{1, \ldots, L = 40\}, t \in \{1, \ldots, T = 10\}$, let $s_{i,\ell,t}$ denote the input tone for the i^{th} individual in the ℓ^{th} trial in block t. Likewise, let $d_{i,\ell,t}$ and $\tau_{i,\ell,t}$ denote, respectively, the selected Mandarin tone and the time taken to reach the corresponding threshold by the i^{th} individual in the ℓ^{th} trial in block t. We now have

$$g\{\tau_{i,\ell,t} \mid s_{i,\ell,t} = s, \boldsymbol{\theta}_{d,s}^{(i)}(t)\} = \frac{b_{d,s}^{(i)}(t)}{\sqrt{2\pi}(\tau_{i,\ell,t} - \delta_s^{(i)})^{3/2}} \exp\left[-\frac{\{b_{d,s}^{(i)}(t) - \mu_{d,s}^{(i)}(t)(\tau_{i,\ell,t} - \delta_s^{(i)})\}^2}{2(\tau_{i,\ell,t} - \delta_s^{(i)})}\right]. \tag{5}$$

The drift rates $\mu_{d,s}^{(i)}(t)$ and the decision boundaries $b_{d,s}^{(i)}(t)$ now also vary with the blocks t. In addition, we accommodate random effects by allowing $\delta_s^{(i)}$, $\mu_{d,s}^{(i)}(t)$ and $b_{d,s}^{(i)}(t)$ to also depend on the subject index i. We let $y_{i,\ell,t} = (d_{i,\ell,t}, \tau_{i,\ell,t})$, $\mathbf{y} = \{y_{i,\ell,t}\}_{i,\ell,t}$, and $d_0 = 4$ be the number of possible decision categories (T1, T2, T3, T4). The likelihood function of our longitudinal drift-diffusion mixed model thus takes the form

$$L(\mathbf{y} \mid \mathbf{s}, \boldsymbol{\theta}) = \prod_{d=1}^{d_0} \prod_{s=1}^{d_0} \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left(g\{\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d,s}^{(i)}(t)\} \prod_{d' \neq d} [1 - G\{\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d',s}^{(i)}(t)\}] \right)^{1\{d_{i,\ell,t} = d, s_{i,\ell,t} = s\}}.$$

3.1 Modeling the Offsets

The offset parameters $\delta_s^{(i)}$, we recall, signify the times spent on encoding the different input tones, the time to press computer keys to record the responses, etc., and hence are not directly relevant to the actual decision making processes. These parameters are thus biologically not very interesting but may still vary between individuals and have an important effect on the estimates of drift rates and boundaries (Teichert *et al.*, 2016). We thus let them vary between input stimuli and participants but assume them to remain stable across blocks as in (5).

We assign uniform priors on $\delta_s^{(i)} \sim \text{Unif}(0, \delta_{s,i,\max})$, where $\delta_{s,i,\max}$ is the minimum of all response times under stimulus s for individual i, that is, $\delta_{s,i,\max} = \min_{\{(\ell,t):s_{i,\ell,t}=s\}} \tau_{i,\ell,t}$.

3.2 Modeling the Drifts and the Boundaries

Our modeling efforts concentrate henceforth on flexibly characterizing the longitudinal evolution of the mixed effects parameters $\mu_{d,s}^{(i)}(t), b_{d,s}^{(i)}(t)$. Variations in these parameters over training blocks explain perceptual learning in the participants. Variations across participants, on the other hand, explain their performance heterogeneity. Following the discussion in the introduction, of particular interest are the local similarities and differences between these parameters for different input-response tone combinations (d, s) in different learning phases.

To this end, we propose essentially identical modeling strategies for $\mu_{d,s}^{(i)}(t)$ and $b_{d,s}^{(i)}(t)$. For ease of exposition avoiding unnecessary repetition, we describe below only these common strategies using simplified generic notations. With $x = (d, s) \in \mathcal{X} = \{(1,1), (1,2), \ldots, (4,4)\} \equiv \{1, 2, \ldots, x_{\max}\}$, $x_{\max} = 4 \times 4$, succinctly representing the input-response tone combinations and, with some abuse, $\theta_x^{(i)}(t)$ being a generic for $\mu_{d,s}^{(i)}(t)$ and $b_{d,s}^{(i)}(t)$, we let

$$\theta_x^{(i)}(t) = \exp\{f_x(t) + u_x^{(i)}(t)\}, \qquad u_x^{(i)}(t) \sim f_u\{u_x^{(i)}(t)\}. \tag{6}$$

The exponentiation in (6) enforces positivity constraints; $f_x(t)$ and $u_x^{(i)}(t)$ denote, respectively, additive fixed and random effects components in the exponential scale; f_u denotes the underlying random effects distribution. When needed, the fixed and random effects components for the drifts and the boundaries, as well as associated parameters and hyper-parameters, will be distinguished by reintroducing the subscripts as $f_{\mu,x}(t)$, $f_{b,x}(t)$, $u_{\mu,x}^{(i)}(t), u_{b,x}^{(i)}(t)$ etc. To further simplify notation, generic data recording experimental blocks in $\{1, \ldots, T\}$ as well as other generic time points in $[1, T]$ will both be denoted by t. Likewise, generic input-response tone combinations as well as their particular values will both be denoted by x and so forth.

We model the components $f_x(t)$ and $u_x^{(i)}(t)$, and hence $\theta_x^{(i)}(t)$, to all be smoothly varying functions over $t \in [1, T]$. A functional approach is not strictly necessary if inference is restricted only to the T data recording blocks blocks $t \in \{1, \ldots, T\}$. Learning may, however, be viewed as a continuous process - the brain synthesizes information from relevant past experiences even when not being actively engaged in actual decision making. A functional approach to modeling $f_x(t)$ and $u_x^{(i)}(t)$ for any $t \in [1, T]$, not just the experimental blocks $t \in \{1, \ldots, T\}$, thus facilitates parameter interpretability. A functional approach is also practically convenient in characterizing smoothly varying longitudinal parameter trajectories.

In modeling the fixed effects components $f_x(t)$, we are not only interested in characterizing their overall trajectories over time t for different input-response combinations $x = (d, s)$ but also how they might vary locally between different values of x in different learning stages. Compared to the fixed effects, we have to, however, rely on much less data to estimate the random effects $u_x^{(i)}(t)$ for different $x = (d, s)$ and different participant i, especially for $d \neq s$ toward later stages of the experiment when most participants identify the input tones with high accuracies. Our models and inferential goals for the random effects $u_x^{(i)}(t)$ will therefore be relatively modest.

3.2.1 Locally Varying Functional Fixed Effects

We now propose a novel approach to modeling the latent functions $f_x(t)$ using basis decomposition methods that allow them to smoothly vary with the blocks t while also depending locally on the indexing variable x. To begin with, we let

$$f_x(t) = \sum_{k=1}^{K} \beta_k^{(x)} B_k(t), \tag{7}$$

where $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^{\mathrm{T}}$ are a set of known locally supported basis functions spanning $[1, T]$, $\boldsymbol{\beta}^{(x)} = (\beta_1^{(x)}, \ldots, \beta_K^{(x)})^{\mathrm{T}}$ are associated unknown coefficients to be estimated from the data. In this article, we use quadratic B-spline bases with knot points coinciding with the block locations. B-splines are non-negative, continuous and have desirable local supports (Figure 3). Mixtures of B-splines are highly flexible (de Boor, 1978). Allowing the $\beta_k^{(x)}$'s to flexibly vary with x, the model can accommodate widely different shapes for different input-response tone combinations.

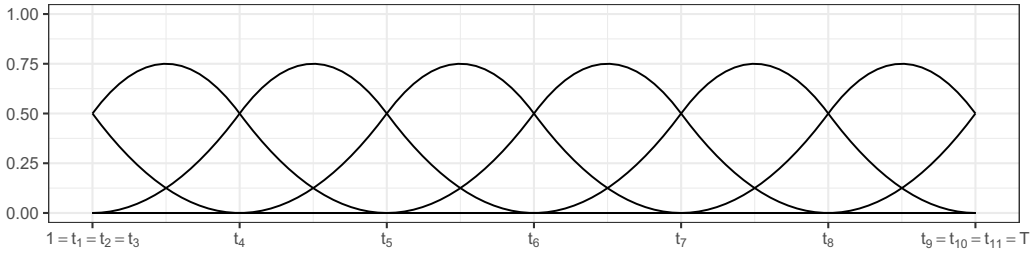

Figure 3: Plot of 8 quadratic B-splines on an interval $[1, T]$ defined by 11 knot points that divide $[1, T]$ into $K = 6$ equal subintervals.

It is difficult to assess how similar or different these functions are using such unstructured models. One potential solution is to cluster the spline coefficients $\boldsymbol{\beta}^{(x)}$ associated with different input-response tone combinations x. If, for example, $\boldsymbol{\beta}^{(x_1)} = \boldsymbol{\beta}^{(x_2)}$ for two combinations x_1 and x_2, then we have $f_{x_1}(t) = f_{x_2}(t)$ for all t.

Such global clustering of all elements of $\boldsymbol{\beta}^{(x)}$ together does not, however, allow us to straightforwardly assess the local similarities and differences between these functions

in different learning phases. To induce a desirable local cluster inducing mechanism, we introduce a set of latent variables $z_k^{(x)}$ for each input-response tone combination x with a shared state space $\mathcal{X}$, and associated core coefficients $\beta^\star_{k,z}$ and let

$$(\beta_k^{(x)} \mid z_k^{(x)} = z_k) = \beta^\star_{k,z_k}, \quad \text{implying} \quad \{f_x(t) \mid z_k^{(x)} = z_k, k = 1, \ldots, K\} = \textstyle\sum_{k=1}^K \beta^\star_{k,z_k} B_k(t). \tag{8}$$

The set of B-spline coefficients to be estimated at the k^{th} location now comprises the $\beta^\star_{k,z_k}$'s that are indexed by $z_k^{(x)} = z_k$ at that location k. When $z_k^{(x_1)} = z_k^{(x_2)}$ for two different levels x_1 and x_2 of x, we have $\beta_k^{(x_1)} = \beta_k^{(x_2)}$ and the implied functions $f_{x_1}(t)$ and $f_{x_2}(t)$ will tend to be similar at location k. Indeed, for quadratic B-splines with knots at the blocks $\{1, \ldots, T\}$, $f_{x_1}(t)$ and $f_{x_2}(t)$ will be exactly equal at block t when $z_t^{(x_1)} = z_t^{(x_2)}$ and $z_{t+1}^{(x_1)} = z_{t+1}^{(x_2)}$.

In theory, we could use B-splines of other small degrees as they all enjoy local support properties. With linear splines, however, smoothness becomes harder to control, and with cubic splines, three latent variables would be needed to determine the cluster configuration at each block t. We found quadratic B-splines to be a good compromise between the two for modeling smoothly varying curves while also maintaining easy interpretability of the latent variables.

Letting $\mathcal{Z}_k = \{z_k : z_k^{(x)} = z_k \text{ for some } x \in \mathcal{X}\}$, the case $|\mathcal{Z}_k| = 1$ then characterizes the scenario when the the spline coefficients for all input-response tone combinations x are the same at location k. On the other end, when $|\mathcal{Z}_k| = x_{\max} = 4 \times 4$, the spline coefficients are all different for different x at location k. In our tone learning application, $|\mathcal{Z}_k|$ tend to be much smaller than $x_{\max}$ uniformly for all k and the restricted support $z_k^{(x)} \in \{1, \ldots, z_{\max}\} \subset \mathcal{X}$ with $z_{\max} = 8 < x_{max} = 16$ will suffice.

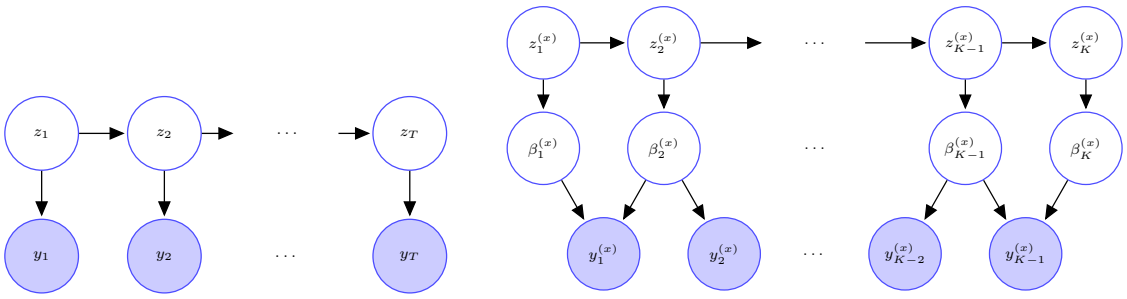

Figure 4: Left panel: Graph of a conventional HMM. Right panel: Graph of our proposed functional HMM model (8) with quadratic B-splines (Figure 3) with knots points coinciding with the data recording time blocks ($T = K - 1$).

We model the temporal evolution of the latent local cluster indicators $z_k^{(x)}, k = 1, \ldots, K$, using hidden Markov models (HMMs) (Figure 4). We consider two types of dynamics for the latent states corresponding to correct (C) and incorrect (I) identification of the tones. That is,

$$(z_k^{(d,s)} \mid z_{k-1}^{(d,s)} = z_{k-1}) \sim \text{Mult}(\pi^{(C)}_{z_{k-1},1}, \ldots, \pi^{(C)}_{z_{k-1},z_{\max}}) \quad \text{when } d = s,$$
$$(z_k^{(d,s)} \mid z_{k-1}^{(d,s)} = z_{k-1}) \sim \text{Mult}(\pi^{(I)}_{z_{k-1},1}, \ldots, \pi^{(I)}_{z_{k-1},z_{\max}}) \quad \text{when } d \neq s.$$

The latent cluster inducing variables $z_k^{(x)}$'s are shared between $f_{\mu,x}(t)$ and $f_{b,x}(t)$, reducing computational complexities while also facilitating model interpretability. We assign Dirichlet priors on the transition probabilities

$$\boldsymbol{\pi}_z^{(C)} = (\pi^{(C)}_{z,1}, \ldots, \pi^{(C)}_{z,z_{\max}})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(C)}/z_{\max}, \ldots, \alpha^{(C)}/z_{\max}) \quad \text{with} \quad \alpha^{(C)} \sim \text{Ga}(a_\alpha, b_\alpha),$$
$$\boldsymbol{\pi}_z^{(I)} = (\pi^{(I)}_{z,1}, \ldots, \pi^{(I)}_{z,z_{\max}})^{\mathrm{T}} \sim \text{Dir}(\alpha^{(I)}/z_{\max}, \ldots, \alpha^{(I)}/z_{\max}) \quad \text{with} \quad \alpha^{(I)} \sim \text{Ga}(a_\alpha, b_\alpha).$$

We next consider priors for the atoms $\beta^\star_{k,z_k}$. Conditional on the $z_k^{(x)}$'s and the coefficients at the previous locations, for $k = 2, \ldots, K$, we construct the priors sequentially as

$$\beta^\star_{k,z_k} \sim \begin{cases} \prod_{\{z^{(x)}_{k-1}:\ x \in \mathfrak{X}_k^{(z_k)}\}} \text{Normal}\left(\beta^\star_{k-1,z^{(x)}_{k-1}}, \sigma^2_{\beta,1}\right) & \text{if } |\mathfrak{X}_k^{(z_k)}| > 0, \\ \text{Normal}(\mu_{\beta,0}, \sigma^2_{\beta,0}) & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathfrak{X}_k^{(z_k)} = \{x : z_k^{(x)} = z_k\}$ is the set of values of x that, at the location k, are assigned the label z_k. In constructing the prior in this manner, we center the core coefficients around the ones that are 'expressed' at the previous location (Figure 5), penalizing their first order differences. The coefficients that are not associated with any levels of x are assigned a normal prior with a large variance $\sigma^2_{\beta,0}$. The initial coefficients are assigned non-informative flat priors as $\beta^\star_{1,z_k} \sim 1$. Additional illustrations on these smoothness inducing priors on the core coefficients can be found in Section S.2 of the supplementary materials.

The smoothness of the curves is controlled by the parameter $\sigma^2_{\beta,1}$ and is assigned a prior, allowing it to be informed by the data. We let

$$\sigma^2_{\beta,1} \sim \text{C}^+(0, 1),$$

where $\text{C}^+(a, b)$ denotes a half-Cauchy distribution (Gelman, 2006; Polson and Scott, 2012) with location parameter a and scale parameter b. The half-Cauchy distribution, which attains its mode at zero, is capable of capturing strong smoothness, while also having heavy tails, thus being capable of capturing wiggly functions. The choice of the scale hyper-parameter is discussed in Section S.5.1 in the supplementary materials.

Importantly, although our basic building blocks for the fixed effects components comprise conventional HMMs, one for each input-response tone combination

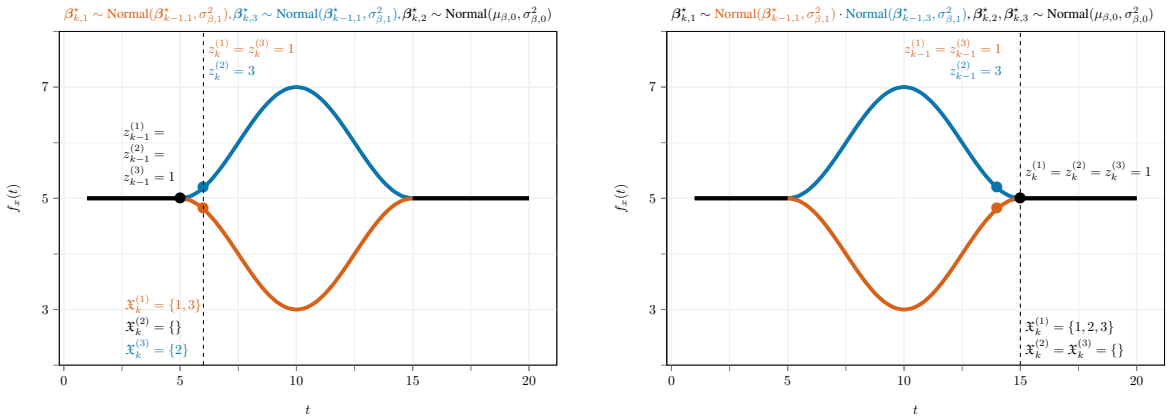

Figure 5: An illustration of the prior on the spline core coefficients $\beta^{\star}_{k,z_k}$ at location k (marked by the dashed vertical lines) in the fixed effects model developed in Section 3.2.1 for a synthetic scenario with $x \in \{1, 2, 3\}$, where the curves corresponding to the three levels of x are initially equal, the curves for $x = 1, 3$ (in red) and $x = 2$ (in blue) then diverge at $t = 6$, merging back again at $t = 15$.

$x = (d, s)$, for any input tone s, all four latent variables $z_k^{(1,s)}, z_k^{(2,s)}, z_k^{(3,s)}, z_k^{(4,s)}$ simultaneously appear in equation (2). For each input tone, the graph for our tone learning model (Figure 6 and Figure S.6 in the supplementary materials) thus resembles a factorial HMM (Ghahramani and Jordan, 1997, fHMM) with four hidden layers. In the posterior, a latent state $z_k^{(d,s)}$ is thus informed by all responses generated under the tone s, not just the subset corresponding to $x = (d, s)$. This has important consequences for posterior inference, as we discuss in Section 4.

3.2.2 Locally Varying Functional Random Effects

We now focus on flexibly modeling the functional random effects components. For reasons outlined before Section 3.2.1, estimating $u_x^{(i)}(t)$ for each different x is a challenging task. For any participant, the random effects for correct and incorrect identification of the tones may, however, be expected to be on the opposite sides of the corresponding population level curves. Taking a middle path, we thus allow different random effects $u_C^{(i)}(t)$ and $u_I^{(i)}(t)$ for correct (C) and incorrect (I) identifications, respectively, as

$$u_{d,s}^{(i)}(t) = u_C^{(i)}(t) \quad \text{when } d = s, \qquad u_{d,s}^{(i)}(t) = u_I^{(i)}(t) \quad \text{when } d \neq s.$$

We adopt a common strategy to model both $u_C^{(i)}(t)$ and $u_I^{(i)}(t)$. Suppressing the subscripts to simplify notation and avoid repetition, we model the time-varying random effects components $u^{(i)}(t)$ as

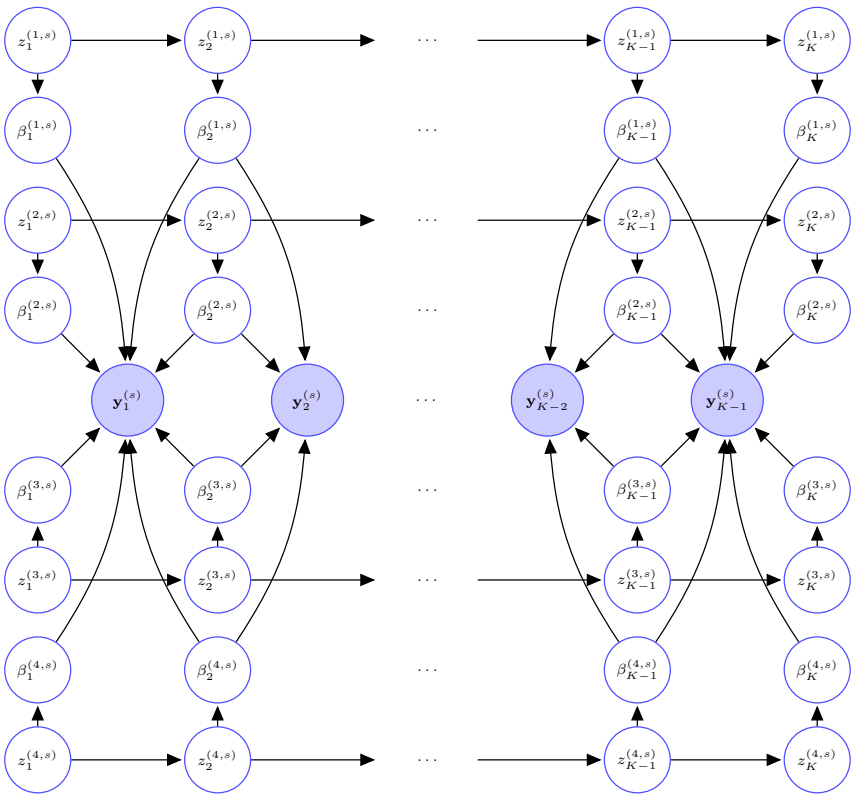

Figure 6: Graph of the proposed fixed effects model for tone learning.

$$
\begin{aligned}
u^{(i)}(t) &= \textstyle\sum_{k=1}^{K} \beta_{k,u}^{(i)} B_k(t), \\
\boldsymbol{\beta}_u^{(i)} &\sim \mathrm{MVN}_K\{\mathbf{0}, (\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P}_u)^{-1}\},
\end{aligned}
\tag{10}
$$

where $\boldsymbol{\beta}_u^{(i)} = (\beta_{1,u}^{(i)}, \ldots, \beta_{K,u}^{(i)})^{\mathrm{T}}$ are subject-specific spline coefficients, $\mathrm{MVN}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a K dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We choose $\mathbf{P}_u = \mathbf{D}_u^{\mathrm{T}}\mathbf{D}_u$, where the $(K-1) \times K$ matrix $\mathbf{D}_u$ is such that $\mathbf{D}_u\boldsymbol{\beta}_u^{(i)}$ computes the first order differences in $\boldsymbol{\beta}_u^{(i)}$. The model thus penalizes $\sum_{k=1}^{K}(\nabla\beta_{k,u}^{(i)})^2 = \boldsymbol{\beta}_u^{(i)\mathrm{T}}\mathbf{P}_u\boldsymbol{\beta}_u^{(i)}$, the sum of squares of first order differences in $\boldsymbol{\beta}_u^{(i)}$ (Eilers and Marx, 1996). The random effects variance parameter $\sigma_{u,s}^2$ models the smoothness of the random effects curves, smaller $\sigma_{u,s}^2$ inducing smoother $u^{(i)}(t)$'s. Additional variations from the constant zero curve are explained by $\sigma_{u,a}^2$ (Figure 7). The absence of random effects is signified by the limiting case $\sigma_{u,s}^2 = \sigma_{u,a}^2 = 0$. We assign half-Cauchy priors on the variance parameters as

$$
\sigma_{u,s}^2 \sim \mathrm{C}^+(0,1), \qquad \sigma_{u,a}^2 \sim \mathrm{C}^+(0,1).
$$

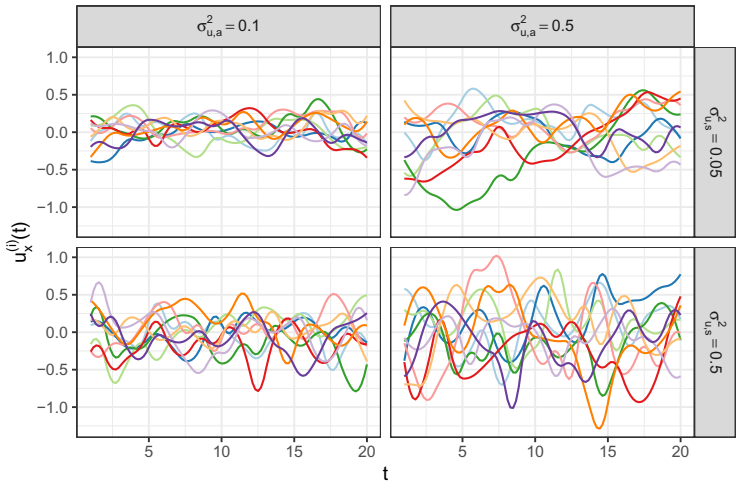

Figure 7: An illustration of the functional random effects model proposed in Section 3.2.2. Each panel shows a collection of 10 random draws from the random effects distribution for a combination of values of $(\sigma^2_{u,s}, \sigma^2_{u,a})$.

Modeled in the same space of quadratic B-splines, the fixed and the random effects curves thus share similar smoothness properties. Having different smoothness controlling parameters, they are, however, allowed to have different smoothness levels. A similar approach, but with additional assumptions on the covariance matrix of the random effects, has previously been developed in Guo (2002). To our knowledge, model (10) for the random effects is thus also novel to the literature.

Integrating out the random effects, the corresponding population level parameters $\theta_x(t)$ are obtained as

$$\theta_x(t) = \int \exp\{f_x(t) + u_x^{(i)}(t)\} f_u\{u_x^{(i)}(t)\} du_x^{(i)}(t) = \exp\left[f_x(t) + \frac{\text{var}\{u_x^{(i)}(t)\}}{2}\right].$$

4 Posterior Inference

Posterior inference for conventional HMMs can generally be based on samples drawn from the posterior using dynamic message passing MCMC algorithms (Rabiner, 1989; Scott, 2002). The nonstandard inverse Gaussian likelihood and the fHMM type model structure of our proposed longitudinal drift-diffusion mixed model, however, bring in significant additional complexities. We adapt recent advances in MCMC algorithms for discrete spaces (Neal, 2003; Van Gael *et al.*, 2008; Titsias and Yau, 2014; Zanella, 2019) in novel non-trivial ways, designing locally informative slice sampling moves that carefully exploits the conditional independence relationships encoded in

the model to overcome the computational challenges. Due to space constraints, the details are deferred to Section S.5 in the supplementary materials.

5 Application to Tone Categorization Data

In this section, we discuss the results produced by our method applied to the tone category learning data described in Section 2. Our primary inference goals, we recall, include understanding systematic longitudinal variations in perceptual categorization decision as the participants get better at identifying the four Mandarin tones with there being some additional interests in assessing individual specific trajectories, especially how they differ between good and bad learners.

Figure 8 shows the posterior mean trajectories and associated 90% credible intervals for the boundaries $b_{d,s}(t)$ and the drift rates $\mu_{d,s}(t)$ estimated by our method for different combinations of (d, s). Figure 9 reports the estimated posterior probabilities of each of the $\binom{4}{2} = 6$ pairs of success $(d = s)$ parameters to cluster together in different blocks. Figure S.16 in the supplementary materials additionally presents the drift curves for successful identifications $(d = s)$ superimposed on each other. These results suggest that after an initial learning phase, where the underlying processes are all similar across all input tones, there are two main learning groups. Two of the tones {T1, T3} seem to be easier to learn, as the corresponding drift parameters are larger, and tones {T2, T4} are more challenging. These findings are corroborated by empirical evidence and have significant biological relevance. The similarity groups of the mandarin tones are in fact {T1, T3}, which are characterized by the height of the pitch, and tones {T2, T4}, which are characterized by the direction of the pitch and are more challenging to learn. Tone T3, in particular, has a unique 'dipping' pitch pattern that is rarely encountered in English (Song *et al.*, 2008), and therefore is easier to categorize. Our proposed method allows similar inferential questions to be answered for the drift parameters corresponding to misclassifications, as well as for all the boundary parameters. The misclassification drift curves are mostly similar to each other, although some minor local differences can be found. Notable exceptions are $\mu_{1,3}(t)$ and $\mu_{3,1}(t)$ which are significantly smaller than all other drifts after the third block. As the participants get trained and experienced, for input tone T1, evidence in favor of tone T3 is thus collected more slowly compared to evidence in favor of T2 and T4, and vice versa. Likewise, while the boundary curve estimates mostly remain constant over the training blocks and similar to each other, $b_{1,3}(t)$ and $b_{3,1}(t)$ again differ from the rest and actually increase over the blocks. As the participants get trained and experienced, more evidence in favor of tone T3 is thus needed to misclassify tone T1 as tone T3 and vice versa. These suggest that, as the participants get trained and experienced, tones T1 and T3 become harder to misclassify for one another.

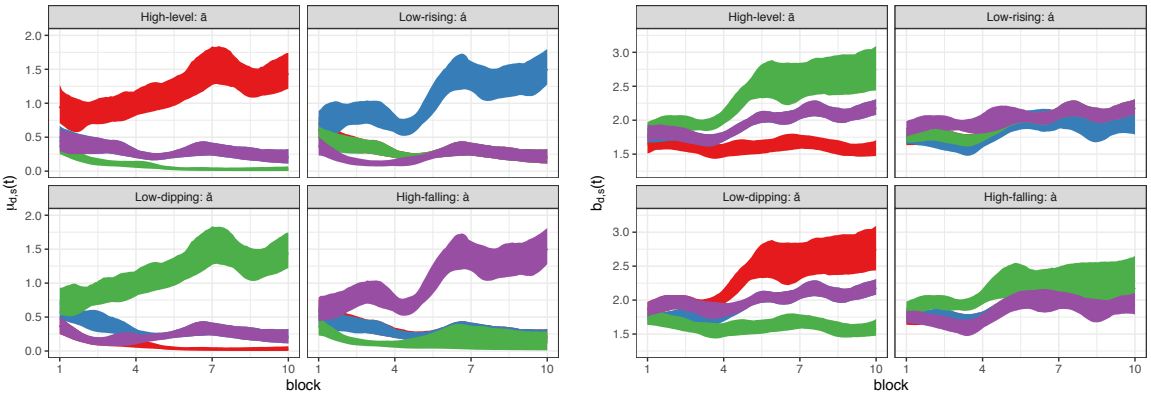

Figure 8: Results for tone learning data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ (left panel) and boundaries $b_{d,s}(t)$ (right panel) for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

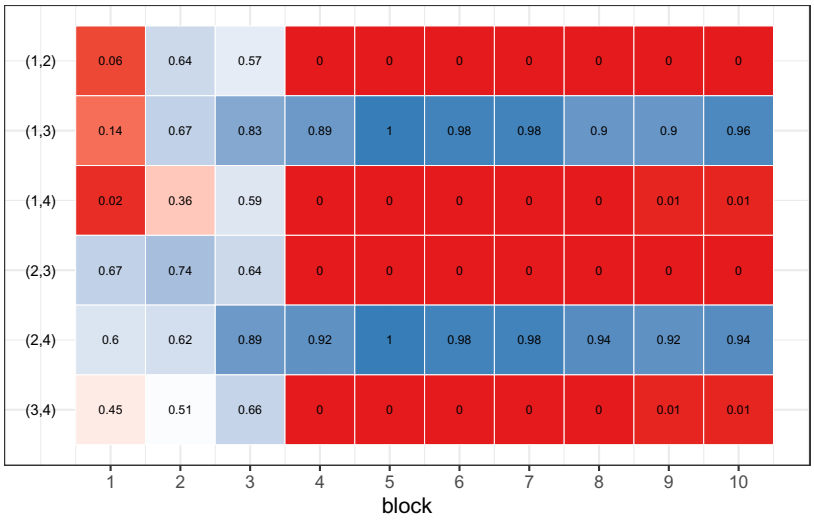

Figure 9: Results for tone learning data: Pairwise posterior co-clustering probabilities of the parameter trajectories for successful identification ($d = s$) of different input tones in different learning phases. The estimated posterior probability of $(\mu_{2,2}, b_{2,2})$ and $(\mu_{3,3}, b_{3,3})$ being clustered together, and hence being equal, in the 3^{th} block is thus 0.74, as shown in row $(2, 3)$ and column 3. Equivalently, the estimated posterior probability of $(\mu_{2,2}, b_{2,2})$ and $(\mu_{3,3}, b_{3,3})$ being different in the 3^{th} block is 0.26.

Importantly, our proposed drift-diffusion mixed model not only allows population level inference about the underlying processes but also allows us to assess individual specific parameter trajectories. Figure 10 shows the posterior mean trajectories and associated 90% credible intervals for the drift rates $\mu_{s,d}^{(i)}$ and the boundaries $b_{s,d}^{(i)}$ estimated by our method for the different success combinations of (d, s) for two par-

ticipants - the one with the best accuracy averaged across all blocks, and the one with the worst accuracy averaged across all blocks. These results suggest significant individual specific heterogeneity. Importantly, the differences in the performances can again be explained mostly by differences in the drift trajectories. For the well performing participant, the drift trajectories increase rapidly with the training blocks before plateauing down around block 6 at which stage the participant has already attained native-like proficiency. For the poorly performing candidate on the other hand, the drift trajectories remain approximately constant across all 10 blocks.

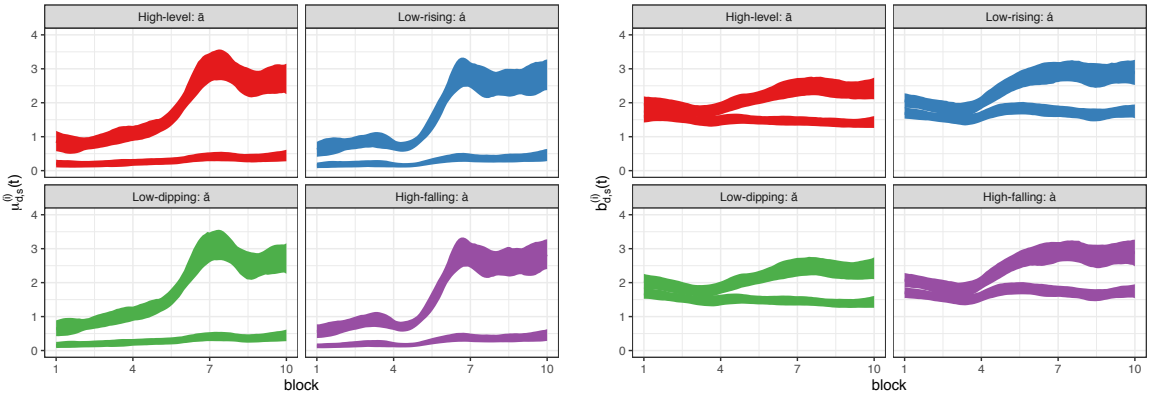

Figure 10: Results for tone learning data: Estimated posterior mean trajectories for individual specific drifts $\mu_{d,s}^{(i)}(t) = \exp\{f_{\mu,d,s}(t) + u_{\mu,C}^{(i)}(t)\}$ (left panel) and boundaries $b_{d,s}^{(i)}(t) = \exp\{f_{b,d,s}(t) + u_{b,C}^{(i)}(t)\}$ (right panel) for successful identification ($d = s$) for two different participants - one performing well (dotted line) and one performing poorly (dashed line). The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

We compare the performance of our method with that of the linear ballistic accumulator (LBA) model (Brown and Heathcote, 2008). Similar to our model, the LBA uses independent evidence accumulators starting at δ that continue until a response threshold b is reached. The accumulator that first reaches the boundary corresponds to the decision outcome, and the time taken to reach this decision boundary is the observed response time. The LBA model, however, assumes that the evidence accumulates linearly at the rate μ, reaching the boundary b precisely at time $\tau = b/\mu$. Unlike in drift-diffusion models, where trial-by-trial variability is explained by stochastically different diffusion paths, the LBA model explains trial-by-trial variability assuming the slopes μ for different trials to be drawn from a Normal$(m_{d,s}, v_{d,s})$ distribution. (Figure S.9 in the supplementary materials).

The literature on LBA models has many serious limitations. The normality assumption on the slopes μ clearly does not satisfy any non-negativity constraints. Existing LBA models are also limited in their use of a common boundary b_s for all decision categories d. There is also no principled way to incorporate systematic stim-

ulus and decision category specific fixed or individual specific random effects into the LBA model. Existing literature is also limited to static settings, there is no mechanism to estimate smoothly varying longitudinal parameter trajectories as the participants get trained and experienced in their decision tasks. In our implementation, we thus fitted the LBA model separately for each block. Finally, the likelihood function of the LBA model is non-convex in the parameters. Parameter estimation based on optimization of the likelihood function is thus fraught with convergence issues. We used the `rtdists` package (Singmann *et al.*, 2019) in `R`, using several random initializations and tracking the objective function to ensure convergence. A more detailed review of the LBA model can be found in Section S.7 of the supplementary materials.

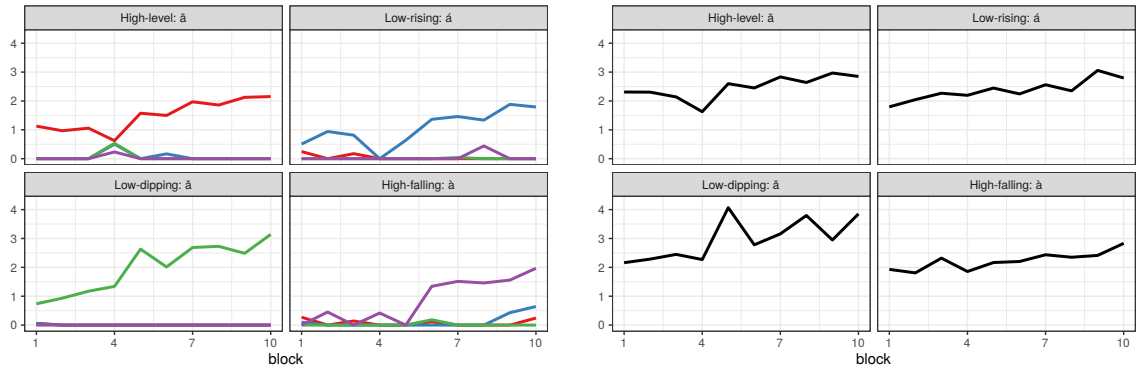

Figure 11: Results for tone learning data: Left: Estimated mean slopes $m_{d,s,t}$ for the LBA model. Right: Estimated boundaries $b_{s,t}$ for the LBA model. In the left panel, $m_{d,s,t}$'s for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

Results produced by the LBA model applied to our motivating tone-learning data are reported in Figure 11. Owing to the limitations discussed above, the inference we make with such models is very limited. For instance, only non-smooth population level estimates are available, individual specific trajectories can not be assessed, etc. Some of our findings can, however, be confirmed by the LBA method. For example, looking at the drift parameter estimates, one can see that tone T3 is consistently associated with larger drifts. As was also seen in the estimates returned by our method, tones {T2, T4} have similar values for the drift and the boundary parameters. Except such general overall findings, the LBA model, however, can not answer scientific questions related to the dynamics of category learning with fine detail.

Our method, on the other hand, provides a biologically interpretable, statistically principled approach to accommodate fixed effects of input stimuli and decision categories as well as random subject specific heterogeneity, allows MCMC algorithm based efficient estimation of longitudinally smoothly evolving parameter trajectories, borrowing information across sample subgroups, participants as well as adjacent time stamps through many layers of hierarchy. Crucially, building on a novel local cluster

inducing mechanism, our method also allows automated assessment of local similarities and differences in the parameter trajectories in very fine detail as the participants get trained and experienced in their decision tasks.

On the scientific side, the detailed insights obtained here point toward interesting and novel hypotheses about learning. For example, we demonstrate that a difference in drift rates, associated with the speed of sensory evidence accumulation, is critical in determining good vs poor learners. Evidence thresholds, on the other hand, remain relatively stable over training blocks as well as across participants. Recent studies have shown that the process of evidence accumulation can be selectively targeted by brain stimulation (Van der Groen *et al.*, 2018). Novel tone learning studies are currently being designed to test if such neurostimulation primarily improves the drift rates but not the evidence thresholds.

On the practical side, the insights obtained above can have important implications for developing advanced training regimens in language learning platforms used by millions of adults. Due to poor understanding of the temporal dynamics of learning, especially in multi-category learning problems, current training regimens are neither time adaptive nor individualized. Similar to personalized medicine, next-generation speech training paradigms seek to optimize and individualize training to reduce vast inter-individual differences in learning success (Wong *et al.*, 2017; Birdsong, 2004). With our ability to assess detailed longitudinal confusion patterns, we can set up efficient training paradigms that can change the dynamics of learning in specific ways. For example, learners may generally benefit from introducing greater variability in pitch height that allows them to shift their focus on pitch direction and hence can reduce disparities in tone confusions like that between T2 and T4; poor learners may additionally benefit from 'perceptual fading' - beginning with easy tones like {T1,T3} and making the training more challenging afterward with the introduction of tones like {T2,T4}; etc. As mentioned before, non-invasive and safe brain stimulation approaches like transcranial random noise stimulation and vagus nerve stimulation can be leveraged to selectively improve the process of sensory accumulation that could enhance the performance in poor learners.

6 Discussion

Summary: In this article, we proposed a novel longitudinal drift-diffusion mixed model for perceptual decision making, allowing the underlying mechanisms to be similar or different at different longitudinal stages. Our research was motivated primarily by auditory neuroscience experiments where scientists are interested in understanding how the decision making mechanisms evolve as the participants get more training in the decision tasks. Our model was built on a novel statistical framework for longitudinal data that exploited local support properties of B-spline bases and (factorial)

HMMs to allow automated assessment of local similarities and differences in the underlying parameter trajectories.

Application to our motivating tone categorization experiments provided interesting novel insights into the underlying learning mechanisms. Notably, we discovered that the improvements and the local variations in tone categorization performance can be explained mostly by variations in the underlying drift parameters while the boundaries mostly remain constant. We also discovered local groupings among the underlying parameter curves in various phases of the learning experiments, how they differ between well and poorly performing participants etc. Such inferences were outside the scope of the previously existing literature.

Methodological extensions: Methodological extensions and topics of our ongoing research include adapting the proposed models to time constrained learning experiments, developing nested models to capture the dynamics within the blocks, accommodating sleep induced overnight 'consolidation' effects, fully developing the inverse-probit model (4) for accuracies introduced in Section 3, etc.

Broader scientific impact: The proposed approach, we believe, takes the existing literature on drift-diffusion decision making models many significant steps forward, enabling neuroscientists to study the longitudinal behavior of biologically interpretable model parameters in much finer detail than what previous methods could achieve.

As reported in Section 5, the findings of our motivating speech learning experiment help formulate interesting novel scientific hypotheses about speech learning. The findings are also practically highly significant in providing exciting opportunities for developing time adaptive and individualized training regimens for language learning.

Efficient estimation of group and individual level trajectories also open exciting avenues for potential adaptations in clinical settings, especially in conjunction with simultaneously performed imaging studies.

Finally, the scope of proposed method is also not restricted to auditory neuroscience problems but the approach can be readily applied to study decision making mechanisms in other areas of neuroscience as well.

Supplementary Materials

Supplementary materials present substantive additional details. These include brief reviews of fHMMs, B-splines, locally informed Hamming ball samplers, the linear ballistic accumulator model, etc. to make the article relatively self-contained. The supplementary materials also discuss the choice of hyper-parameters for our model, the MCMC algorithm used to sample from the posterior of our model and its convergence diagnostics. The supplementary materials also present simulation studies and a comparison with a reduced model further illustrating the efficacy and the advantages of our proposed method. In separate files, the supplementary materials additionally

include the tone categorization data set described in Section 2 and analyzed in Section 5, audio recordings of the four input Mandarin tones, and `R` programs implementing the longitudinal drift-diffusion mixed model developed in this article.

Acknowledgments

We thank the editor, Dr. Heping Zhang, for comments leading to a significantly improved version of the initial manuscript. We also thank Dr. Peter Mueller, Dr. Mario Peruggia, Dr. Rachel Reetzke and Dr. Tobias Teichert for helpful discussions on the research presented here.

Funding

This work was supported by the National Science Foundation grant NSF DMS-1953712 and National Institute on Deafness and Other Communication Disorders grants R01DC013315 and R01DC015504 awarded to Sarkar and Chandrasekaran.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Birdsong, D. (2004). Second language acquisition and ultimate attainment. *Handbook of Applied Linguistics*, pages 82–105.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., and Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, **33**, 10–16.

Borooah, V. K. (2002). *Logit and probit: ordered and multinomial models*. Sage.

Brody, C. D. and Hanks, T. D. (2016). Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology*, **37**, 149–157.

Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, **57**, 153–178.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag, Berlin.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., and Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, **14**, 1462–1467.

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, **128**, 456–465.

Chandrasekaran, B., Yi, H.-G., and Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, **21**, 488–495.

Chhikara, R. (1988). *The inverse Gaussian distribution: Theory, methodology, and applications*. CRC Press.

Chib, S. and Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, **110**, 67–89.

Cox, D. R. and Miller, H. D. (1965). *The theory of stochastic processes*. CRC Press.

Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika*, **75**, 613–632.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**, 553–566.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., *et al.* (2002). *Analysis of longitudinal data*. Oxford University Press.

Ding, L. and Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. *Neuron*, **79**, 640–649.

Dufau, S., Grainger, J., and Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 1117–1128.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–102.

Feng, G., Yi, H. G., and Chandrasekaran, B. (2019). The role of the human auditory corticostriatal network in speech learning. *Cerebral Cortex*, **29**, 4077–4089.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, **26**, 1099–1121.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–534.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, **29**, 245–273.

Glimcher, P. W. and Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, **30**, 535–574.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, **58**, 121–128.

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, **431**, 859–862.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, **87**, 47–57.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, **21**, 60–99.

Kim, S., Potter, K., Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2017). A Bayesian race model for recognition memory. *Journal of the American Statistical Association*, **112**, 77–91.

Kunkel, D., Potter, K., Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2019). A Bayesian race model for response times under cyclic stimulus discriminability. *The Annals of Applied Statistics*, **13**, 271–296.

Leite, F. P. and Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, **72**, 246–273.

Li, Y., Lin, X., and Müller, P. (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics*, **66**, 70–78.

Lu, J. (1995). *Degradation processes and related reliability models*. Ph.D. thesis, McGill University, Montreal, Canada.

Maddox, W. T. and Chandrasekaran, B. (2014). Tests of a dual-system model of speech category learning. *Bilingualism: Language and Cognition*, **17**, 709–728.

McDonald, S. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall, London.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, **5**, 437–449.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Müller, P., Quintana, F. A., Rosner, G. L., and Maitland, M. L. (2013). Bayesian inference for longitudinal data with non-parametric treatment effects. *Biostatistics*, **15**, 341–352.

Navarro, D. J. and Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, **53**, 222–230.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, **31**, 705–767.

Nguyen, X. and Gelfand, A. E. (2011). The Dirichlet labeling process for clustering functional data. *Statistica Sinica*, **21**, 1249–1289.

Nguyen, X. and Gelfand, A. E. (2014). Bayesian nonparametric modeling for functional analysis of variance. *Annals of the Institute of Statistical Mathematics*, **66**, 495–526.

Paulon, G., Reetzke, R., Chandrasekaran, B., and Sarkar, A. (2019). Functional logistic mixed-effects models for learning curves from longitudinal binary data. *Journal of Speech, Language, and Hearing Research*, **62**, 543–553.

Pedersen, M. L., Frank, M. J., and Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, **24**, 1234–1251.

Peters, J. and D'Esposito, M. (2020). The drift diffusion model as the choice rule in inter-temporal and risky choice: A case study in medial orbitofrontal cortex lesion patients and controls. *PLOS Computational Biology*, **16**.

Petrone, S., Guindani, M., and Gelfand, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B*, **71**, 755–782.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.

Purcell, B. A. (2013). *Neural mechanisms of perceptual decision making*. Vanderbilt University.

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and B. Gold, E. (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, **111**, 1168–1181.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, **77**, 257–286.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59–108.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, **20**, 873–922.

Ratcliff, R. and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, **9**, 347–356.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, **20**, 260–281.

Reetzke, R., Xie, Z., Llanos, F., and Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, **28**, 1419–1427.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). *Stochastic processes*. Wiley New York.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, **2**, 33–42.

Scott, S. L. (2002). Bayesian methods for hidden Markov models recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

Singer, J. D., Willett, J. B., Willett, J. B., *et al.* (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, **27**, 161–168.

Smith, P. L. and Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, **32**, 135–168.

Song, J. H., Skoe, E., Wong, P. C., and Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, **20**, 1892–1902.

Teichert, T., Grinband, J., and Ferrera, V. (2016). The importance of decision onset. *Journal of Neurophysiology*, **115**, 643–661.

Titsias, M. K. and Yau, C. (2014). Hamming ball auxiliary sampling for factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 2960–2968.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, **36**, 702–716.

Tuerlinckx, F., Maris, E., Ratcliff, R., and De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior Research Methods, Instruments, & Computers*, **33**, 443–456.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550–592.

Van der Groen, O., Tang, M. F., Wenderoth, N., and Mattingley, J. B. (2018). Stochastic resonance enhances the rate of evidence accumulation during combined brain stimulation and perceptual decision-making. *PLOS Computational Biology*, **14**, 1–17.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM.

Vandekerckhove, J. and Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, **14**, 1011–1026.

Vandekerckhove, J., Tuerlinckx, F., and Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision-making. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1429–1434. Washington, DC.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, **106**, 3649–3658.

Whitmore, G. and Seshadri, V. (1987). A heuristic derivation of the inverse gaussian distribution. *The American Statistician*, **41**, 280–281.

Wong, P. C., Vuong, L. C., and Liu, K. (2017). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia*, **98**, 192–200.

Xie, Z., Reetzke, R., and Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of Neurophysiology*, **117**, 1409–1424.

Zanella, G. (2019). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, pages 1–14.

Supplementary Materials for

Bayesian Semiparametric Longitudinal Drift-Diffusion Mixed Models for Tone Learning in Adults

Giorgio Paulon[1] (giorgio.paulon@utexas.edu)
Fernando Llanos[2,3] (f.llanos@pitt.edu)
Bharath Chandrasekaran[3](b.chandra@pitt.edu)
Abhra Sarkar[1] (abhra.sarkar@utexas.edu)

[1]Department of Statistics and Data Sciences,
University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA

[2]Department of Linguistics,
University of Texas at Austin,
305 East 23rd Street B5100, Austin, TX 78712, USA

[3]Department of Communication Science and Disorders,
University of Pittsburgh,
4028 Forbes Tower, Pittsburgh, PA 15260, USA

Supplementary materials present brief reviews of B-splines, additional illustrations of our proposed smoothness inducing priors, brief reviews of fHMMs and associated computational machinery, details of the MCMC algorithm we designed to sample from the posterior, MCMC performance diagnostics, a review of linear ballistic accumulator models, comparisons with a simpler sub-model, results of simulation experiments, and some additional figures. Separate files additionally include the tone categorization data set described in Section 2 and analyzed in Section 5, audio recordings of the four input Mandarin tones, and `R` programs implementing the longitudinal drift-diffusion mixed model developed in this article.

S.1 B-splines

In the main article, we employed quadratic B-spline bases in the construction of functional factorial HMMs. The construction of quadratic B-spline bases is detailed below (de Boor, 1978). Consider knot-points $t_1 = t_2 = t_3 = A < t_4 < \cdots < B = t_{K+3} = t_{K+4} = t_{K+5}$, where $t_{3:(K+3)}$ are equidistant with $\delta = (t_4 - t_3)$. For $j = 3, 4, \ldots, (K+2)$, quadratic B-splines B_j are then defined as

$$B_j(X) = \begin{cases} \{(X - t_{J-1})/\delta\}^2/2 & \text{if } t_{J-1} \leq X < t_J, \\ -\{(X - t_J)/\delta\}^2 + (X - t_J)/\delta + 1/2 & \text{if } t_J \leq X < t_{j+2}, \\ \{1 - (X - t_{j+2})/\delta\}^2 & \text{if } t_{j+2} \leq X < t_{j+3}, \\ 0 & \text{otherwise.} \end{cases}$$

The components at the ends are likewise defined as

$$B_1(X) = \begin{cases} \{1 - (X - t_1)/\delta\}^2/2 & \text{if } t_3 \leq X < t_4, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_2(X) = \begin{cases} -\{(X - t_3)/\delta\}^2 + (X - t_4)/\delta + 1/2 & \text{if } t_3 \leq X < t_4, \\ \{1 - (X - t_4)/\delta\}^2/2 & \text{if } t_4 \leq X < t_5, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{K+1}(X) = \begin{cases} \{(X - t_{K+1})/\delta\}^2/2 & \text{if } t_{K+1} \leq X < t_{K+2}, \\ -\{(X - t_{K+2})/\delta\}^2 + (X - t_{K+2})/\delta + 1/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{K+2}(X) = \begin{cases} \{(X - t_{K+2})/\delta\}^2/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3 in the main paper provides a graphical illustration of these functions.

S.2 Illustration of the Smoothness Inducing Prior

Our proposed model penalizes the difference between the pairs of core coefficients in the same latent state. Figure 5 in the main paper, reproduced here as Figure S.1 for easy access, shows the effect of the smoothing prior on the core coefficients in a synthetic scenario with $x \in \{1, 2, 3\}$.

In the example in the left panel, at location $k - 1 = 5$, all of the levels for the covariate

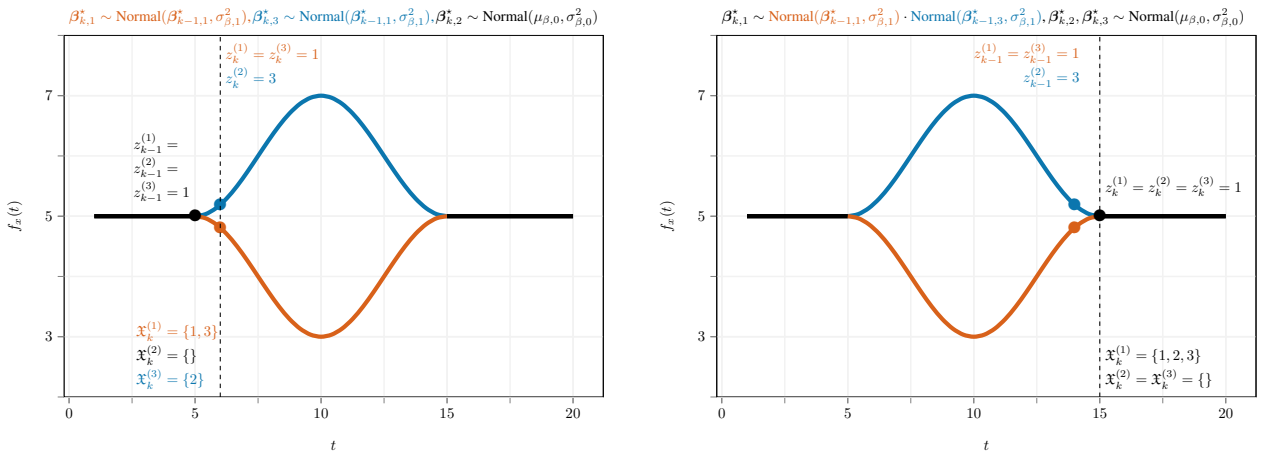

Figure S.1: An illustration of the prior on the spline core coefficients $\beta^\star_{k,z_k}$ at location k (marked by the dashed vertical lines) in the fixed effects model developed in Section 3.2.1 for a scenario with $x \in \{1, 2, 3\}$, where the curves corresponding to the three levels of x are initially equal, the curves for $x = 1, 3$ (in red) and $x = 2$ (in blue) then diverge at $t = 6$, merging back again at $t = 15$.

x are assigned to the first latent state, yielding the same curve for the three levels of x. At location $k = 6$, levels 1 and 3 are assigned to the first latent state, whereas level 2 is assigned to the third latent state. This corresponds to the case in which the curves for $x = 1, 3$ and $x = 2$ diverge. Therefore, using (9),

- $\mathfrak{X}_k^{(1)} = \{x : z_k^{(x)} = 1\} = \{1, 3\}$ and the conditional prior for the core coefficient of the first latent state is $\beta^\star_{k,1} \sim \prod_{j \in \{z_{k-1}^{(1)}, z_{k-1}^{(3)}\}} \text{Normal}(\beta^\star_{k-1,j}, \sigma^2_{\beta,1}) = \text{Normal}(\beta^\star_{k-1,1}, \sigma^2_{\beta,1})$,
- $\mathfrak{X}_k^{(2)} = \{x : z_k^{(x)} = 2\} = \emptyset$ and the conditional prior for the core coefficient of the second latent state is $\beta^\star_{k,2} \sim \text{Normal}(\mu_{\beta,0}, \sigma^2_{\beta,0})$,
- $\mathfrak{X}_k^{(3)} = \{x : z_k^{(x)} = 3\} = \{2\}$ and the conditional prior for the core coefficient of the third latent state is $\beta^\star_{k,3} \sim \prod_{j \in \{z_{k-1}^{(2)}\}} \text{Normal}(\beta^\star_{k-1,j}, \sigma^2_{\beta,1}) = \text{Normal}(\beta^\star_{k-1,1}, \sigma^2_{\beta,1})$.

In the example in the right panel, at location $k - 1 = 14$, levels 1 and 3 are assigned to the first latent state, whereas level 2 is assigned to the third latent state. At location $k = 15$, all of the levels for the covariate x are assigned to the first latent state. This corresponds to the case in which the curves for $x = 1, 3$ and $x = 2$ merge back. Therefore,

- $\mathfrak{X}_k^{(1)} = \{x : z_k^{(x)} = 1\} = \{1, 2, 3\}$ and the conditional prior for the core coefficient of the first latent state is $\beta^\star_{k,1} \sim \prod_{j \in \{z_{k-1}^{(1)}, z_{k-1}^{(2)}, z_{k-1}^{(3)}\}} \text{Normal}(\beta^\star_{k-1,j}, \sigma^2_{\beta,1}) = \prod_{j \in \{1,3\}} \text{Normal}(\beta^\star_{k-1,j}, \sigma^2_{\beta,1})$,
- $\mathfrak{X}_k^{(2)} = \{x : z_k^{(x)} = 2\} = \emptyset$ and the conditional prior for the core coefficient of the second latent state is $\beta^\star_{k,2} \sim \text{Normal}(\mu_{\beta,0}, \sigma^2_{\beta,0})$,
- $\mathfrak{X}_k^{(3)} = \{x : z_k^{(x)} = 3\} = \emptyset$ and the conditional prior for the core coefficient of the third latent state is $\beta^\star_{k,3} \sim \text{Normal}(\mu_{\beta,0}, \sigma^2_{\beta,0})$.

S.3 Factorial HMM (fHMM)

The basic HMM (Frühwirth-Schnatter, 2006; McDonald and Zucchini, 1997, etc.) consists of two processes: an *observed* process $\{\mathbf{y}_t\}$ recorded sequentially over a set of discrete time points $t = 1, 2, \ldots, T$ and an associated *hidden* process $\{z_t\}$ which evolves according to a first order Markov chain with discrete state space. Specifically, an HMM makes the following set of conditional independence assumptions to model the hidden and the observed processes

$$p(z_t \mid \mathbf{z}_{1:(t-1)}) = p(z_t \mid z_{t-1}),$$
$$p(\mathbf{y}_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{z}_{1:t}) = p(y_t \mid z_t).$$

The distributions $p(z_t \mid z_{t-1})$ and $p(y_t \mid z_t)$ are often referred to as the *transition distribution* and the *emission distribution*, respectively.

In factorial HMMs (Ghahramani and Jordan, 1997), the latent states are represented by a collection of variables $\{\mathbf{z}_t\} = \{(z_t^{(1)}, \ldots, z_t^{(L)})\}$ where each component $\{z_t^{(\ell)}\}$ now evolves according to a first order Markov chain with discrete state spaces, and the *observed* process $\{y_t\}$ is observed sequentially as before over a set of discrete time points $t = 1, 2, \ldots, T$. An fHMM thus makes the following set of conditional independence assumptions to model the hidden and the observed processes

$$p(\mathbf{z}_t \mid \mathbf{z}_{1:(t-1)}) = \textstyle\prod_{\ell=1}^{L} p(z_t^{(\ell)} \mid z_{t-1}^{(\ell)}),$$
$$p(y_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{z}_{1:t}) = p(\mathbf{y}_t \mid \mathbf{z}_t) = p(\mathbf{y}_t \mid z_t^{(1)}, \ldots, z_t^{(L)}).$$

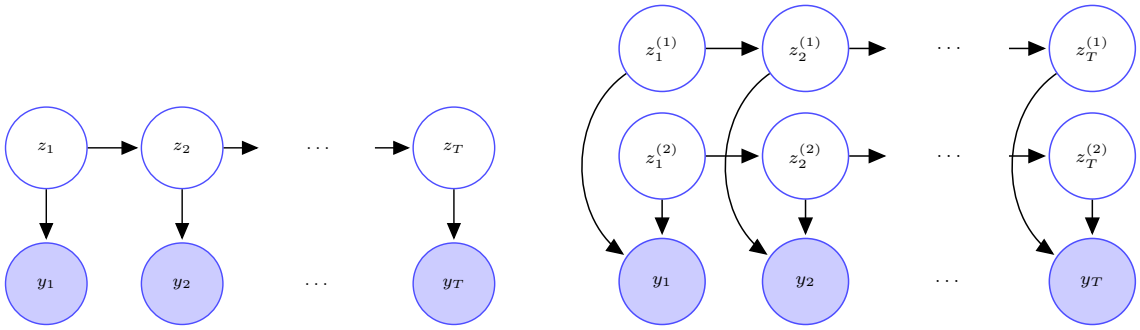

Figure S.2: Left panel: Graph of an HMM. Right panel: Graph of an fHMM with two layers.

In our work, we adapted the basic fHMM to characterize local influences of the categorical predictor in longitudinal functional models. In the drift-diffusion model of Section 3, for each input tone $s \in \{1, \ldots, d_1\}$, we introduced an fHMM $\{\mathbf{z}_k^{(s)} = (z_k^{(1,s)}, \ldots, z_k^{(d_0,s)})\}$ with d_0 layers, one for each level of the response d. Conditional on $z_k^{(d,s)} = z_k$, we then associated the coefficients $\beta_{k,d,s}$ of a predictor dependent B-spline mixture model with atoms $\beta^{\star}_{k,z_k}$. Specifically, we let

$$p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}) = \textstyle\prod_{d=1}^{d_0} p(z_k^{(d,s)} \mid z_{k-1}^{(d,s)}),$$
$$\{\beta_{k,d,s} \mid z_k^{(d,s)} = z_k\} = \beta^{\star}_{k,z_k}.$$

S.4 Locally Informed Hamming Ball Sampler

Forward-backward (or backward-forward) algorithms for HMMs rely on passing messages forward (or backward) and then sampling backward (or forward) (Rabiner, 1989; Scott, 2002). While adapting such algorithms to fHMMs, the requirement to sum over all possible configurations in computing the messages becomes a challenge. Hamming ball samplers for fHMMs (Titsias and Yau, 2014) avoid this computationally expensive step by introducing and conditioning on an auxiliary variable that restricts the sampling to only a slice (Neal, 2003) of the entire high-dimensional space. In doing so, the sampler also allows localized joint updating of all constituent chains, making it less prone to get trapped in local modes.

Let $h(\mathbf{z}_t, \mathbf{v}_t) = \sum_{\ell=1}^{L} 1\{z_t^{(\ell)} \neq v_t^{(\ell)}\}$ denote the Hamming distance between the vectors $\mathbf{z}_t = (z_t^{(1)}, \ldots, z_t^{(L)})^{\mathrm{T}}$ and $\mathbf{v}_t = (v_t^{(1)}, \ldots, v_t^{(L)})^{\mathrm{T}}$ and $\mathcal{H}_m(\mathbf{z}_t) = \{\mathbf{v}_t : h(\mathbf{z}_t, \mathbf{v}_t) \leq m\}$ denote a Hamming ball of radius m around $\mathbf{z}_t$.

Consider an fHHM, as shown in Figure S.2 but with L component chains each with state space $\{1, \ldots, d\}$. Introducing an auxiliary variable $\mathbf{v}$ following a conditional probability distribution $p(\mathbf{v} \mid \mathbf{z}) = \prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)$, the augmented joint model becomes $p(\mathbf{y}, \mathbf{z}, \mathbf{v}) = p(\mathbf{v} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z}) = \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}p(\mathbf{z}_1)\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})$. Sampling $\mathbf{v}$ from the posterior can then be done by sampling independently from the full conditionals $p(\mathbf{v}_t \mid \mathbf{z}_t)$. Sampling $\mathbf{z}$ from the posterior can still be carried out using forward-backward (or backward-forward) message passing algorithms but with the augmented full conditional $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v}) \propto \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}\big\{\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\big\}p(\mathbf{z}_1)$. The set of possible configurations needed to compute the messages at time t is now restricted to the support of $p(\mathbf{v}_t \mid \mathbf{z}_t)$. If this can be made much smaller compared to the original size of the state space, computational burden can be greatly reduced.

The Hamming ball algorithm does this by setting $p(\mathbf{v}_t \mid \mathbf{z}_t) \propto 1\{\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)\}$, that is, by sampling the $\mathbf{v}_t$'s uniformly from $\mathcal{H}_m(\mathbf{z}_t)$. By symmetry, since $\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)$ if and only if $\mathbf{z}_t \in \mathcal{H}_m(\mathbf{v}_t)$, the support of each $\mathbf{z}_t$ in the full conditional $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v})$ is then restricted only to $\mathcal{H}_m(\mathbf{v}_t)$.

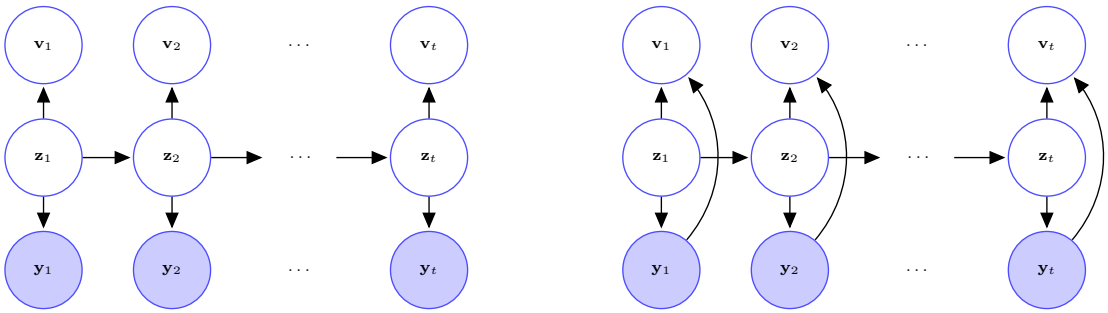

Figure S.3: Graph of a Hamming ball sampler (left panel) and a locally informed Hamming ball sampler (right panel) for fHMM.

The Hamming ball sampler is still limited in its ability to efficiently explore the neighborhood of $\mathbf{z}_t$ as it blindly proposes new values along arbitrarily chosen directions within the ball. More informed moves can be proposed utilizing the information contained in the likeli-

hood function (Zanella, 2019). For instance, $p(\mathbf{v}_t \mid \mathbf{z}_t, \mathbf{y}_t) \propto g\{p(\mathbf{y}_t \mid \mathbf{v}_t)\}1\{\mathbf{v}_t \in \mathcal{H}_m(\mathbf{z}_t)\}$, for proper choices of $g(\cdot)$, favors moves along directions that increase the conditional likelihood $p(\mathbf{y}_t \mid \mathbf{v}_t)$ (Figure S.3). The augmented joint model now becomes $p(\mathbf{y}, \mathbf{z}, \mathbf{v}) = p(\mathbf{v} \mid \mathbf{y}, \mathbf{z})p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z}) = \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{y}_t, \mathbf{z}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}\big\{\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\big\}p(\mathbf{z}_1)$. Sampling $\mathbf{z}$ from the posterior can be carried out using message passing algorithms as before with each $\mathbf{z}_t$ restricted to $\mathcal{H}_m(\mathbf{v}_t)$ but with the updated full conditionals $p(\mathbf{z} \mid \mathbf{y}, \mathbf{v}) \propto \big\{\prod_{t=1}^{T} p(\mathbf{v}_t \mid \mathbf{z}_t, \mathbf{y}_t)p(\mathbf{y}_t \mid \mathbf{z}_t)\big\}\big\{\prod_{t=2}^{T} p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\big\}p(\mathbf{z}_1)$.

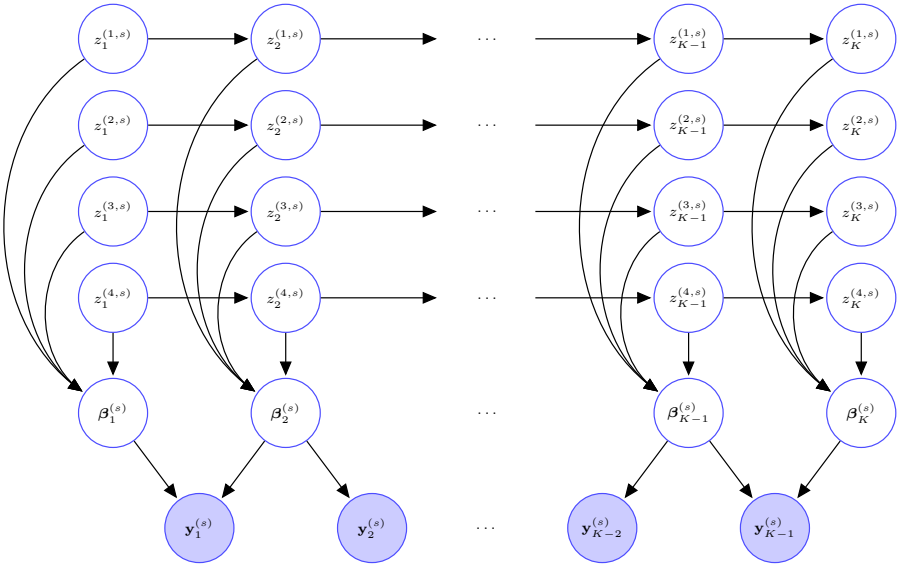

Figure S.4: Graph of the proposed longitudinal drift-diffusion mixed model for tone learning with $\boldsymbol{\beta}_k^{(1,s)}, \ldots, \boldsymbol{\beta}_k^{(4,s)}$ collected in single nodes $\boldsymbol{\beta}_k^{(s)}$ for each k.

S.5 Posterior Inference

S.5.1 Prior Hyper-parameters and MCMC Initializations

The fixed effects parameters of the drift-diffusion mixed effects model (6) are initialized with an empirical Bayes type approach. As discussed in Section 3, the boundary and the drift parameters are related to the first two moments of the response times. Thus, we can use the empirical distribution of the response times to choose the initial guess for both drift and boundary parameters for each combination of input stimulus and response. The random effects are instead initialized at zero. The clustering configuration is initialized with all the success curves in different clusters, and all the failure curves in the same cluster.

Other crucial hyper-parameters are the mean and the standard deviation for the prior term of the unassigned components of $\boldsymbol{\beta}_{\mu}^{(x)}$ and $\boldsymbol{\beta}_{b}^{(x)}$, that is, the second term in the prior (9) in the main paper. We use the empirical distributions of the response times at every time point to set $\mu_{\beta,0}, \sigma^2_{\beta,0}$.

The hyper-parameters in the Gamma(a_α, b_α) prior for the concentration parameters $\alpha^{(C)}$ and $\alpha^{(I)}$ of the Dirichlet distributions characterizing the latent variable dynamics are set at $a_\alpha = b_\alpha = 1$, as recommended in Escobar and West (1995).

The half-Cauchy priors $C^+(0,1)$ on the smoothness parameters are non-informative for the smoothness of the corresponding longitudinal curves. The $C^+(0,1)$ distribution attains its mode at zero and hence is capable of capturing strong smoothness but also has heavy tails and is thus also capable of capturing wiggly functions. The left panel of Figure S.5 shows some draws from $\mu_x(t) \mid \sigma^2_{\beta_\mu,1}$ with independent draws of the corresponding smoothness controlling parameter $\sigma^2_{\beta_\mu,1}$ from a $C^+(0,1)$ prior. A wide variety of curves are clearly sampled - some very smooth, some very wiggly, and many in between. Also, as the right panel of Figure S.5 illustrates, the posterior distributions of the smoothness parameters in our model all concentrate well within a region of flat $C^+(0,1)$ prior probability density. This is additional evidence that our prior is not producing any consistent bias in the posterior estimates.

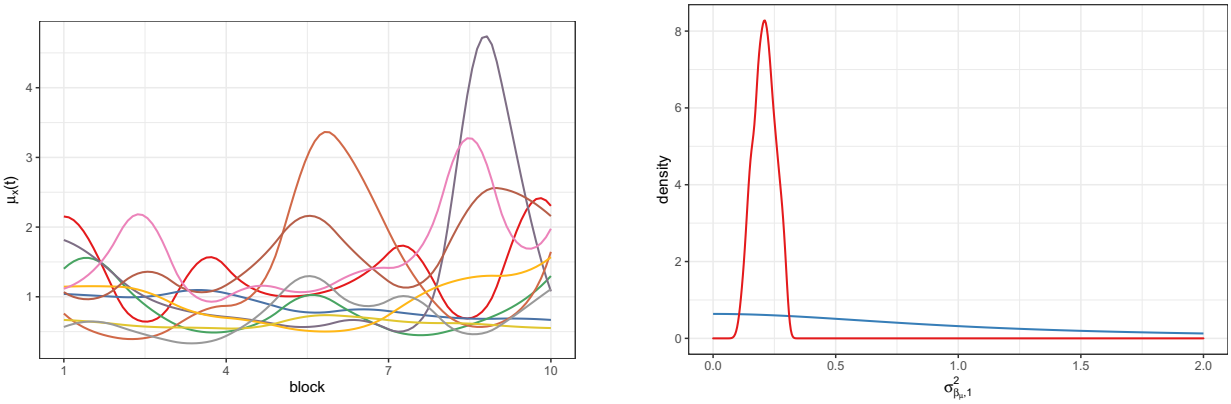

Figure S.5: Left: 10 conditionally independent draws from $\mu_x(t) \mid \sigma^2_{\beta_\mu,1}$ with independent draws of $\sigma^2_{\beta_\mu,1}$ from a $C^+(0,1)$ prior. Right: The $C^+(0,1)$ prior distribution (in blue) and the corresponding posterior distribution (in red) for the smoothness parameter $\sigma^2_{\beta_\mu,1}$.

S.5.2 Posterior Computation

Posterior inference for the longitudinal drift-diffusion mixed model, described in Section 3 in the main paper, is based on samples drawn from the posterior using a message passing MCMC algorithm.

In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned, including the data points. Also, p_0 will sometimes be used as a generic for a prior distribution without explicitly mentioning its hyper-parameters. The sampler for the drift diffusion model of Section 3 comprises the following steps.

1. Update the offset parameters $\delta_s^{(i)}, s = 1, \ldots, d_0$. The full conditionals $p(\delta_s^{(i)} \mid \boldsymbol{\zeta}) \propto p_0(\delta_s^{(i)}) L(\mathbf{y} \mid \mathbf{s}, \boldsymbol{\theta})$ do not have closed forms. Metropolis-Hastings (MH) steps with log-normal proposals centered on the previous sampled values are used to update these parameters.

2. Jointly update the drift and boundary spline coefficients $(\beta^\star_{\mu,k,z_k}, \beta^\star_{b,k,z_k}), k = 1, \ldots, K$.

(a) If the parameters are assigned to one of the clusters, the full conditionals do not have closed forms. MH steps are therefore used with the smoothness inducing priors (9) on $(\beta^{\star}_{\mu,k,z_k}, \beta^{\star}_{b,k,z_k})$ as the proposal distributions.

(b) If the parameters are not assigned to any of the clusters, the full conditional distribution is the second term of the prior in (9).

3. Update the latent cluster assignments $\mathbf{z}_k^{(s)} = (z_k^{(1,s)}, \ldots, z_k^{(4,s)})^{\mathrm{T}}$:

(a) Sample the auxiliary variables $\mathbf{v}_k^{(s)} = (v_k^{(1,s)}, \ldots, v_k^{(4,s)})^{\mathrm{T}}$ as

$$p(\mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \mathbf{y}_k^{(s)}, \boldsymbol{\zeta}) \propto g\{p(\mathbf{y}_k^{(s)} \mid \mathbf{v}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta})\} 1\{\mathbf{v}_k^{(s)} \in \mathcal{H}_m(\mathbf{z}_k^{(s)})\}, \qquad k = 1, \ldots, K-1,$$
$$p(\mathbf{v}_K^{(s)} \mid \mathbf{z}_K^{(s)}, \boldsymbol{\zeta}) \propto 1\{\mathbf{v}_K^{(s)} \in \mathcal{H}_m(\mathbf{z}_K^{(s)})\}.$$

(b) Back-propagate the messages $m_k(\mathbf{z}_k^{(s)}) = p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta})$ using the recursion

$$\begin{aligned}
&m_k(\mathbf{z}_k^{(s)}) = p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_k^{(s)}, \mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{(k+1):(K-1)}^{(s)}, \mathbf{v}_{(k+1):K}^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{y}_k^{(s)}, \mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{(k+1):(K-1)}^{(s)}, \mathbf{v}_{(k+1):K}^{(s)} \mid \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) \\
&= \sum_{\mathbf{z}_{k+1}^{(s)}} p(\mathbf{v}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \mathbf{y}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) m_{k+1}(\mathbf{z}_{k+1}^{(s)}), \\
&\propto \sum_{\mathbf{z}_{k+1}^{(s)} \in \mathcal{H}_m(\mathbf{v}_{k+1}^{(s)})} g\{p(\mathbf{y}_k^{(s)} \mid \mathbf{v}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta})\} 1\{\mathbf{v}_k^{(s)} \in \mathcal{H}_m(\mathbf{z}_k^{(s)})\} p(\mathbf{y}_k^{(s)} \mid \mathbf{z}_k^{(s)}, \mathbf{z}_{k+1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_{k+1}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) m_{k+1}(\mathbf{z}_{k+1}^{(s)}),
\end{aligned}$$

starting with the final condition $m_K(\mathbf{z}_K^{(s)}) = 1\{\mathbf{z}_K^{(s)} \in \mathcal{H}_m(\mathbf{v}_K^{(s)})\}$.

(c) Sample the latent cluster assignments forward one step at a time from

$$p(\mathbf{z}_{1:K}^{(s)} \mid \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) = p(\mathbf{z}_K^{(s)} \mid \mathbf{z}_{1:(K-1)}^{(s)}, \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) \ldots p(\mathbf{z}_1^{(s)} \mid \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}),$$

where

$$\begin{aligned}
&p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \mathbf{y}_{1:(K-1)}^{(s)}, \mathbf{v}_{1:K}^{(s)}, \boldsymbol{\zeta}) \propto p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_{1:k}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) \\
&= p(\mathbf{y}_{1:(k-2)}^{(s)}, \mathbf{v}_{1:(k-2)}^{(s)} \mid \mathbf{z}_{1:k}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k-1}^{(s)}, \mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k:(K-1)}^{(s)}, \mathbf{v}_{k:K}^{(s)} \mid \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) \\
&\propto p(\mathbf{y}_{k-1}^{(s)}, \mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{1:(k-1)}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)})
\end{aligned}$$

$$= p(\mathbf{v}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \mathbf{y}_{k-1}^{(s)}, \boldsymbol{\zeta}) p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)})$$
$$\propto g\{p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{v}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta})\} p(\mathbf{y}_{k-1}^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \mathbf{z}_k^{(s)}, \boldsymbol{\zeta}) p(\mathbf{z}_k^{(s)} \mid \mathbf{z}_{k-1}^{(s)}, \boldsymbol{\zeta}) m_k(\mathbf{z}_k^{(s)}).$$

4. Update the cluster specific fixed effects spline coefficients:

$$(\beta_{\mu,k}^{(x)} \mid z_k^{(x)} = z_k, \boldsymbol{\zeta}) \sim 1\{\beta_{\mu,k}^{(x)} = \beta_{\mu,k,z_k}^{\star}\}, \quad k = 1, \ldots, K.$$
$$(\beta_{b,k}^{(x)} \mid z_k^{(x)} = z_k, \boldsymbol{\zeta}) \sim 1\{\beta_{b,k}^{(x)} = \beta_{b,k,z_k}^{\star}\}, \quad k = 1, \ldots, K.$$

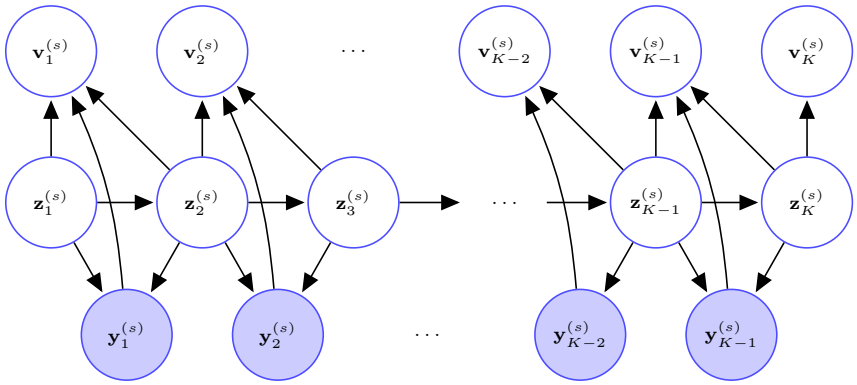

Figure S.6: Locally informed Hamming ball sampling of the latent states in our tone-learning longitudinal drift-diffusion mixed model. See also Figure 6 in the main paper.

5. Update the transition probability matrices:

$$\left(\boldsymbol{\pi}_z^{(C)} \mid \boldsymbol{\zeta}\right) \sim \text{Dir}(\alpha^{(C)}/z_{\max} + n_{z,1}^{(C)}, \ldots, \alpha^{(C)}/z_{\max} + n_{z,z_{\max}}^{(C)})$$
$$\left(\boldsymbol{\pi}_z^{(I)} \mid \boldsymbol{\zeta}\right) \sim \text{Dir}(\alpha^{(I)}/z_{\max} + n_{z,1}^{(I)}, \ldots, \alpha^{(I)}/z_{\max} + n_{z,z_{\max}}^{(I)}),$$

where $n_{z,z'}^{(C)} = \sum_k 1\{z_k^{(x)} = z, z_{k+1}^{(x)} = z'\}$ is the number of transitions from z to z' for the HMMs associated with the correct identification of the tones, that is, with x s.t. $d = s$. A similar definition holds for $n_{z,z'}^{(I)}$.

6. Update the cluster specific smoothness parameter

$$p(\sigma_{\beta_\mu,1}^2 \mid \boldsymbol{\zeta}) \propto \left(\sigma_{\beta_\mu,1}\right)^{-Kx_{\max}} \exp\left(-\frac{1}{2\sigma_{\beta_\mu,1}^2} \sum_x \boldsymbol{\beta}_\mu^{(x)\mathrm{T}} \mathbf{P}_u \boldsymbol{\beta}_\mu^{(x)}\right) p_0(\sigma_{\mu,u,a}^2).$$

MH steps with log-normal proposals centered on the previous sampled values are used to update these parameters.

7. Update the random effects spline coefficients $\beta_{\mu,k,u}^{(i)}$ and $\beta_{b,k,u}^{(i)}$: The full conditional does not have a closed form. An MH step with a normal proposal centered on the previous value was used.

8. Update the random effects variance parameters $\sigma^2_{\mu,u,a}$, $\sigma^2_{\mu,u,s}$, $\sigma^2_{b,u,a}$ and $\sigma^2_{b,u,s}$:
The full conditional for $\sigma^2_{\mu,u,a}$ is given by

$$p(\sigma^2_{\mu,u,a} \mid \boldsymbol{\zeta}) \propto \det(\sigma^{-2}_{\mu,u,s}\mathbf{P}_u + \sigma^{-2}_{\mu,u,a}\mathbf{I}_K)^{n/2} \exp\left(-\frac{1}{2\sigma^2_{\mu,u,a}} \sum_{i=1}^{n} \boldsymbol{\beta}^{(i)\mathrm{T}}_{\mu,u}\mathbf{P}_u\boldsymbol{\beta}^{(i)}_{\mu,u}\right) p_0(\sigma^2_{\mu,u,a}).$$

Analogous expressions can be found for the full conditionals of $\sigma^2_{\mu,u,s}$, $\sigma^2_{b,u,a}$ and $\sigma^2_{b,u,s}$. MH steps with log-normal proposals centered on the previous sampled values are used to update these parameters.

The main challenge here arises from the nonconjugacy of the inverse Gaussian distribution based likelihood function, requiring MH steps for updating $\delta^{(i)}_s, \beta^\star_{b,k,z_k}, \beta^\star_{\mu,k,z_k}$. We employed the adaptive MH algorithm (Roberts and Rosenthal, 2009) for updating $\delta^{(i)}_s$ and the variance parameters, avoiding the difficult task of choosing the parameters of their proposal distributions while also improving mixing. Specifically, for every batch of 50 iterations, we inflate or deflate the standard deviation of the proposal distribution such that the optimal acceptance rate of 44% is achieved (Roberts *et al.*, 2001). The adaptive MH could not be employed for the cluster specific parameters $(\beta^\star_{b,k,z_k}, \beta^\star_{\mu,k,z_k})$ due to label switching, so we used tempered MH steps instead. For the proposal distributions for $(\beta^\star_{b,k,z_k}, \beta^\star_{\mu,k,z_k})$, we used the smoothness inducing conditional prior distributions $p_0(\beta^\star_{\mu,k,z_k} \mid \boldsymbol{\beta}^\star_{\mu,k-1}) \times p_0(\beta^\star_{b,k,z_k} \mid \boldsymbol{\beta}^\star_{b,k-1})$. Since the conditioning variables $\boldsymbol{\beta}^\star_{\mu,k-1}$ and $\boldsymbol{\beta}^\star_{b,k-1}$ are also updated at every iteration, the values sampled from the smoothness inducing priors are frequently accepted.

Based on M thinned samples $\{\boldsymbol{\theta}^{(m)}\}_{m=1}^M$ drawn from the posterior after the burn-in, the individual level drift parameters in the drift-diffusion mixed model are estimated as

$$\mu^{(i)}_x(t) = \exp\{f_{\mu,x}(t) + u^{(i)}_\mu(t)\} = \tfrac{1}{M}\textstyle\sum_{m=1}^M \exp\{\widehat{f}^{(m)}_{\mu,x}(t) + \widehat{u}^{(i,m)}_\mu(t)\},$$

where $\widehat{f}^{(m)}_{\mu,x}(t) = \sum_{k=1}^K \beta^{\star(m)}_{\mu,k,z^{(x,m)}_k} B_k(t)$, $\widehat{u}^{(i,m)}_\mu(t) = \sum_{k=1}^K \beta^{(i,m)}_{k,u,\mu} B_k(t)$ etc. The population level drift parameters are likewise estimated as

$$\begin{aligned}\mu_x(t) = \textstyle\int \exp\{f_{\mu,x}(t) + u^{(i)}_\mu(t)\} f\{u^{(i)}_\mu(t)\} du^{(i)}_\mu(t) &= \exp\left[f_{\mu,x}(t) + \frac{\mathrm{var}\{u^{(i)}_\mu(t)\}}{2}\right] \\ &= \tfrac{1}{M}\textstyle\sum_{m=1}^M \exp\left\{\widehat{f}^{(m)}_{\mu,x}(t) + \frac{\mathrm{var}\{\widehat{u}^{(i,m)}_\mu(t)\}}{2}\right\},\end{aligned}$$

S.5.3 Software, Runtime, etc.

The results reported in this article are all based on $5,000$ MCMC iterations with the initial $2,000$ iterations discarded as burn-in. The remaining samples were further thinned by an interval of 5. We programmed in `R` and `C++`. The codes are available as part of the supplementary materials. The MCMC algorithm takes 10 hours on a Dell machine with 16 Gb RAM. A 'readme' file, providing additional details for a practitioner, is also included in the supplementary materials.

S.6 MCMC Diagnostics

This section presents some convergence diagnostics for the MCMC sampler described in the main manuscript. The results presented here are for the tone learning data set. Diagnostics for the simulation experiments were similar and hence omitted.

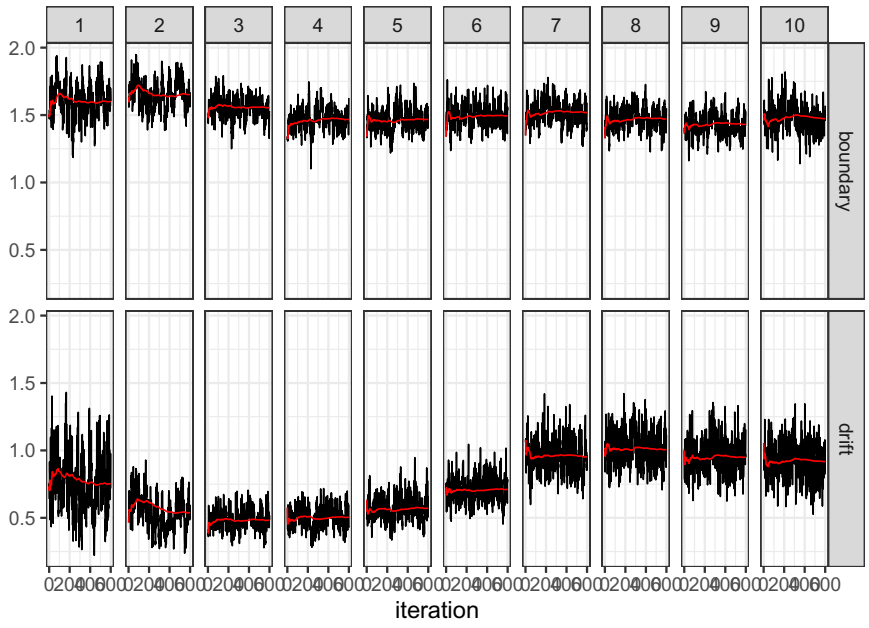

Figure S.7: Trace plots of the individual drift rates $\mu_{1,1}^{(i)}(t)$ and boundary parameters $b_{1,1}^{(i)}(t)$ corresponding to the success categorization of tone T1 evaluated at each of the training blocks. The two rows correspond to the two different classes of parameters, and the ten columns to the training blocks. In each panel, the solid red line shows the running mean. Results for other drift and boundary parameters were very similar.

Figure S.7 shows the trace plots of some individual level parameters at different training blocks. Figure S.8 shows the trace plots of some individual level offset parameters. These results are based on the MCMC thinned samples. As these figures show, the running means are very stable and there seems to be no convergence issues. Additionally, the Geweke test (Geweke, 1991) for stationarity of the chains, which formally compares the means of the first and last part of a Markov chain, was also performed. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. The results of the test, reported in Table S.1 and Table S.2, indicate that convergence was satisfactory for the parameters considered. Only one parameter, $\mu_{1,1}^{(i)}(2)$ in the second row of Table S.1, had a significant p-value. Some chance rejections are expected in multiple hypothesis testing scenarios. A visual inspection of the corresponding trace plot, however, does not indicate any serious issue.

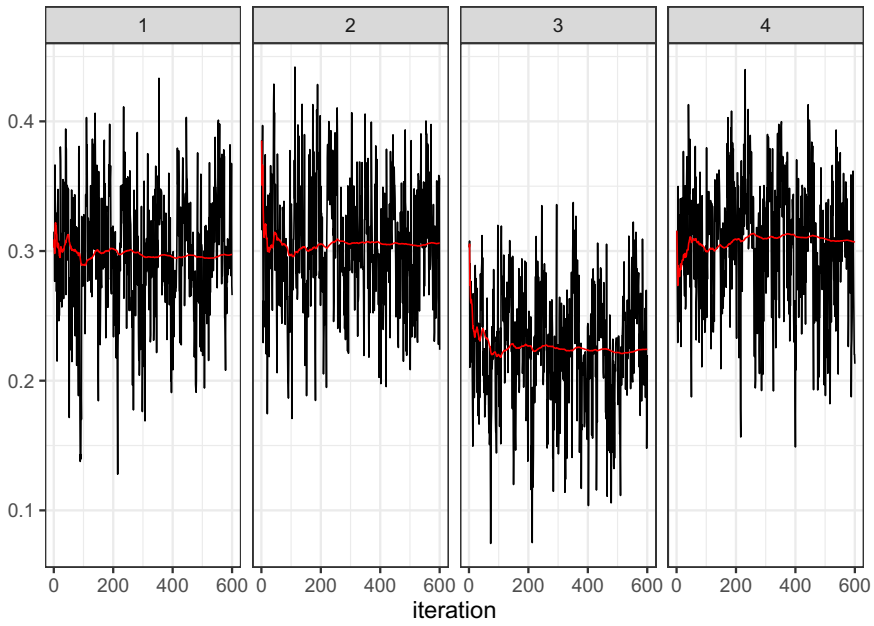

Figure S.8: Trace plots of the individual level offset parameters $\delta_s^{(i)}$ for the four possible input tones. The four columns correspond to the input stimuli s. In each panel, the solid red line shows the running mean. Results for other offset parameters were very similar.

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$
boundary	1.161	0.973	1.162	-1.287	-1.080	-0.554	0.164	-0.285	0.481	0.894
	(0.25)	(0.33)	(0.25)	(0.20)	(0.28)	(0.58)	(0.87)	(0.78)	(0.63)	(0.37)
drift	1.884	3.467	-0.102	-0.863	-1.171	-0.845	0.445	0.821	0.362	0.607
	(0.06)	(0.00)	(0.92)	(0.39)	(0.24)	(0.40)	(0.66)	(0.41)	(0.72)	(0.54)

Table S.1: Geweke statistics and associated p-values assessing convergence of the individual drift rates $\mu_{1,1}^{(i)}(t)$ and boundary parameters $b_{1,1}^{(i)}(t)$ corresponding to the success categorization of tone T1 evaluated at each of the training blocks. Results for other drift and boundary parameters were very similar.

$s=1$	$s=2$	$s=3$	$s=4$
-0.395	-0.848	-0.019	-0.217
(0.69)	(0.40)	(0.98)	(0.83)

Table S.2: Geweke statistics and associated p-values assessing convergence of the of the individual level offset parameters $\delta_s^{(i)}$ for the four possible input tones. Results for other offset parameters were very similar.

S.7 Linear Ballistic Accumulator Model

We present here a review of the LBA model (Brown and Heathcote, 2008) for easy reference with some repetition from the main paper to make this section relatively self-contained.

The LBA model is a popular framework for studying neural mechanisms underlying choice between multiple alternatives. Similar to our model, it uses independent evidence accumulators starting at δ_s that continue until a response boundary b_s is reached. The accumulator that first reaches the boundary corresponds to the decision outcome, and the time at which the boundary is reached is the response time. The evidence, however, accumulates linearly at the rate $\mu_{d,s}$, reaching the boundary b_s precisely at time $\tau_d = b_s/\mu_{d,s}$. To explain trial-by-trial variability, the LBA model assumes that the slopes μ for different trials are random draws from a Normal$(m_{d,s}, v_{d,s})$ distribution. The cumulative distribution function for the boundary crossing time τ_d for the d^{th} category is thus given by

$$F_{LBA}(\tau_d \mid \boldsymbol{\theta}_{d,s}) = 1 - \Phi\left(b_s/\tau_d \mid m_{d,s}, v_{d,s}\right),$$

where $\boldsymbol{\theta}_{d,s} = (m_{d,s}, v_{d,s}, b_s)^{\mathrm{T}}$. The likelihood of the LBA model at the t^{th} time point is thus

$$L_t(\mathbf{y}_t \mid \mathbf{s}, \boldsymbol{\theta}) = \prod_{d=1}^{d_0} \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left[f_{LBA}(\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d,s,t}) \prod_{d' \neq d} \{1 - F_{LBA}(\tau_{i,\ell,t} \mid \boldsymbol{\theta}_{d',s,t})\} \right]^{1\{d_{i,\ell,t}=d\}},$$

where $\boldsymbol{\theta}_{d,s,t} = (m_{d,s,t}, v_{d,s,t}, b_{s,t})^{\mathrm{T}}$, and $f_{LBA}(\tau) = \frac{dF_{LBA}(\tau)}{d\tau}$ is the pdf of τ.

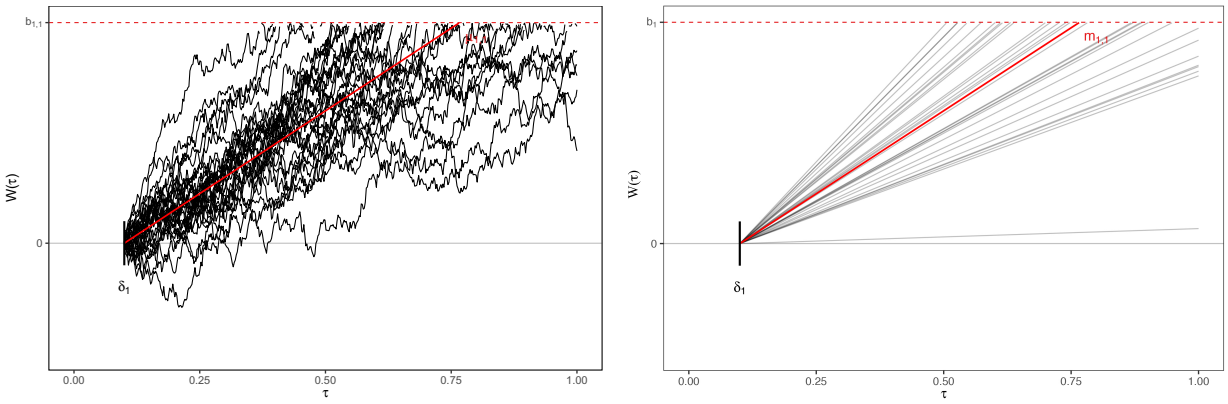

Figure S.9: Representation of the underlying evidence accumulation processes for our drift-diffusion model (left) and the LBA model (right) for 30 independent trials with fixed stimulus and decision categories $d = s = 1$. The red line represents the drift parameter $\mu_{1,1}$ for the drift-diffusion model (left) and the mean of the drift parameters $m_{1,1}$ for the LBA (right). In drift-diffusion models, trial-by-trial variability is explained by stochastically different diffusion paths for different trials. In the LBA model, trial-by-trial variability is explained by stochastically varying slopes drawn from a Normal distribution.

The existing literature on LBA models has many serious limitations. The normality assumption on the slopes μ in the LBA model does not satisfy a non-negativity constraint. A common boundary b_s for all decision categories d is also inflexible. Importantly, there is no principled method to incorporate systematic stimulus and decision category specific fixed or individual specific random effects into the LBA model. Existing literature is also limited to static settings, there is no mechanism to estimate smoothly varying longitudinal trajectories as the participants get trained and experienced in their decision tasks. In our implementation, we thus fitted these models separately for each time stamp. Finally, the likelihood function of the LBA model described above is non-convex in the parameters. Parameter estimation based on optimization of the likelihood function is thus fraught with convergence issues. We used the `rtdists` package (Singmann *et al.*, 2019) in `R`, using several random initializations and tracking the objective function to ensure convergence.

S.8 Comparison with a Simpler Sub-Model

In this section, we summarize the results produced by a simpler alternative model, specifically, a reduced static version of our proposed longitudinal drift-diffusion mixed model fitted separately to data from each block as in the case of the LBA model. Using notation similar to those in our proposed longitudinal mixed model, we now let $\mu_{x,t}^{(i)} = \exp\{f_{\mu,x,t} + u_{\mu,x,t}^{(i)}\}$ be the drift rates and $b_{x,t}^{(i)} = \exp\{f_{b,x,t} + u_{b,x,t}^{(i)}\}$ be the boundary parameters. The time index t now appears in subscript, as opposed to as an argument within parenthesis in our original longitudinal functional model. Other relevant parts of the model, including the priors, remain unchanged.

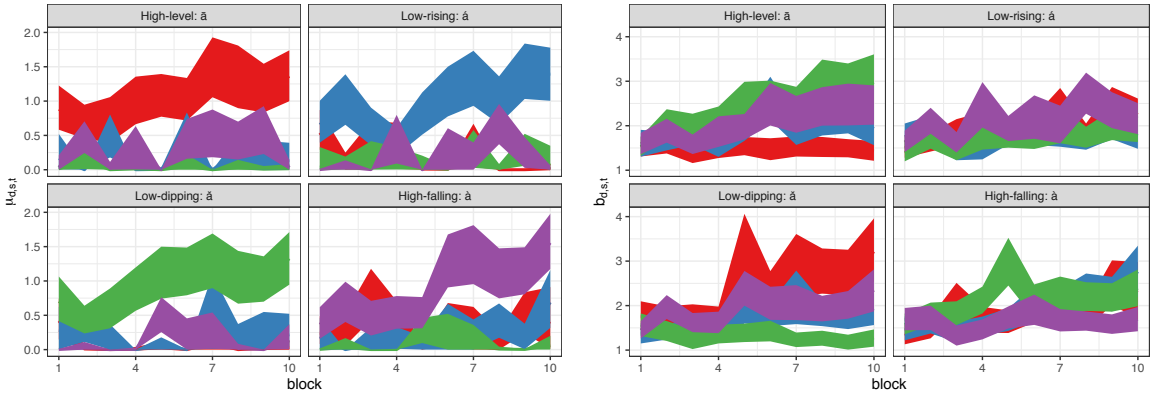

Figure S.10: Results for tone learning data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s,t}$ (left panel) and boundaries $b_{d,s,t}$ (right panel) for the inverse Gaussian drift-diffusion mixed model applied independently for each block. The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

Figure S.10 shows the posterior means and associated 90% credible intervals for the population level boundaries $b_{d,s,t}$ and drift rates $\mu_{d,s,t}$ estimated by fitting the above described static drift-diffusion model fitted separately to data from each block. These results are generally consistent with the ones illustrated in Figure 8 in the main paper. However, this reduced model yields less interpretable results for at least three reasons. First, the absence of functional dependence makes it harder to pinpoint a general trend because the estimates are not smooth but very wiggly across the training blocks. Second, the fixed effects parameters are not allowed to cluster across input-response combinations, which results in many redundant configurations. Third, the parameter estimates under our proposed model seem to have smaller uncertainty due borrowing of information across adjacent blocks as well as across input-output tone combinations via local clustering.

S.9 Simulation Studies

In this section, we discuss the results of some synthetic numerical experiments. We are not aware of any other method from the existing literature that can be readily applied or at least be easily adapted to our data settings and inferential challenges. We thus restrict our focus mostly on evaluating the performances of the proposed longitudinal inverse Gaussian drift-diffusion mixed model. We do present a comparison with the LBA model though, applying it separately for each block as in Section 5 in the main paper.

In designing the simulation scenarios, we have tried to closely mimic our motivating tone learning data set. We thus chose $n = 20$ participants being trained over $T = 10$ blocks to identify $d_0 = 4$ tones. We set $\mu_{d,s}(t), b_{d,s}(t)$ to values that are very similar to the corresponding estimated values for the real data set. The local differences were all set to be in the drift curves; additionally, some boundary trajectories were globally different from each other. We slightly simplified the local clustering structure, however, to be able to better illustrate the workings of our proposed method. Moreover, we choose $u_\mu^{(C,i)}(t), u_b^{(C,i)}(t), u_\mu^{(I,i)}(t), u_b^{(I,i)}(t), \delta_s$ etc. to be the estimated posterior means obtained for the real data set.

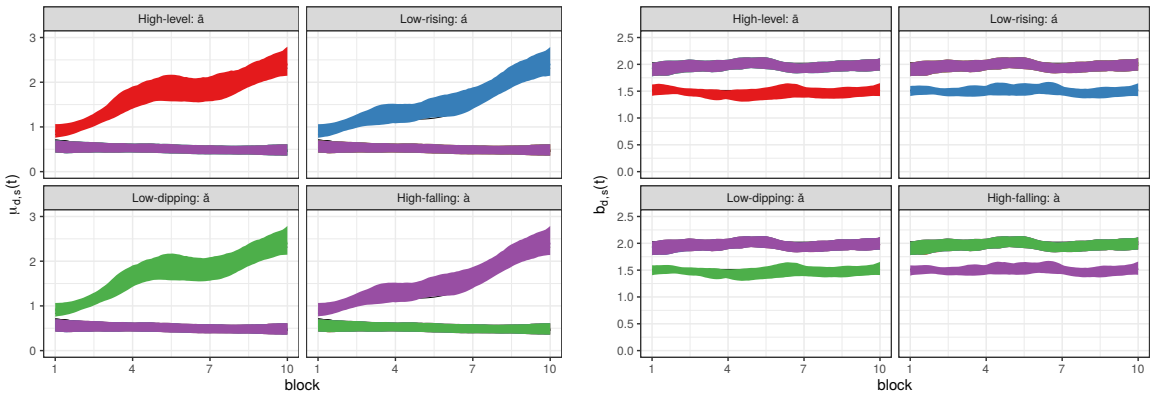

Figure S.11: Results for synthetic data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ (left panel) and boundaries $b_{d,s}(t)$ (right panel) for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. The solid black lines represent underlying true curves. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

We experimented with 50 synthetic data sets generated according to the design described above. The results produced by our method were highly stable and consistent across all data sets. The results summarized below represent a typical scenario.

Figure S.11 shows the posterior mean trajectories and associated 90% credible intervals for the the drift rates $\mu_{d,s}(t)$ and boundaries $b_{d,s}(t)$, for every possible combination of (d, s). Figure S.15 additionally presents the drift curves for successful identifications $(d = s)$ superimposed on each other. These figures suggest that the underlying true curves are all recovered well by our method. In comparison, the results obtained by the LBA model, displayed in Figure S.12, suffer from the same limitations discussed in Section 5. Furthermore,

Figures S.13 and S.14 suggest that the underlying true local partition structure, as well as the individual specific parameter trajectories, are also estimated quite well by our method.

Figure S.12 presents the results obtained by the LBA model applied to the synthetic data set. There is a general agreement between the population level estimates produced by our method and the LBA. However, as discussed in detail in Section 5 in the main paper and Section S.7 in the supplementary materials, the LBA model has many serious limitations, including being incapable of producing individual level estimates, having shared boundary parameters across all input tones, not borrowing any information across adjacent time stamps etc. Only a very limited set of inferential questions can therefore be answered by the LBA model.

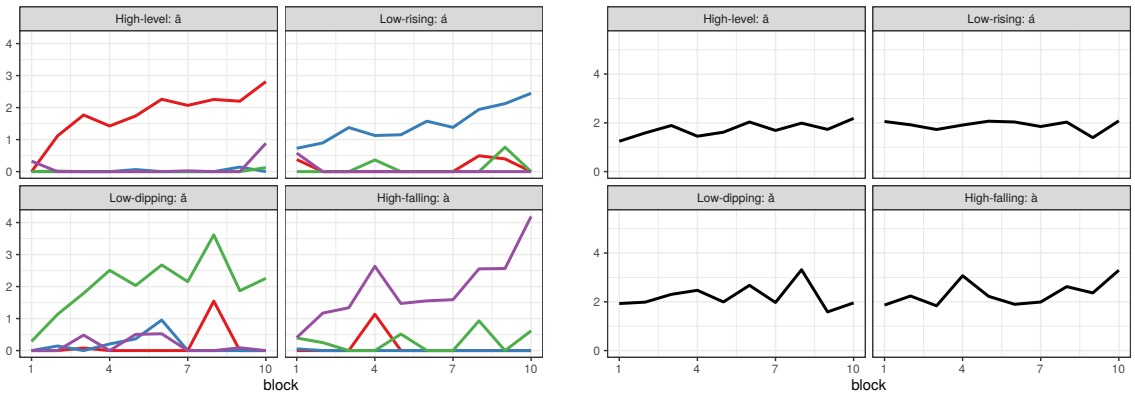

Figure S.12: Results for synthetic data: Left: Estimated mean slopes $m_{d,s,t}$ for the LBA model. Right: Estimated boundaries $b_{s,t}$ for the LBA model. In the left panel, $m_{d,s,t}$'s for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

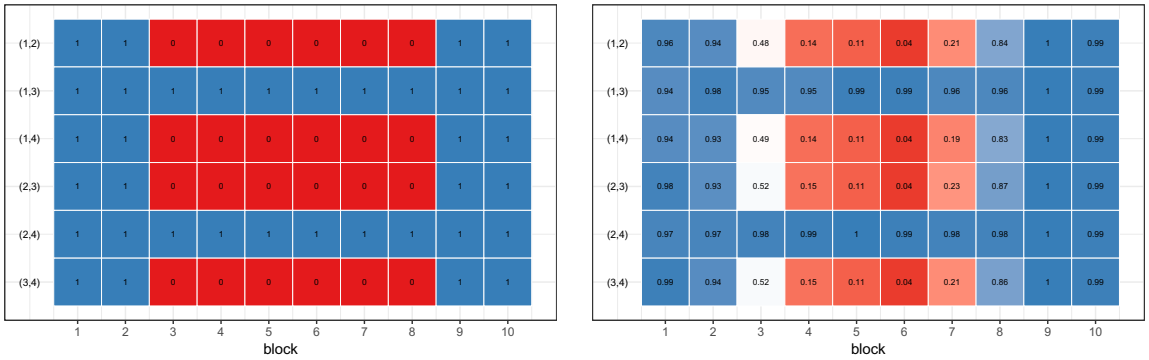

Figure S.13: Results for synthetic data: The left panel shows the true clustering structure of the underlying parameter trajectories for successful identification ($d = s$) of different input tones in different learning phases. The right panel shows the corresponding posterior co-clustering probabilities estimated by our proposed method.

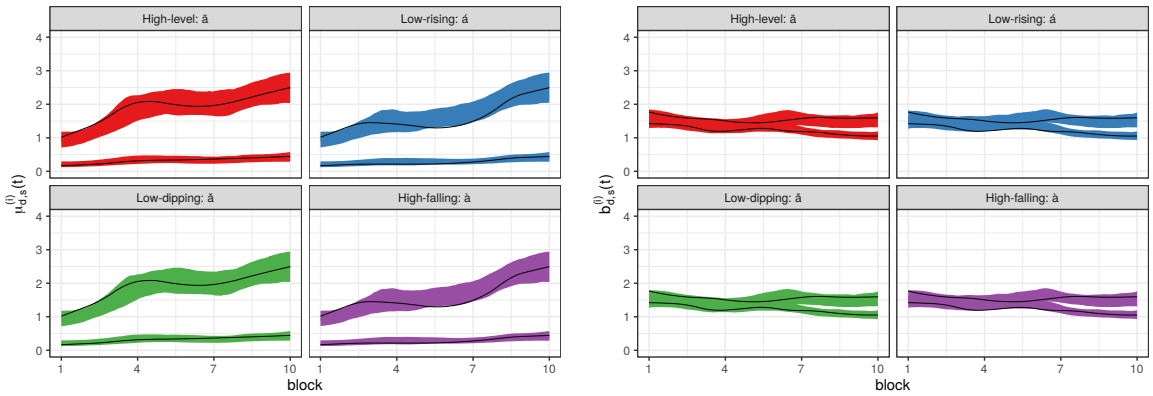

Figure S.14: Results for synthetic data: Estimated posterior mean trajectories for individual specific drifts $\mu_{d,s}^{(i)}(t)$ (left panel) and boundaries $b_{d,s}^{(i)}(t)$ (right panel) for two different participants - one performing well (dotted line) and one performing poorly (dashed line). The shaded areas represent the corresponding 90% point wise credible intervals. The solid black lines represent underlying true curves. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

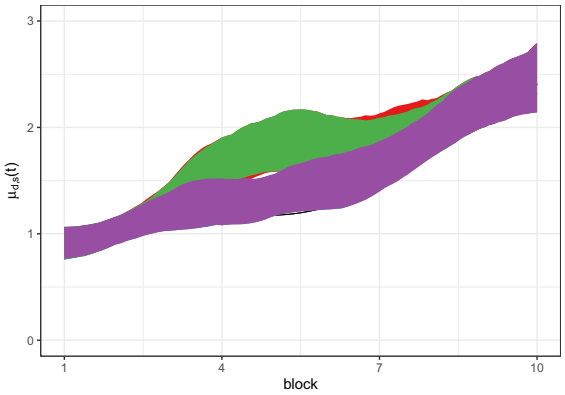

Figure S.15: Results for synthetic data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ for successful identification ($d = s$) of different input tones for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. The solid black lines represent underlying true curves. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

S.10 Additional Figures

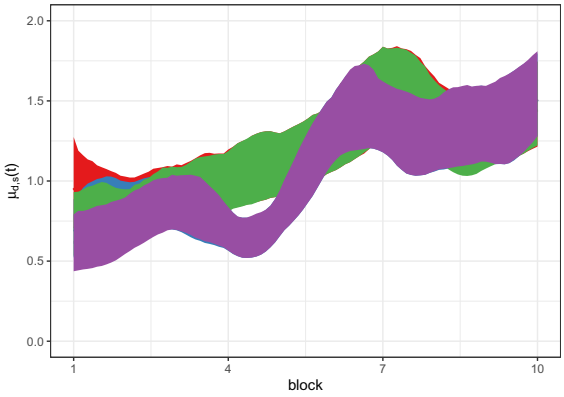

Figure S.16: Results for tone learning data: Estimated posterior mean trajectories of the population level drifts $\mu_{d,s}(t)$ for successful identification ($d = s$) of different input tones for the proposed longitudinal inverse Gaussian drift-diffusion mixed model. The shaded areas represent the corresponding 90% point wise credible intervals. Parameters for the high-level tone response category T1 are shown in red; low-rising T2 in blue; low-dipping T3 in green; and high-falling T4 in purple.

References

Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, **57**, 153–178.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics*, pages 169–193.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, **29**, 245–273.

McDonald, S. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall, London.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, **31**, 705–767.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, **77**, 257–286.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349–367.

Roberts, G. O., Rosenthal, J. S., *et al.* (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16**, 351–367.

Scott, S. L. (2002). Bayesian methods for hidden Markov models recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

Singmann, H., Brown, S., Gretton, M., and Heathcote, A. (2019). rtdists: Response time distributions. R package version 0.10-0.

Titsias, M. K. and Yau, C. (2014). Hamming ball auxiliary sampling for factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 2960–2968.

Zanella, G. (2019). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association*, pages 1–14.