STATISTICAL INFERENCE IN AUDITORY PERCEPTION

by
Benjamin Michael Skerritt-Davis

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
September 2020

Abstract

The human auditory system effortlessly parses complex sensory inputs despite the ever-present randomness and uncertainty in real-world scenes. To achieve this, the brain tracks sounds as they evolve in time, collecting contextual information to construct an internal model of the external world for predicting future events. Previous work has shown the brain is sensitive to many predictable (and often complex) patterns in sequential sounds. However, real-world environments exhibit a broader spectrum of predictability, and moreover, the level of predictability is constantly in flux. How does the brain build robust internal representations of such stochastic and dynamic acoustic environments?

This question is addressed through the lens of a computational model based in statistical inference. Embodying theories from Bayesian perception and predictive coding, the model posits the brain collects statistical estimates from sounds and maintains multiple hypotheses for the degree of context to include in predictive processes. As a potential computational solution for perception of complex and dynamic sounds, this model is used to connect sensory inputs with listeners' responses in a series of human behavioral and electroencephalography (EEG) experiments incorporating uncertainty. Experimental results point toward the underlying sufficient statistics collected by the brain, and the extension of these statistical representations to multiple dimensions is

examined along spectral and spatial dimensions. The computational model guides interpretation of behavioral and neural responses, revealing multiplexed responses in the brain corresponding to different levels of predictive processing. In addition, the model is used to explain individual differences across listeners highlighted by uncertainty.

The proposed computational model was developed based on first principles, and its usefulness is not limited to the experiments presented here. The model was used to replicate a range of previous findings in the literature, unifying them under a single framework. Moving forward, this general and flexible model can be used as a broad-ranging tool for studying the statistical inference processes behind auditory perception, overcoming the need to minimize uncertainty in perceptual experiments and pushing what was previously considered feasible for study in the laboratory towards what is typically encountered in the "messy" environments of everyday listening.

Thesis Readers

Dr. Mounya Elhilali (Primary Advisor)
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. Jason Fischer (Second Reader)
Assistant Professor
Department of Psychological and Brain Sciences
Johns Hopkins University

Dr. Hynek Hermansky
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. James West
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

To my husband Steven, for his patience,

and to Margo, for arriving at just the right time.

Acknowledgements

I have many people to thank for helping me, both directly and indirectly, get to this point. There were many uncertainties in the pursuit of my PhD, but one thing that was certain was the supportive presence of my advisor, Mounya Elhilali. Her dedication and persistent attention to our research, her patience with my many explorations, and her guidance in my career as a whole has pushed me to become a better critical thinker, a better communicator, and a better researcher. Her mentorship was the core of my graduate experience, and I am very grateful for our time working together. I'd also like to thank my dissertation committee for their useful insights and clarifying questions about my research as I prepared this dissertation: Professors Jason Fischer, Hynek Hermansky, and Jim West. They form the latest cohort in a long string of fantastic teachers who helped me appreciate the joy of learning and discovery, among them: Professors Sanjeev Khudanpur, Reza Shadmehr, Daniel Naiman, Joy Ko, Laurie Heller, Jim Valles, John Marston, and Mark Steinbach.

I was lucky to have a consistently great group of lab mates throughout my time at Hopkins. I'm very grateful to Susan Shuai, Merve Kaya, and Nick Huang for helping me learn how to use EEG; to Dimitra Emmanouilidou for helping me through my forays into undergraduate teaching; to Ashwin Bellur for providing constant commiseration as we moved through grad school together; to Sandeep Kothinti and Stephanie Graceffo for being excellent “beta-testers” for my work; and to the rest of the lab for helping create a positive and supportive workspace. I’d like to thank Kate Fischl for being a good friend and co-conspirator. I am also grateful for the help provided by two undergraduate research assistants: Katherine Simeon and Audrey Chang. And I am very appreciative of the many collaborators and researchers I met outside of Hopkins who provided generous feedback as I developed this work, especially: Maria Chait, Sijia Zhao, Malcolm Slaney, and Shihab Shamma.

I would also like to thank the people who enriched my life immensely outside of my academic pursuits. I am very lucky to have an incredible family who have been constant cheerleaders (and not just at Pride parades): my parents Joan and Mike, and my three sisters Liz, Jen, and Kim (two of which are Hopkins graduates). I am grateful for a fabulous community here in Baltimore—this city became our home almost immediately because of them—especially: Steve Martin, Joe Johnson, Abby & Peter Jackson, the Tisch-Hines, Morgan & Joe Horvath, Alish Wolf & Chas Phillips, and many more. And I’d especially like to thank Blake Zachary and Jeff Hochstetler for the extra pizzazz they’ve brought to our life in B’more.

Finally, I am beyond grateful for the many forms of love and support I’ve received from my husband, Steven. Beginning with moving down to Baltimore, he has been right by my side throughout this journey. I wouldn’t want to be here without him.

Contents

Abstract ii

Dedication v

Acknowledgements vi

Contents viii

List of Figures xi

Chapter 1 Introduction 1

1.1 Approach 4

1.2 Contributions 6

1.3 Overview 7

Chapter 2 Modeling statistical inference of sequential sounds . . . 9

2.1 Introduction 9

2.2 D-REX model 11

2.2.1 Model assumptions 11

2.2.2 Robust prediction of dynamic inputs 13
2.2.3 Model outputs 17
2.2.4 Model parameters 20
2.3 Examples from real-world audio 22
2.4 Replication of results from the literature 26
2.5 Discussion 31

Chapter 3 Statistical inference along a single dimension 36

3.1 Introduction 36
3.2 Methods 38
3.2.1 Participants 38
3.2.2 Stimuli 39
3.2.3 Procedure 41
3.2.4 EEG recording and data analysis 42
3.2.5 Model 44
3.3 Results 46
3.3.1 Perceptual experiments 46
3.3.2 Computational Model 50
3.3.3 Electroencephalography 54
3.4 Discussion 59

Chapter 4 Statistical inference along multiple dimensions 66

4.1 Introduction 66
4.2 Methods 69

4.2.1 Participants 69

4.2.2 Stimuli 70

4.2.3 Procedure 72

4.2.4 EEG data recording and analysis 73

4.2.5 Model 75

4.3 Results 76

4.3.1 Perceptual experiments 76

4.3.2 Computational model 80

4.3.3 Electroencephalography 89

4.4 Discussion 95

Chapter 5 Conclusion **100**

References **104**

Appendix I Related publications & abstracts **117**

Appendix II Statistical inference & working memory **118**

Appendix III Computer code **124**

Vita **145**

List of Figures

Figure 1-1 Examples of regularities 3

Figure 2-1 D-REX model 13

Figure 2-2 Model outputs for examples from real-world audio clips . . . 23

Figure 2-3 Model outputs for examples from real-world audio clips, cont'd 25

Figure 2-4 Replication of neural results from the literature 27

Figure 2-5 Replication of neural results from the literature, cont'd . . . 29

Figure 3-1 Random fractal stimuli 46

Figure 3-2 Psychophysics results from Experiments 1 and 2 48

Figure 3-3 Range of model behavior in Experiment 1 51

Figure 3-4 Model fit to subject behavior from Experiments 1–2 53

Figure 3-5 Contextual effects on tone ERP 56

Figure 3-6 Phase-locking analysis at model changepoints 58

Figure 4-1 Multidimensional random fractal stimuli 70

Figure 4-2 Behavioral results for experiments SP and TP 77

Figure 4-3 Multidimensional model schematic 81

Figure 4-4 Model comparison for Experiments nSP and nTP 86

Figure 4-5 Memory parameters of the `Late_D22_MAX` model 89

Figure 4-6 Linearity of surprisal response in experiment nSP and nTP . 92

Figure 4-7 Multiplexed neural responses aligned to model outputs . . . 96

Figure II-1 SI & WM: Single feature results 121

Figure II-2 SI & WM: Multifeature results across task 122

Figure II-3 SI & WM: Multifeature results across feature 123

Chapter 1

Introduction

Real-world listening environments are constantly in flux, giving rise to multiple layers of uncertainty in auditory perception. Consider a forest or a city street: each scene exhibits uncertainty due to a changing ensemble of sounds entering and exiting the scene (e.g., animal calls, rustling trees, car engines, footsteps), compounded by the uncertainty due to randomness in each individual sound source (in the pitch of a bird call or in the path of a car or pedestrian). To interpret these complex surroundings, the brain constantly sifts through all of this uncertainty, adapting to the dynamics of the scene as it evolves over time.

Sound sources often unfold as a series of discrete events, and the brain sequentially collects information from these sounds over time, gradually building up a mnemonic representation of the underlying sound source. Predictive coding theory offers an explanation for how the brain encodes past sensory information to tackle the uncertainty in dynamic scenes. Broadly, the theory proposes the brain uses the recent context to build an internal model of the external world, and this internal representation is used to make predictions of future events [1–3]. These internal representations must be invariant to the randomness inherent in real-world environments, while simultaneously allowing for flexibility to change with the dynamics of the acoustic scene. Extracting robust representations from ongoing sound is automatic and effortless for the average listener, but the underlying neural computations that accomplish this in everyday listening are largely unknown.

Invariant properties of sound sources are typically referred to in the predictive coding literature as *regularities*, and *regularity extraction* is the brain's ability to access these properties for use in auditory scene analysis [4, 5]. We differentiate between two types of regularities in sequential sounds: *deterministic regularities* that describe static characteristics or predictable patterns, and *stochastic regularities* that exist in the continuum between perfectly predictable and completely random. The key distinction lies in the presence or absence of *uncertainty*: with deterministic regularities, a new sound can immediately be interpreted as conforming to or deviating from the regularity *with certainty*, while for stochastic regularities this is not the case.

Consider, for example, the musical score in Fig 1-1, which contains various types of regularities within this short excerpt: Fig 1-1a and b highlight deterministic regularities, a single repeating note and a repeating sequence of notes, respectively; Fig 1-1c highlights an example of a stochastic regularity, where there is a statistical pattern that does not repeat exactly; and Fig 1-1d indicates a segment with stochastic regularities that have more randomness and are less visually apparent. The brain is remarkably sensitive to this range of predictability in music, and, although music is highly structured compared to everyday scenes, this ability to extract and exploit regularities in sequential sounds is used broadly in auditory perception in general.

Typically, studies in predictive coding manipulate listener expectations by embedding regularities in sequences of sounds, and behavioral and neural responses are

Figure 1-1. Examples of different types of regularities embedded in a musical excerpt. Deterministic regularities (a and b) are repeating patterns that can be interpreted with certainty. Stochastic regularities (c and d) can only be described abstractly, and involve some level of uncertainty. The regularity in c) comprises of near repetitions transposed down by a single step, while the regularity in d) is less visually apparent.

examined at violations of or changes in these regularities; the oddball paradigm is the prototypical example of this in the literature [6, 7]. Previous studies have shown the brain is sensitive to a vast array of deterministic regularities in sound sequences, from simple repetitions to more complex patterns, for example: two interleaved deterministic sequences [8], an abstract pattern within a single acoustic feature ("falling pitch within tone-pairs"[9]) or one spanning multiple features ("the higher the pitch, the louder the intensity"[10]). Studies using stochastic regularities have demonstrated that listeners can discriminate between sound sequences based on statistical structure using both behavioral and neural responses [11–13], that neural responses to deviance are modulated by increases in uncertainty modulate [14–16], and that the brain is sensitive to Markov structure within small sets of stimuli [17–19]. One possible mechanism for how the brain represents stochastic regularities is through statistical estimates, which entails extracting representative parameters from observed sensory cues [20, 21]. However, the nature and extent of statistics collected by the brain is an open question.

The aim of this dissertation is to investigate how the brain uses statistical representations to interpret real-world sounds, where regularities exhibit a broad spectrum of predictability. How does the brain build robust internal representations from such stochastic and dynamic sensory inputs?

1.1 Approach

To investigate the predictive processing of dynamic stochastic sounds, we use a combination of human behavioral and electroencephalography (EEG) experiments alongside computational modeling. With the certainty afforded by deterministic regularities, the connection between inputs (i.e., stimuli) and outputs (i.e., responses) is straightforward; however, as uncertainty is introduced into the experimental paradigm, uncertainty unavoidably manifests in the experimental data collected. Stochastic regularities render the connection between stimuli and response (especially neural responses) *very* tenuous. This complexity necessitates the use of a computational model to guide both the analysis and interpretation of behavioral and neural responses to stochastic stimuli.

We developed a novel computational model that incorporates Bayesian theories of predictive processing, incrementally predicting future sensory inputs given the preceding context [5, 22–24]. From sequential inputs extracted from audio along any continuous-valued acoustic or perceptual dimension (e.g., pitch, spatial location, spectral centroid), the model outputs a probabilistic prediction of the next input given its context. Just as in natural listening scenarios, the model does not assume stationarity in the incoming sound; rather, it infers the amount of context from the observed inputs. Additionally, the model outputs measures of prediction mismatch and posterior beliefs that are easily interpretable in terms of predictive coding theory. We use this model to compare different internal representations in predictive processing to behavioral responses, and in turn use the model to guide analysis of neural responses.

We applied this model in a series of human experiments to examine predictive processing of stochastic regularities in sequential sounds. Stimuli were sound sequences exhibiting random fractal structure (also known as $1/f$ or power-law noise), which is notable for its ecological relevance, as such structures have been found in music [25], speech [26], and natural sounds [13]. We used a change detection paradigm, tasking listeners with detecting changes in entropy of stimuli sequences. This paradigm mirrors the challenges presented to the auditory system in everyday listening, where the dynamics of emergent regularities must be inferred from sensory inputs.

The general experimental approach was as follows: We first established through behavior the extent of listeners' ability to detect statistical structure embedded in stochastic sounds, and we used the proposed computational model to test alternative computational mechanisms that could give rise to these behaviors. We then used the computational model to further interpret behavioral differences across individual listeners and analyzed neural responses in similar experiments. We applied this experimental approach first using stimuli that varied along a single dimension (pitch), and we then expanded this approach to investigate the perception of sounds that evolve along multiple dimensions simultaneously.

1.2 Contributions

The goal of this dissertation is to expand our understanding of the mechanisms behind predictive processing of sequential auditory inputs in the presence of uncertainty. The main contributions can be summarized as follows:

(i) The computational model provides a framework for probing specific components of predictive processing. Rather than being developed for a specific paradigm or domain of stimuli, the model was designed using first principles from predictive coding theory. This gives the model broad applicability to interpret predictive processing of sequential sounds, not only with the controlled stimuli typically used in perceptual experiments, but in music and speech listening as well, all under a unified computational framework. We demonstrate several uses of the model—namely, to test alternative computational mechanisms giving rise to individual behavioral and neural responses—but the usefulness of the model goes beyond the experimental studies described in this dissertation. We additionally used the model to replicate a range of existing results from the predictive coding literature, and we explored the model's flexibility in interpreting various real-world audio examples using different statistical representations along a variety of input dimensions.

(ii) The extent to which the brain collects statistics from sequential sounds has not been sufficiently explored in previous work. Aided by the model, human behavioral and neural evidence establish that the brain collects higher-order temporal dependencies between sounds as they unfold over time. Moreover, these statistics are collected independently across multiple dimensions simultaneously.

(iii) The behavioral paradigm reveals individual differences in the perceptual system that are amplified by uncertainty from statistical inference processes. Through the lens of the model, variability across listeners was interpreted in terms of individual perceptual and memory limitations that are not directly accessible through listeners' behavioral or neural responses.

(iv) Uncertainty also leads to trial-by-trial variability in response timing, which is particularly problematic for time-locked analyses in EEG, where low SNR necessitates many repetitions and precise temporal alignment across trials and subjects to obtain meaningful results. To account for variability due to the stochastic nature of each stimulus, EEG epochs are anchored according to model outputs to reveal neural responses time-locked to the underlying predictive processes.

1.3 Overview

The remainder of this dissertation is organized in three main chapters, each building on the results from the previous chapters.

Chapter 2 presents a description of the proposed computational model in its entirety, without target application or experimental paradigm. This chapter includes two demonstrations of the generality of the model: (i) illustrations of model outputs in response to a variety of real-world audio examples to inspire deeper inquiry into predictive processing of natural sounds, and (ii) replication of various results from the predictive coding literature under the same computational framework.

In Chapter 3, an experimental paradigm for investigating statistical inference along a single dimension is developed. Behavioral evidence for statistical processing is presented, and the computational model from Chapter 2 is applied to determine the internal statistical representation that best explains experimental results. The model is then used to interpret neural responses.

In Chapter 4, the experimental paradigm from Chapter 3 is expanded to investigate statistical inference along multiple dimensions (pitch, timbre, and spatial location). Behavioral results demonstrate listeners' ability to flexibly exploit and integrate stochastic regularities across spectral and spatial dimensions, and the model is used to compare many hypotheses for how this integration occurs. Neural responses reflect different levels of predictive processing revealed by the model.

Finally, Chapter 5 synthesizes these results and offers potential avenues for future work.

Chapter 2

Modeling statistical inference of sequential sounds

2.1 Introduction

Computational modeling has been used previously to expand the realm of investigation in predictive coding. It has facilitated the interpretation of trial-by-trial variability in listener responses [27], the link between individual spiking neurons and neural responses to deviance measured at the scalp [28], and the recasting of various listening phenomena, such as streaming and object perception, in terms of predictive coding [22, 29, 30]. Computational modeling is particularly useful for studying statistical processing in the brain, where stimulus-driven analyses are often constrained by uncertainty in the stimulus and in the elicited response [14, 15, 31]. A common limitation of these models is that they are designed for a particular experimental paradigm. One notable exception is the IDyOM model, initially formulated for musical expectation [32], which has been used to decode neural responses to music [33, 34] as well as describe statistical learning of sound sequences in general [19, 35]. Additionally, the ARTSTREAM model, based on Gestalt principles of perception, incorporates predictive coding into a broader framework for auditory scene analysis [36]. These models, however, place various limitations on the domain of sensory inputs: the IDyOM model operates on a discrete set of inputs (i.e., an alphabet), ignoring any ordering or distance between elements, and the ARTSTREAM model assumes smoothness and harmonicity. These provisions hinder the ability of these models to apply broadly across different listening scenarios or explore the internal representations used in predictive processing *in general*.

In this chapter, we present a computational model that provides a potential algorithmic solution for the predictive processes employed in everyday listening. It is agnostic to experimental paradigm or listening scenario and makes minimal assumptions on the sensory input, instead offering a framework to compare different assumptions and internal representations in the brain using experimental responses. This model is grounded in theoretical accounts of predictive coding based in Bayesian inference [37–39], and its mathematical underpinnings have previously been explored in predictive-inference tasks using sequences of numbers [40, 41]. In lieu of modeling neural mechanisms directly, we use neurally plausible computations to model the cognitive processes that map sensory inputs to decision and action. This approach favors simplicity in relating model inputs, outputs, and parameters to perceptual processes, facilitating the exploration of underlying predictive mechanisms and their connection to neural and behavior responses in a broad range of experimental studies and realistic listening environments.

This chapter is organized as follows. First, we describe the model in its general form, along with use cases relating the model to various experimental paradigms employed in auditory research. Then, to demonstrate the flexibility of the model in capturing different statistical structures in auditory inputs, we illustrate the predictive processing of real-world audio examples along a variety of input dimensions. Finally, we use the model to replicate and reinterpret existing results from the predictive coding literature under a unified framework.

2.2 D-REX model

The Dynamic Regularity Extraction (D-REX) model is a computational model for predictive processing of sequential sounds. This model has its roots in Bayesian changepoint detection [42], which has previously been cast as a neurally plausible framework for predictive processing of sensory inputs in the brain [40]. In this section, we describe the model in general terms with ideas interspersed regarding possible applications of the model to specific experimental paradigms. Source code for the D-REX model is available online at `http://www.github.com/jhu-lcap/drex-model`, as well as in Appendix III.

2.2.1 Model assumptions

The D-REX Model builds a predictive distribution at time t, Ψ_t, for the next input x_{t+1} given all previously observed inputs:

$$\Psi_t = \mathbb{P}(x_{t+1}|x_{1:t})$$

where the input observations $\{x_t\}_{t\in\mathbb{Z}^+}$ are continuous-valued and sampled discretely in time. The input sequence $\{x_t\}$ can be any acoustic or perceptual feature extracted from the acoustic waveform (e.g., pitch, RMS energy, spectral spread, loudness, spatial location). For example, the input to the model could be the sequence of pitches extracted from a melody.

The input sequence is assumed to be stochastic, drawn from a probability distribution f with unknown parameters θ, i.e., at each time t, $x_t \sim f_\theta$. For example, if f is a univariate Gaussian distribution, θ would be the unknown mean and variance. While the form of the distribution f is constant, the model does not assume stationarity in this distribution, i.e., the parameters θ can change at unknown times. Fig 2-1a shows an example input sequence generated from a Gaussian distribution with two changes in the parameters θ (changes indicated by arrows). The D-REX model currently includes built-in support for the following distributions: Gaussian, Log-normal, Gaussian mixture, and Poisson; note that this list is not exhaustive, and additional distributions can be easily incorporated into the model code.

With Gaussian and Log-normal distributions, the distribution is additionally specified by D, the extent of temporal dependence between observations. For $D > 1$, the model assumes successive observations are drawn from a joint distribution with dimensionality D, and the form of the unknown parameters θ reflect this dependence. For example, a multivariate Gaussian distribution with $D = 2$ assumes dependence (and non-zero covariance) between adjacent observations, while with $D = 1$, observations are assumed to be statistically independent. As D increases, the model assumes temporal dependence across wider spans of the input observations.

The choice of distribution f (and temporal dependence D) is crucial, as they determine what statistical structures are captured by the model. When modeling perceptual processes, the choice of distribution represents an implicit hypothesis that the brain is sensitive to these same statistical structures or regularities, therefore it can be used to compare different internal representations in the brain.

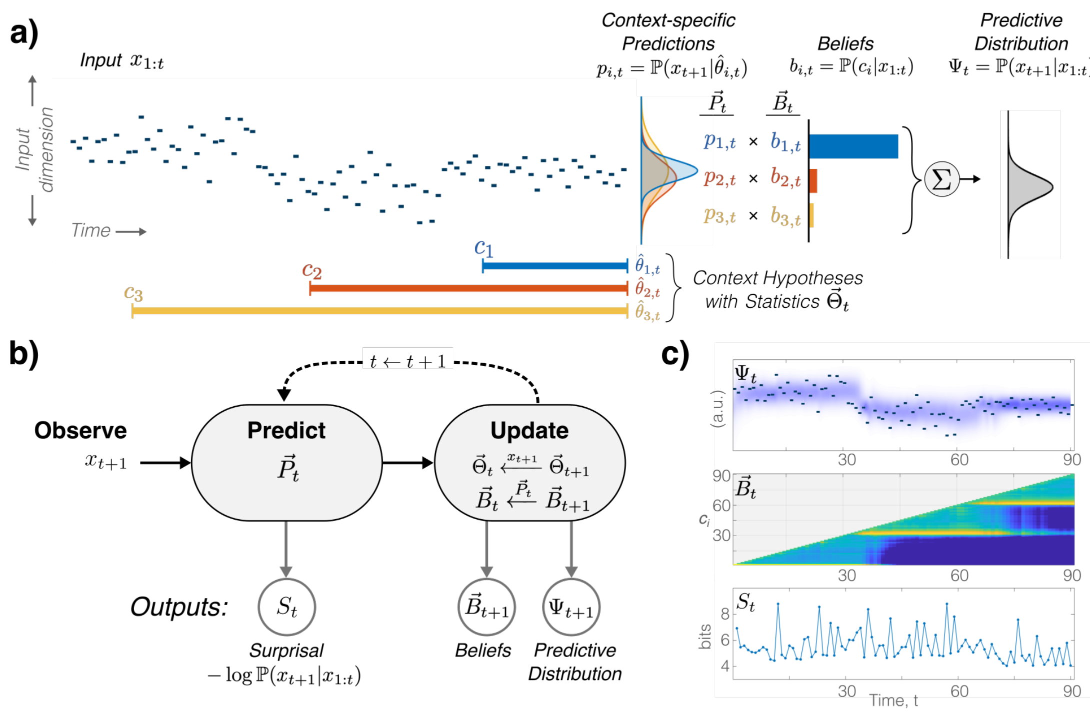

Figure 2-1. Model description. a) The model uses multiple context hypotheses to account for unknown changes in the observed sequence. Context-specific predictions $\vec{P}_t$ based on sufficient statistics $\vec{\Theta}_t$ are combined, weighted by corresponding beliefs $\vec{B}_t$, to yield the predictive distribution Ψ_t for the next input x_{t+1}. b) Upon observing x_{t+1}, the predictions and new input are used to update the statistics and beliefs, which are used in turn to predict the next input, and so on. There are three principal outputs from the model at each time: the surprisal of the newly observed input based on its prediction, the predictive distribution for the next input, and the beliefs (or posterior distribution over contexts). c) Outputs from the model for the example sequence in a). Note the predictive distribution and beliefs reflect the underlying change in statistics inferred by the model.

2.2.2 Robust prediction of dynamic inputs

The model makes minimal assumptions on the input sequence, constraining only the parametric form of the generating distribution but not the parameters themselves. The challenge is then to make predictions of the next input x_{t+1} that are robust both to unknown dynamics in the underlying generating distribution and to uncertainty stemming from stochastic inputs.

Sufficient statistics $\hat{\theta}$

The model represents past predictive information via sufficient statistics $\hat{\theta}$ collected from the observed inputs. These sufficient statistics are estimates of the unknown parameters θ and depend on the distribution choice f: for example, a Gaussian distribution with $D = 1$ has sufficient statistics $\hat{\theta}$ comprised of the sample mean and sample variance. The prediction from the model then depends on these statistical estimates in lieu of the past observations themselves:

$$\mathbb{P}(x_{t+1}|x_{1:t}) = \mathbb{P}(x_{t+1}|\hat{\theta}_t) \tag{2.1}$$

where $\hat{\theta}_t$ are the sufficient statistics for distribution f estimated from the previous observations $x_{1:t}$. Here, we refer to the extent of past observations used to estimate statistics $\hat{\theta}$ as the *context window* for the prediction.

Multiple hypotheses for the unknown context

Because the dynamics of the underlying distribution are unknown, the choice of context window impacts the quality of the prediction. For example, if the underlying parameters θ have changed at any point in the observed sequence, collecting sufficient statistics $\hat{\theta}$ over a context that includes *all* past observations will result in poor statistical estimates of the current parameters. Without *a priori* knowledge of when these changes occur, the model must infer the appropriate context window from the data. To do this, the model makes predictions using multiple contexts simultaneously, each referred to as a *context hypothesis*. Each hypothesis forms a potential parsing of the past into observations that are relevant for the current prediction and those that are not.

Let the set of context hypotheses be $\vec{C} = \{c_i\},\ i \in \{1, \ldots, M\}$, where c_i is the leading boundary of the i^{th} context and M is the total number of hypotheses. At each time t, the model maintains a corresponding set of sufficient statistics collected over each context, $\vec{\Theta}_t = \{\theta_{i,t}\}$, and produces a set of predictions for the next observation given each context, $\vec{P}_t = \{p_{i,t}\}$. For the i^{th} context hypothesis:

$$
\begin{aligned}
p_{i,t} &= \mathbb{P}(x_{t+1}|c_i, x_{c_i:t}) \\
&= \mathbb{P}(x_{t+1}|\hat{\theta}_{i,t})
\end{aligned} \tag{2.2}
$$

where c_i, $\hat{\theta}_{i,t}$, and $p_{i,t}$ are the i^{th} context boundary, the statistics collected over that hypothesis, and the context-specific prediction based on these statistics, respectively. Note that compared to Eq.Eq. (2.1), the context-specific prediction of x_{t+1} in Eq.Eq. (2.2) only depends on observations after the context boundary c_i, because observations before c_i are independent of x_{t+1}.

The model also maintains a set of *context beliefs* $\vec{B}_t = \{b_{i,t}\}$, each representing the evidence for the i^{th} context at time t given all previously observed inputs:

$$
b_{i,t} = \mathbb{P}(c_i|x_{1:t}) \tag{2.3}
$$

These beliefs form a discrete posterior distribution over context hypotheses.

By default, the model produces a new context hypothesis at each time-step, entertaining the possibility of a change at *any time*. This can be adjusted using the input parameters of the model to represent prior knowledge about when changes occur or to decrease computational cost of maintaining an exhaustive set of context hypotheses.

"Integrating out" the unknown context

To build the predictive distribution Ψ_t, the context-specific predictions $p_{i,t}$ are combined across context hypotheses, weighted by their beliefs $b_{i,t}$ (see Fig 2-1a-right). We then have the final predictive distribution at time t:

$$\begin{aligned}\Psi_t = \mathbb{P}(x_{t+1}|x_{1:t}) &= \sum_{i=1}^{M} \mathbb{P}(x_{t+1}, c_i|x_{1:t}) \\ &= \sum_{i=1}^{M} \mathbb{P}(x_{t+1}|c_i, x_{c_i:t})\mathbb{P}(c_i|x_{1:t}) \\ &= \sum_{i=1}^{M} p_{i,t} b_{i,t} \end{aligned} \tag{2.4}$$

This weighted summation "integrates out" the unknown context in a Bayesian fashion, building a prediction for x_{t+1} that adapts to changes in the underlying statistics of the observed sequence.

Fig 2-1a shows an illustration of how the model builds the prediction for x_{t_1} given an example input sequence $x_{1:t}$ using three context hypotheses (with leading boundaries c_1, c_2, c_3 and statistics $\hat{\theta}_{1,t}$, $\hat{\theta}_{2,t}$, $\hat{\theta}_{3,t}$). Context-specific predictions ($p_{1,t}$, $p_{2,t}$, $p_{3,t}$) show how the distributions differ by context, and the beliefs ($b_{1,t}$, $b_{2,t}$, $b_{3,t}$) show the relative evidence for the three context hypotheses at time t. In this example, the model uses a Gaussian distribution with $D = 1$ (i.e., no temporal dependence). Note that c_1 is the only context that does not span an unknown change in distribution parameters θ: its prediction $p_{1,t}$ more closely matches the statistics of the recently observed inputs, and it has the highest belief $b_{1,t}$. The final predictive distribution Ψ_t is a weighted summation of the context-specific predictions.

Iterative processing

Fig 2-1b shows the main processing stages that the model undertakes in each time-step:

Observe. The new input x_{t+1} is observed.

Predict. The probability of x_{t+1} under each context hypothesis is computed using the context-specific predictive distributions $\vec{P}_t$ (see Eq Eq. (2.2)).

Update. Sufficient statistics $\vec{\theta}_t$ are updated with the newly observed input [43], and beliefs $\vec{B}_t$ are updated using the predictive probabilities [42].

The updated statistics and beliefs, $\vec{\Theta}_{t+1}$ and $\vec{B}_{t+1}$, are used in turn to process the subsequent input x_{t+2}, and so on.

2.2.3 Model outputs

There are three main outputs from the model, as shown in Fig 2-1b, which can each be used to relate the model to behavioral and neural responses in various experimental paradigms. Importantly, the model is causal, so all outputs depend only on previously observed inputs.

(i) S_{t+1} is the **surprisal** of the input x_{t+1}. After x_{t+1} has been observed, the surprisal S_{t+1} indicates the mismatch between this observation and its predictive probability:

$$S_{t+1} = -\log \mathbb{P}(x_{t+1}|x_{1:t}) \tag{2.5}$$

where the probability is computed from Eq Eq. (2.4). Observations with a low probability of occurring have high surprisal, whereas those with a high probability have low surprisal, and observations with probability 1 (i.e., completely predictable) have zero surprisal. The term *surprisal* used here is related to information content, or the information gained when a random variable is observed [44].

Surprisal is analogous to a probabilistic deviance response. In particular, surprisal can be related to the Mismatch Negativity (MMN) in electrophysiology responses (for comparisons of D-REX surprisal to MMN results in the literature, see [45]). Surprisal can also be related to discrimination paradigms where the contrastive property in the stimulus relates to predictability. For example, average surprisal can be used to discriminate between sequences with different entropy [11, 35].

(ii) Ψ_{t+1} is the **predictive distribution** of the next observation x_{t+2}, or the weighted sum of context-specific predictions (see Eq Eq. (2.4)). As a probability distribution, quantities such as the expected value (i.e., the predicted value of the next input), the entropy, or the precision can be derived from Ψ_{t+1} and used to connect neural event-related or oscillatory responses to specific aspects of prediction [46–48]. For example, the predictive distribution can be used to examine the evolution of precision-weighted EEG responses in the brain [35].

(iii) $\vec{B}_{t+1}$, the **beliefs**, forms the posterior probability distribution over context hypotheses (see Eq Eq. (2.3)). The beliefs represent the relative evidence across context hypotheses. Similar to the predictive distribution, measures can be derived from the beliefs to relate to behavioral and neural respones, e.g., the expected context at time t: $\mathbb{E}[c_i] = \sum_{i=1}^{M} c_i b_{i,t}$.

Beliefs can be particularly useful in change detection paradigms. For example, the beliefs can be used to compute the probability at least one change has occurred in the observed sequence, or equivalently, the probability that the context boundary occurs *after* the beginning of the observed sequence:

$$\mathbb{P}(\text{Change}) = \mathbb{P}(c_i > 1 | x_{1:t+1}) = \sum_{i:\, c_i > 1} b_{i,t+1}$$

Or, the beliefs can be used to define a moment-by-moment measure of how much the beliefs shift at each time to adapt to changing statistics:

$$\delta_t = D_{\text{JS}}(\vec{B}_t \mid\mid \vec{B}_{t+1})$$

where $D_{\text{JS}}(\cdot||\cdot)$ is the Jensen-Shannon divergence, or the distance, between beliefs before and after observing x_{t+1}.

To relate model outputs to behavioral responses, a threshold can be applied to any of these measures of change to acquire a binary change-detection decision from the model. This decision response can then be used to fit the model to listener behavior (for example, see [49]). In this case, the threshold represents an additional parameter of the model, where decreasing the threshold results in increased sensitivity in the model to change, and vice-versa.

Fig 2-1c displays model outputs for the example sequence in Fig 2-1a as they evolve over time. Note this same visual representation of the model outputs will be used in the Examples section below. The predictive distribution (Fig 2-1c-top) adapts to changes in the input observations. These correspond to shifts in the context beliefs (Fig 2-1c-middle), displayed as vertical slices at each time t, with color corresponding to the log-probability of each context boundary c_i on the vertical axis. For example, interpreting the vertical slice at $t = 60$ from the bottom-up, beliefs indicate very low probability for context hypotheses with $c_i < 30$, a peak around $c_i = 30$, and medium probability for $c_i > 30$, indicating the context hypothesis with $c_i = 30$ has the highest belief at time $t = 60$ given previous observations (note this matches the ground truth for the most recent change in the input sequence). The diagonal boundary reflects the causal nature of the model: at each time t, there are only context hypotheses with boundaries c_i in the past (i.e., $c_i \leq t$). The surprisal (2-1c-bottom) shows the momentary mismatch of every input after it has been observed. Note that higher surprisal corresponds with observations that fall farther outside of the predictive distribution in the top plot.

The use-cases of the D-REX model mentioned above are not exhaustive, nor are the three principal outputs of the model—surprisal, prediction, beliefs—the extent of possible responses produced by the model. They are presented here as the basic building blocks of the model's response which can be used to derive application-specific outputs to interpret a variety of experimental paradigms and listening tasks related to predictive processing.

2.2.4 Model parameters

The parameters of the D-REX model (not to be confused with the unknown distributional parameters θ) have straightforward interpretations in terms of prior knowledge, individual differences in neural resources, and the underlying computational implications for predictive algorithms in the brain. These parameters give the D-REX model flexibility to serve multiple purposes, from asking specific questions about perceptual processes to tailoring the model to fit behavior of individual subjects.

Priors: π

The priors π are the initial statistical estimates for a new context hypothesis and take the same form as the sufficient statistics $\hat{\theta}$. These priors represent any "prior knowledge" in the model regarding the statistics of the input sequence after a change *before any new inputs have been observed.* In most cases, the priors can be set to sufficient statistics estimated from exposure stimuli with the same statistical properties as the target stimuli. Or the priors can be used to test hypotheses about how prior knowledge affects predictions: for example, how different long-term prior experience affects the listener responses to the same inputs, or how trial-to-trial learning evolves over the course of an experiment.

Hazard rate: h_t

The hazard rate h_t is the probability of a change in underlying statistics occurring at time t *before* any inputs after time t have been observed. If the hazard rate h_t is greater than zero, a new context hypothesis is created at time t with belief equal to h_t, i.e., $b_{1,t} = h_t$. The larger h_t is, the more volatility and change is assumed in the underlying statistics of input. The hazard rate can be constant, i.e., changes in the unknown parameters θ are equally probable at all times, or it can vary over time, encompassing prior knowledge about when changes are expected to occur in the input sequence.

Perceptual parameters: M**,** N

Previous studies have shown that human listeners do not operate as ideal Bayesian Observers [50]. Two perceptual parameters represent neurally plausible constraints to predictive processing in the model:

Memory $\boldsymbol{M}$ is the total number of context hypotheses and represents working memory capacity constraints in the brain [51, 52]. If context hypotheses are created at each time-step (i.e., if $h_t > 0,\ \forall t$), M also represents the maximum context window used by the model to generate predictions, or equivalently, the maximum sample size used to estimate statistics $\hat{\theta}$.

Observation noise $\boldsymbol{N}$ sets a lower bound on prediction uncertainty, representing limitations in perceptual fidelity along the input dimension [53, 54]. Observation noise is equivalent to adding independent Gaussian noise to the observed input with zero-mean and constant variance N, which has the effect of both increasing uncertainty of the prediction *and* decreasing precision of the sufficient statistics $\hat{\theta}$.

Both of these perceptual parameters affect predictive processing and can be used to fit the model to individual listener behavior by defining a model response analogous to the listener response and performing a parameter search to find the parameters that best replicate listener response. An example of this can be found in [49].

2.3 Examples from real-world audio

To illustrate the flexibility of the D-REX model, in this section we show model outputs for example inputs taken from real-world audio clips. Audio examples were selected to represent a range of real-world sound sources from music, speech, and environmental sounds. Across the examples, we demonstrate the model's capacity to capture a variety of statistical structures along an assortment of input dimensions related to spectral, spatial, and temporal processing.

Each panel in Fig 2-2 and 2-3 shows the input sequence (top, in black) with the three model outputs as they evolve over time: predictive distribution (top, in blue), beliefs (middle), and surprisal (bottom). The input feature and distribution used in the model are indicated above each example with annotations of audio events therein. All audio clips were downloaded from publicly available sources, and input sequences for the model were extracted from the acoustic waveform using custom MATLAB scripts.

In each example, an "ideal-observer" model was used with zero observation noise and infinite memory parameters. The distributional choice f (and temporal dependence D, when applicable) was chosen based on the input dimension and/or to illustrate

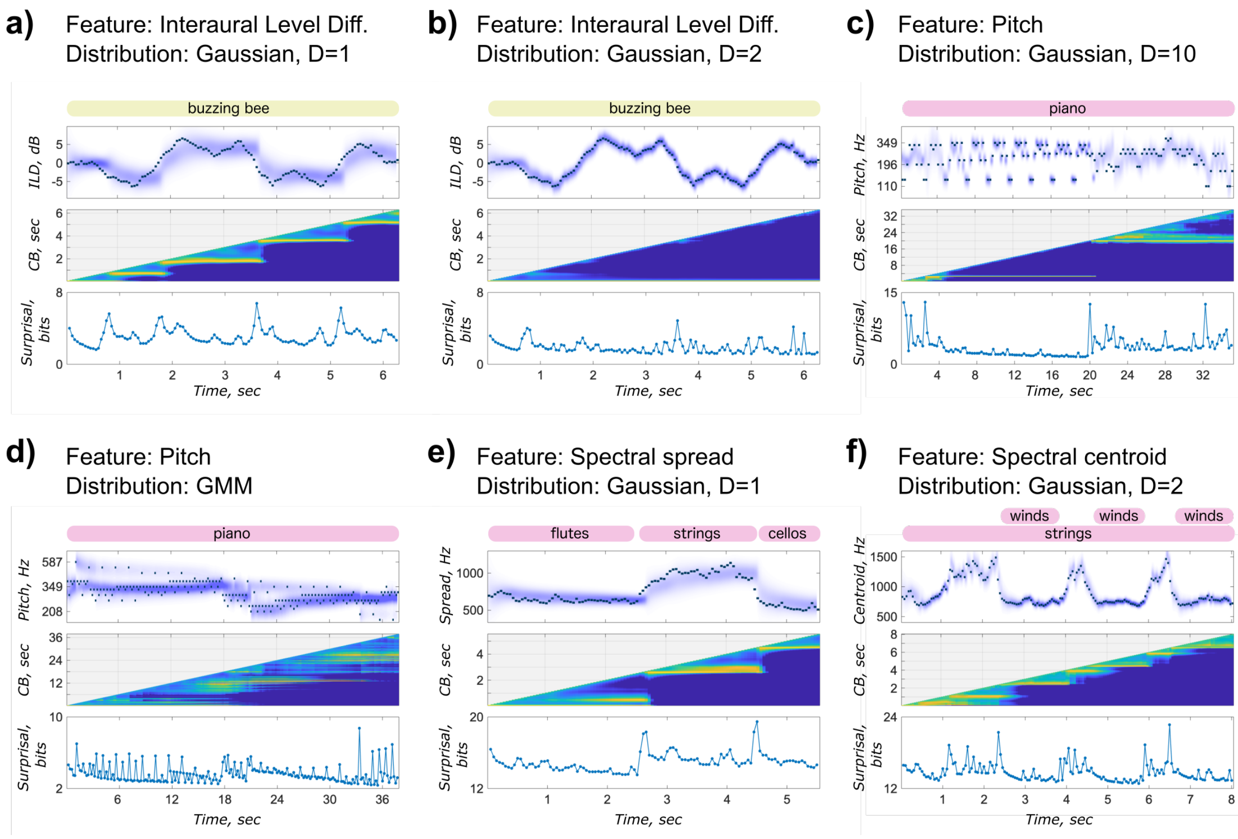

Figure 2-2. Model outputs for example inputs from real-world audio clips. Each panel displays the model predictive distribution (top), context beliefs (middle), and surprisal (bottom) over time, with the input sequence overlaid on the predictive distribution (top, in black). The input dimension (feature), distributional choice in the model, and audio event annotation are indicated above. Includes examples employing Gaussian and Gaussian mixture distributions.

the impact of this choice on the outputs from the model. Examples are organized according to the input dimension.

Spatial location. Fig 2-2a and b show model outputs from a binaural recording of a buzzing bee flying around the head. As an acoustic surrogate for spatial location, the input dimension used here is the Interaural Level Difference (ILD-dB), the dB-ratio of root-mean-squared (RMS) energy between the left and right channels in 50 ms analysis frames. Both Fig 2-2a and b use a Gaussian distribution in the model, but differ in the temporal dependence D. In Fig 2-2a, the model assumes no temporal dependence ($D = 1$), and statistical changes are apparent in the prediction and in the beliefs as the input deviates from the running mean, which can also be seen in peaks in surprisal. In this case, the model interprets the input as a series of segments with static mean and variance; the clear "staircase" image in the beliefs shows this segmentation.

In contrast, when temporal dependence is incorporated as in 2-2b ($D = 2$), no changes are apparent. Here, the model collects covariances between adjacent inputs, tracking the trajectory of the sequence along the input dimension. Note that the precision of the prediction is much higher compared to Fig 2-2a. This offers an alternative interpretation of the same input sequence.

Pitch. Fig 2-2c and d show model outputs from two Bach melodies. Pitch was extracted from source MIDI files using the MATLAB-MIDI toolbox[1]. Pitches are represented in semi-tones to reflect logarithmic tonotopy in the auditory system. Fig 2-2c uses a Gaussian distribution again but with much longer temporal dependence ($D = 10$). The large covariance structure collected by the model is sensitive to the arpeggiated melody in the first half of the input sequence, as can be seen in the coalescing of the prediction around the input, as well as in the low surprisal. The model then adapts to the change in melody motif around $t = 20$. Note that because the model uses statistical representations, exact repetitions were not necessary to capture the regularity in the first half of the sequence.

In Fig 2-2d, the model uses a Gaussian mixture model (GMM) to represent the pitches of another Bach melody. While this distribution does not have temporal dependence, it is more flexible for representing arbitrary distributions in the input.

[1] https://github.com/kts/matlab-midi

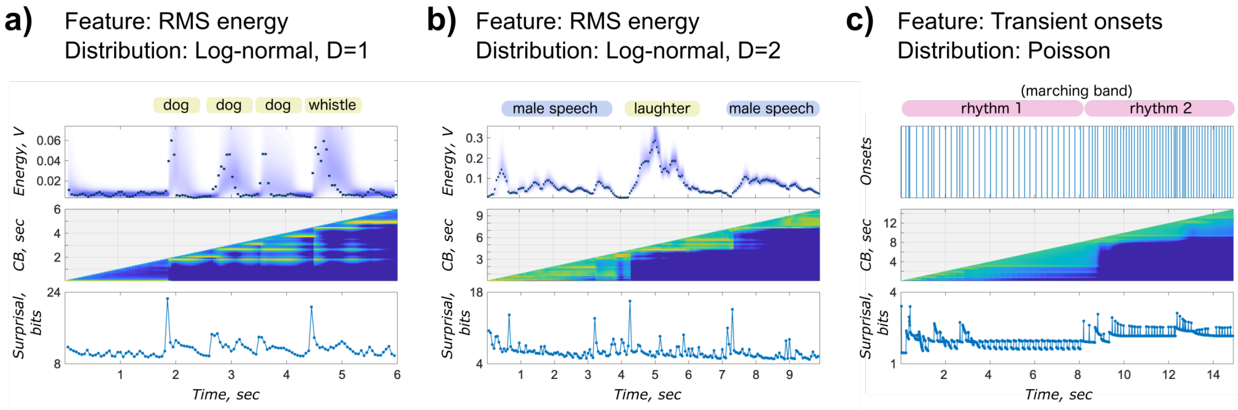

Figure 2-3. Model outputs for example inputs from real-world audio clips, continued. Similar layout to Fig 2-2. Includes examples employing log-normal and Poisson distributions.

The prediction captures the multimodal nature of the input and adapts gradually to changes in the statistics, as can be seen by the dispersal of beliefs across multiple contexts. Note that the peaks in surprisal coincide with lower-probability observations in the high component of the sequence, but the overall surprisal trend is downward, as the model builds better estimates of the underlying statistics.

Spectral profile. Fig 2-2e and f use Gaussian distributions to process two spectral features from orchestral performances: spread and centroid. These spectral features were derived from the cochleogram, a physiologically-inspired spectrogram computed from the acoustic waveform as part of the NSL toolbox[2], using 50 ms analysis frames. With both features, changes in orchestration (i.e., which instruments are playing at each moment) are reflected in the beliefs from the model. These two examples demonstrate how the model can be used to track timbre in the acoustic input.

Energy. Fig 2-3a and b apply a log-normal distribution to the RMS energy measured in frames from two everyday recordings. RMS energy was computed directly from the acoustic waveform in 50 ms analysis frames. In Fig 2-3a, peaks in surprisal correspond

[2] http://nsl.isr.umd.edu/downloads.html

with dog barks and a whistle. Note that the surprisal of the first dog bark is higher than the later events, a consequence of the statistics of the preceding context. In Fig 2-3b, the beliefs capture turn-taking in conversational speech between a male speaker and group laughter.

Onset timing. The final example in Fig 2-3c applies the model to a temporal dimension: the timing of transient onsets extracted from a recording of a marching band drum line. Transient onsets were extracted by finding peaks in the mean power across high-frequency channels from the cochleogram (center frequency$>$ 1760 Hz) using 16 ms analysis frames. The model assumes a Poisson distribution in the input. Note the change in rhythm in the input sequence is reflected in the beliefs, and higher surprisal indicates moments when the rate of transients deviates from the preceding statistics.

These examples illustrate the flexibility of the model to build predictions from a variety of auditory inputs along various dimensions. Importantly, we do not prescribe a particular set of statistics in the model. Rather, the flexibility to utilize different statistics offers an opportunity to compare various statistical representations to see which best explains experimental results.

2.4 Replication of results from the literature

To demonstrate the model's applicability to existing experimental results in predictive coding, we collected surprisal responses from the D-REX model to stimuli found in the literature. Stimuli range in predictability to show the capacity of the model to capture a variety of phenomena under a single framework. Using a Gaussian distribution with different levels of temporal dependence (D), we can ascertain the statistics that are sufficient—i.e., the "simplest explanation"—for responses observed in the brain.

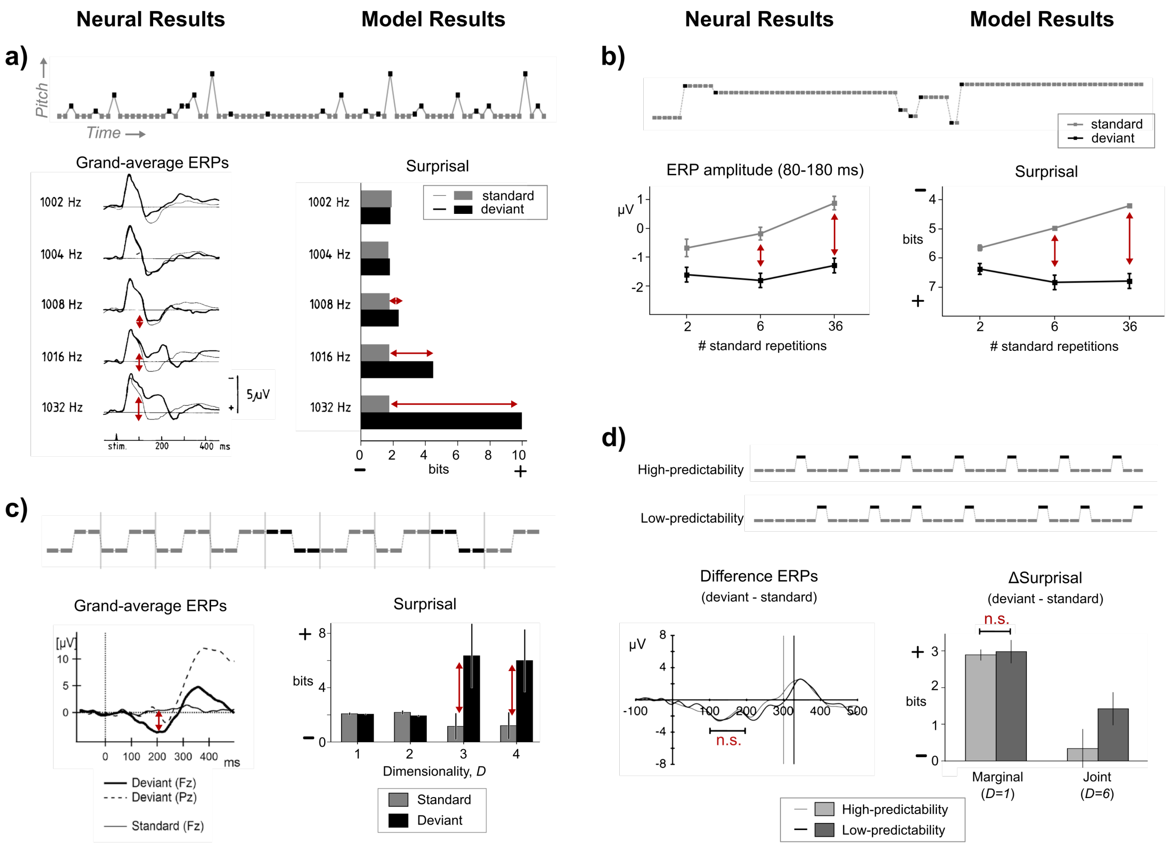

Figure 2-4. Replication of neural results. Results from the literature (left) are compared to surprisal responses from the D-REX model (right) to the same stimuli (above): a) [6], b) [55], c) [56], d) [57]. Arrows indicate replicated trends. Surprisal axis is occasionally inverted to facilitate visual comparison. Experimental figures reproduced with permission from the publishers. Data in b) plotted from published table.

In Figs 2-4 and 2-5, neural results directly from the literature are presented alongside model results for comparison (e.g., MMN amplitude vs. surprisal), with example stimuli shown above each result. Trends shared between neural and model results are indicated by red arrows. To facilitate visual comparison, the surprisal axis is occasionally inverted to align higher surprisal in the model results with lower predictability in the neural results. Figures from the literature are reproduced in their original form, unless otherwise noted.

Oddball. Dating back to 1978, Näätänen and colleagues have used the oddball paradigm to elicit neural markers of deviance from a detected regularity [7, 58]. The paradigm includes a *standard* stimulus exhibiting some regularity and *deviant* stimuli breaking the regularity; if the brain is sensitive to the regularity, the mismatch negativity (MMN) appears around 100–200 ms after onset in the deviant's Event-Related Potential (ERP) response relative to the standard. This negativity increases with frequency distance between the deviant and standard [6]. The D-REX model with $D = 1$, or marginal statistics, similarly shows an increase in surprisal to the deviant as frequency distance increases (see Fig 2-4a).

Roving oddball. The oddball paradigm has been extended using a standard that changes over time, where each deviant becomes the new standard. As the number of standards increases, ERP response to the *standard* increases in the MMN window (80–180 ms), while response to the *deviant* stay relatively the same [55]; similarly, as the number of standards increases, model surprisal with $D = 1$ decreases ($F_{2,147} = 108.1, p < 0.0001$), while surprisal to deviants stays the same ($F_{2,147} = 1.18, p > 0.1$) (see Fig 2-4b[3], surprisal axis flipped for visual comparison).

Pattern oddball. Tone-patterns can also serve as standards in the oddball paradigm. In [56], an MMN response to the first tone of the deviant pattern (BBAA) relative to the first tone of the standard pattern (AABB) indicates the brain is sensitive to the 4-tone pattern. In the model's surprisal response, this is replicated with dimensionality $D > 2$ ($t_{74} = 15.11, p < 0.0001$), indicating the minimal statistics necessary to detect the deviant is actually over a shorter window than the pattern itself; deviance can be detected by the entire 4-tone pattern or by three repetitions of the same tone (see Fig 2-4c).

[3]Neural results from literature reproduced from data published in a table.

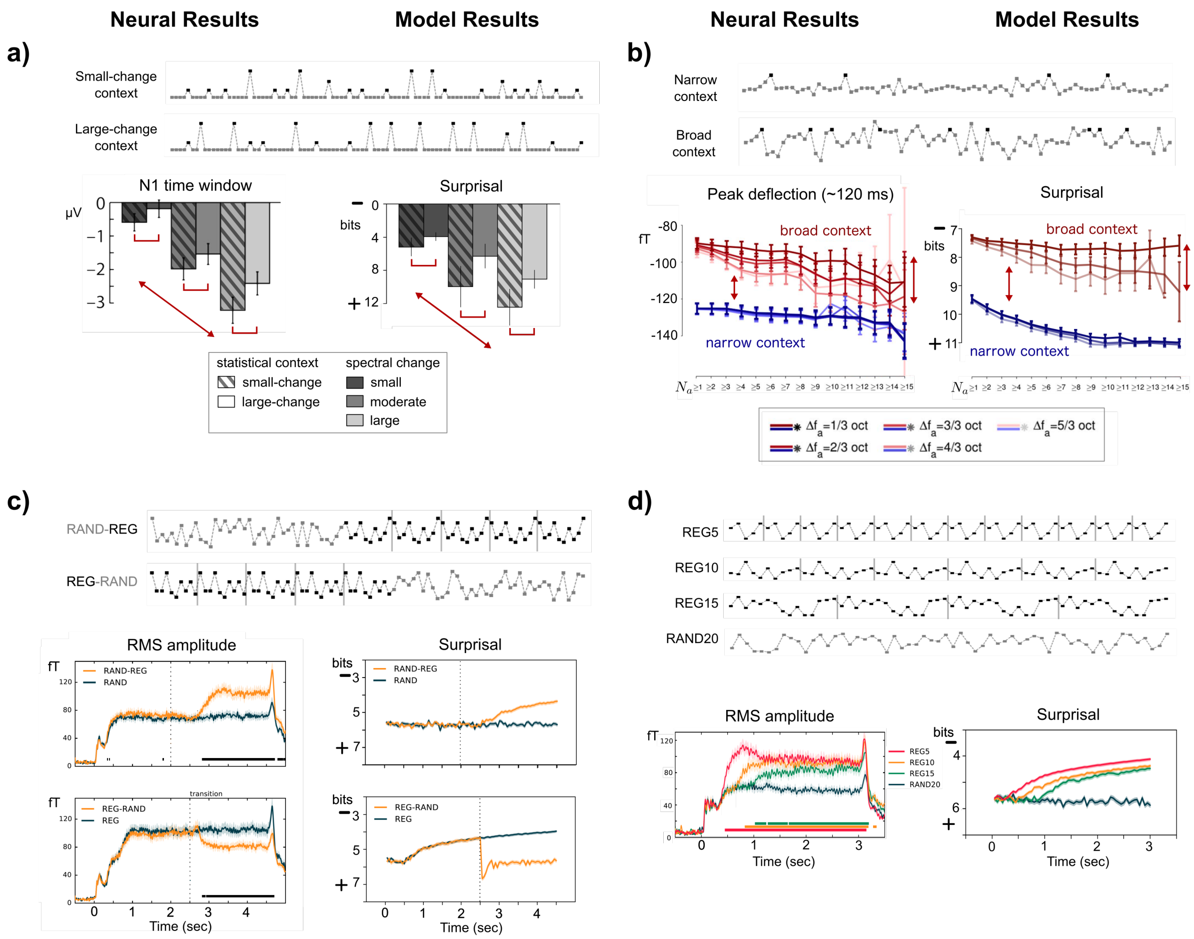

Figure 2-5. Replication of neural results from the literature, continued. a) [15], b) [14], c) and d) [35]. Arrows indicate replicated effects. Surprisal axis is occasionally inverted to facilitate visual comparison. Experimental figures reproduced with permission from the publishers. Data in a) plotted from published table.

High- & low-predictability oddball. Top-down attentional affects have been measured in the MMN response. In [57], the MMN response was measured in two conditions: a high-predictability condition where the number of standards preceding a deviant was usually 4 (AAAAB), and a low-predictability condition where the number of standards was uniformly distributed between 2 and 6. Listeners were tasked with detecting every deviant (B). ERP evidence shows a significant MMN response to deviants but *no difference* in MMN magnitude between predictability conditions; this null result is replicated by differential surprisal between deviant and standard from the model with $D = 1$ collecting only marginal statistics ($t_{23} = 1.27, p > 0.1$) (see Fig 2-4d).

By contrast, a model with $D = 6$ collects temporal covariances that cover the entire AAAAB pattern and no longer finds the final B tone "surprising" (see Fig 2-4d-right). This mirrors a similar study where listeners were tasked with listening for the entire pattern and exhibited no MMN response to the deviant tone [59]. These top-down effects can be described in terms of the statistics being collected—when attending to the B tone only, listeners collect marginal statistics; when attending to the entire AAAAB pattern, listeners collect long-range temporal statistics.

Statistical oddball biased toward large or small changes. Context effects have been observed in the MMN response by manipulating the relative probabilities of deviants, biasing them toward small- or large-change deviants [15]. Deviant effects moduled by statistical context are observed in N1 amplitude: magnitude increases with deviant change and is augmented by the small-change context, where large changes are less probable. An ANOVA applied to model surprisal (with $D = 1$) shows the same significant effects for spectral change ($F_{2,477} = 668.66, p < 0.0001$) and statistical context ($F_{1,477} = 221.14, p < 0.0001$) (see Fig 2-5a[4]).

Gaussian sequences differing in variance. Context effects have also been observed using random stimuli drawn from a Gaussian distribution with different variances [14]. Responses to deviants (presented 2 octaves above the mean) show a negative peak around 120 ms that is larger for narrow relative to broad statistical context. Additionally, there is evidence of adaptation effects in the broad context when comparing deviant responses based on the number of preceding tones (N_a) falling outside a frequency region (ΔF_a) (see [14] for details). The model with $D = 1$ replicates these results (see Fig 2-5b).

Regular vs. random sequences. Repeating patterns are another class of stimuli used to explore regularity extraction in the brain. In particular, RMS power in MEG has been shown to increase with decreasing entropy in the stimulus [35]: RMS power *increases gradually* when the stimulus transitions from random to repeating pattern (RAND-REG), while RMS power *decreases abruptly* for the opposite transition (REG-RAND). The model replicates both of these phenomena in the time-course of surprisal, with D greater than the pattern length (see Fig 2-5c). Additionally, the model replicates effects of pattern length on RMS power [35], again reflecting differences in entropy (see Fig 2-5d).

2.5 Discussion

The D-REX model is a functional instantiation of existing theoretical formulations for predictive processing and object formation in perception, where sound sources are represented probabilistically and sensory inputs are incorporated into the brain's internal representation of the world [5, 22, 24, 60, 61]. The composition of the D-REX

[4] Neural results from literature reproduced from data published in a table.

model aligns with previous literature regarding the underlying computations behind predictive processing: the brain builds statistical representations estimated from sounds over time [20, 21, 62, 63], and the brain maintains multiple hypotheses for how much of the past is relevant to the present moment [64, 65]. These claims are represented explicitly in the model by statistical estimates collected over different time-windows, each of which gives a prediction for future inputs. Prediction errors are then used to update probabilistic beliefs in each context, weighting contexts proportionally by their evidence. This competition between concurrent hypotheses for the relevant context is crucial for robust interpretation with dynamics and uncertainty in the sensory input.

By no means a complete picture of predictive coding in auditory perception, the D-REX model is a flexible computational framework offering several footholds from which facets of predictive processing can be explored. By connecting the model's outputs to experimental responses, the model can act as a "simulated" listener undergoing the same experimental tasks as human listeners. The internal components of the model can then be tinkered with and tuned to explore which configurations of the model give rise to responses that match listener responses. This approach can be used to investigate many open questions in predictive processing in audition.

The model can be used to investigate the nature of the internal statistical representation employed by the brain. What statistics are collected by the brain? How do these statistics differ between perceptual dimensions? To what extent are dependencies over time and across dimensions represented? How do statistical representations vary with listeners' attentional state or long-term experience? These questions can be addressed explicitly using the statistical estimates employed in the model: with existing experimental results, the model can be used with different statistical representations to examine which best replicates listener responses, or the model can be used to drive new investigations specifically designed to tease apart the statistical representation by providing alternative hypotheses for experimental results under certain statistics.

The model can also be used to investigate how context and experience shape perception at different time-scales. At short-term scales, the context windows of the model can be used to ask questions about the granularity of the statistical representation in memory, for example, to set an upper bound on the maximum context window used by listeners, or to find the minimum set of contexts that can replicate listener behavior and whether this is consistent across stimuli with different levels of complexity. At longer time-scales, the priors of the model can be used to represent different prior expectations of the listener learned from previous exposure, where model responses using different priors could be used to investigate how prior experience affects predictions or how listener responses reflect learning over the course of an experiment. Again, these questions can be approached by using the model to give targeted hypotheses for experimental outcomes.

As a surrogate for the computational processes behind predictive processing in individual listeners, the model can be used to explain differences in behavioral or neural responses across listeners. In addition to examining effects of representation and experience on individual perception mentioned above, the perceptual parameters of the model (memory and observation noise) can provide additional insight into how known constraints on neural resources manifest in subject-to-subject variability in behavioral and neural responses. Currently, the connection between these modeling parameters and their neural counterparts is plausible, and early evidence supports this connection (see Appendix II for preliminary results exploring the model's predictions of working memory capacity). Future investigation into the behavioral and neural consequences of the perceptual parameters can add interpretive heft to the model.

An additional strength of the model lies in its ability to combat the noise that invariably creeps into experimental paradigms incorporating uncertainty. Behavioral and neural responses to stochastic stimuli are themselves stochastic, and trial-to-trial variability can cloud results, especially in neural responses where precise time-locking is often a prerequisite to any event-related analysis. The model can be used to reduce jitter by aligning neural responses to events derived from model response *to the same stimulus*. Neural responses can then be correlated with specific aspects of predictive processing (e.g., prediction error, precision, evidence accumulation). The model provides an avenue to take findings established in more tightly-controlled experiments, and see if they hold in more complex settings where well-defined events for time-locking are less apparent.

Finally, the model is modular and extendable. We demonstrated the capacity of the model to capture many possible statistical representations along different sensory dimensions in real-world audio examples, but the input dimensions and probability distributions explored here are not exhaustive. New probability distributions can easily be included in the D-REX model, and the model can be applied along any dimension in the acoustic input. Moreover, the modeling framework can be expanded in other ways to broaden its application. As currently implemented, the model operates at a single level in the sensory input and along a single time-scale, but it could be layered to build heirarchical predictions at different levels of abstraction or multiple time-scales. In addition, while the model was designed for audition, the same sequential prediction computations could be applied in and across other sensory modalities. Future work can also address how the predictive algorithms identified by the model could be implemented in neural circuits.

Beyond retrospective interpretation of existing results, the D-REX model can be used to guide future experiments, probing the temporal processing of complex sounds. As a flexible and general computational model for predictive coding, it can be used as a tool to pursue a deeper understanding of the computational mechanisms behind predictive coding of rich, dynamic sounds in a variety of listening scenarios under a single unifying framework. The D-REX model can be used to push the boundary of what is considered feasible for study in the laboratory towards the complexity encountered in everyday listening.

Chapter 3

Statistical inference along a single dimension

3.1 Introduction

In this chapter, we employ the Dynamic Regularity Extraction (D-REX) model described in Section 2 to model Bayesian inference used by the auditory system to track sensory statistics in pitch. This computational framework, alongside human behavioral and electroencephalography (EEG) experiments, allows us to directly test alternative hypotheses regarding the extent to which auditory statistical information is represented in memory and the optimality of statistical inference in the brain.

The nature of the statistical representation collected by the brain has not been fully explored in the literature. Previous studies have focused on the marginal statistics of tones within a sequence, showing that the brain is sensitive to changes in mean and variance [14, 16]. We refer to these as *lower-order statistics*, describing sounds independent of their context. Here, we investigate whether the brain collects *higher-*

order statistics about the dependencies between sounds over time; namely, we examine how the brain gathers information about the temporal covariance structure in a stochastic sequence of sounds. We use melody stimuli with pitches based on random fractals, which exhibit long-range dependencies and cannot be described solely by lower-order statistics. We specifically use random fractals because of their ecological relevance: previous work has demonstrated the presence of random fractals in music [25], speech [26], and natural sounds [13] and shown the brain is sensitive to the amount of randomness, or entropy, in random fractal melodies [11, 12].

Change detection experiments are well-suited for investigating regularity extraction, where the task is to detect deviation from an established regularity in a sequence of sounds. A detection can be reported behaviorally or recorded in the neural response; for example, in EEG studies the Mismatch Negativity (MMN) is commonly used to index deviance detection in the brain. A correct detection indicates the brain is sensitive to the tested regularity, for a change response is necessarily preceded by knowledge of what is being changed. Compared to discrimination, the change detection paradigm more closely mirrors how the brain processes sounds in the real world, where boundaries between sound sources are not known *a priori*, but must be inferred from changes in ongoing sound.

The mechanisms needed for change detection may differ depending on the type of regularity. With deterministic regularities, the brain can explicitly test whether each incoming sound deviates from the extracted pattern or not with near certainty. Deviation from a stochastic regularity, on the other hand, emerges gradually as evidence is accumulated over time, causing a delay in the perceived moment of change proportional to the amount of evidence needed to detect the change. This uncertainty unavoidably introduces variability in perception across trials and across subjects, which is particularly problematic for time-locked analyses such as in EEG, where low SNR necessitates many repetitions and precise temporal alignment across trials and subjects to get meaningful results. To account for this variability and facilitate the study of stochastic regularities in change detection, we use the D-REX model as a perceptual model of the mechanisms for extracting and using regularities in a changing scene to guide our analysis.

The perceptual parameters of the D-REX model that represent neural resource limitations (i.e., finite working memory and observation noise) provide constraints on performance that are valuable to interpret sub-optimal detection performance and variability across listeners' behavior. By fitting the model to human behavior from a series of change detection experiments, we explore questions regarding auditory stochastic regularity extraction: Which statistics are sufficient to explain human behavior? How do the perceptual parameters of the model account for differences in behavior across subjects? Finally, we use the model to guide analysis of EEG data, revealing effects that would be otherwise hidden using conventional EEG analyses.

3.2 Methods

3.2.1 Participants

All participants reported no history of hearing loss or neurological problems. Participants gave informed consent prior to the experiment and were paid for their participation. All procedures were approved by the Johns Hopkins Institutional Review Board (IRB).

In Experiment 1, ten participants (9 Female) were recruited from an undergraduate population (mean age: 18.7 years). In Experiment 1b, 21 participants (14 Female) were recruited from an undergraduate population (mean age: 20.1 years). In Experiment 2, ten participants (6 Female) were recruited from an undergraduate population (mean age: 18.7 years). Finally, in Experiment 3 (EEG), 14 participants were recruited, and six participants were excluded from EEG analysis because behavioral performance was near chance ($d' < 0.5$). Out of the remaining eight subjects, six were female, and the mean age was 20 years.

3.2.2 Stimuli

Stimuli in Experiments 1–2 were pure-tone melodies with tone frequencies determined by random fractals. Random fractals are stochastic processes with spectrum inversely proportional to frequency and with spectral slope β ($1/f^{\beta}$). β parameterizes the entropy of the random fractal: as β decreases entropy increases, with $\beta = 0$ yielding a white-noise spectrum and the highest entropy. Four levels of entropy were used to create the stimuli, corresponding to $\beta = 0, 1.5, 2, 2.5$. Random fractals were generated by repeatedly applying the inverse Fourier transform to the $1/f^{\beta}$ spectrum with random phase, yielding many unique instances. These random fractals were standardized to remove any differences in mean and variance, then quantized and mapped to 35 frequencies in a quasi-semitone scale (15 frequencies/octave) centered on 330 Hz (range: 150–724 Hz). Melodies were synthesized using pure tones with 150ms duration and 10ms ramped onset and offset (squared cosine). Inter-onset interval between tones was 175ms.

In Experiments 1 and 1b, all melody stimuli had a length of 60 tones. Stimuli with changes in entropy ("change trials") were composed of two equal-length melodies with different entropy, one with the highest entropy ($\beta = 0$) and one with a lower entropy, resulting in three degrees of change ($\Delta\beta = 1.5, 2, 2.5$). Both increasing- and decreasing-entropy trials (referred to as INCR and DECR, respectively) were included, resulting in six change conditions, as well as control trials with constant entropy at each entropy levels. There were 150 trials in total, with 15 trials per condition.

In Experiment 2, stimuli were similar to those in Experiment 1 with an additional manipulation of melody length. Along with the same change degree and direction conditions, there were three length conditions (20, 40, and 60 tones) with the change always occurring in the midpoint of the melody. For each of the 18 change conditions (3 $\Delta\beta$ x 2 direction x 3 length) and each of the 12 control conditions (4 β x 3 length), there were 8 trials, for a total of 240 trials.

In Experiment 3, stimuli were based on an alternative parameterization of entropy using first-order Markov chains, which provided greater control over the distributions used to generate the melodies. Specifically, this allowed us to exclude tone repetitions from the melody stimuli to prevent any correlates in EEG due simply to repetition. Because none of the analyses or results are predicated on properties exclusive to random fractals, and both types of stochastic stimuli are perceptually similar, we treat both stimuli identically.

Melody stimuli were composed of 50 pure-tones with pitches sampled from 11 frequencies on a semitone scale (range: 247–440 Hz). For each melody, the first tone frequency was sampled uniformly from all 11 frequencies. Subsequent tone frequencies were drawn from a probability distribution based on a modified logistic curve centered on the previous observation with entropy parameterized by the logistic slope k,

$$P_k(x_t|x_{t-1}) = \begin{cases} 0, & x_t = x_{t-1} \\ A/(1+e^{-k|x_t - x_{t-1}|}), & \text{otherwise} \end{cases}$$

where x_t and x_{t-1} are the current and former tone frequencies (in semitones) and A is a normalization constant. As k increases, this distribution becomes more biased towards smaller frequency steps and lower entropy, and it has maximum entropy at $k = 0$, a uniform distribution across the 10 frequencies (excluding the previous frequency). High-entropy sequences and low-entropy sequences were generated with $k = 0$ and $k = 0.7$, respectively. For change trials, k transitioned smoothly between the two extremes in the middle 10 tones of the melody (tones 21–30) to avoid obvious outliers from an abrupt change in the distribution.

In Experiment 3, there were 150 melody trials in this experiment: 50 trials for each change direction (INCR and DECR), and 25 control trials per entropy level (LOW and HIGH). Tones were 125 ms in duration and presented with inter-onset interval of 160 ms.

3.2.3 Procedure

For all experiments, stimuli were presented in randomized order by subject with self-paced breaks between blocks. During each melody trial, listeners were instructed to listen for a change in the melody. Feedback was given after each response in order to guard against task misunderstanding and ensure listeners had as much information as possible to perform the task well.

Listeners were not given explicit instructions about what they were listening for, but rather learned the task implicitly over the course of a training block prior to testing. Incorrect responses in the training block caused the same stimulus to be replayed with feedback (including an indication of when the change occurs, in the case of missed detections). Participants advanced to testing after completing at least 15 trials and correctly answering 5 consecutive trials (all participants completed training in under 30 trials).

In Experiments 1, 2, and 3, participants responded via keyboard (or response box for Experiment 3) whether or not they heard a change *after the melody finished*. In Experiment 1b, listeners responded *in the middle of the melody trial* as soon as a change was heard by pressing the space-bar. If the space-bar was not pressed before the end of the melody presentation, this was recorded as a negative response. Responses before the nominal changepoint of change trials (i.e., the midpoint) were considered false-alarms.

In psychophysics experiments (1, 1b, 2), Stimuli were synthesized offline as 16-bit, 44.1 kHz wav-files and presented via over-ear headphones (Sennheiser HD 595) at a comfortable listening level using PsychToolbox (`psychtoolbox.org`) and custom scripts in MATLAB (The Mathworks). Participants were seated in an anechoic booth in front of the presentation computer. The experiment duration was approximately 50 minutes.

In the Experiment 3, subjects were seated in an anechoic chamber with stimuli presented via in-ear earphones (Etymotic ER-2) at a comfortable listening level. Before each melody trial, a cross appeared in the center of the screen, and subjects were instructed to fixate on the cross to reduce eye movement artifacts.

3.2.4 EEG recording and data analysis

In Experiment 3, EEG was recorded using a BioSemi ActiveTwo system (Biosemi) with 32 electrodes placed in central and frontal locations on the scalp selected to maximize signal-to-noise ratio for neural signals originating in auditory centers of the brain [66, 67]. Six additional electrodes were placed on left and right mastoids, the nose, and alongside the eyes for re-referencing and blink artifact removal. Data was recorded at a sampling rate of 4096 Hz.

For each subject, EEG data were preprocessed with custom scripts in MATLAB using the FieldTrip toolbox (www.fieldtriptoolbox.org) and NoiseTools [68]. Continuous EEG was re-referenced to the left mastoid, filtered to 1–100 Hz (two-pass Butterworth, 3rd-order for high-pass and 6th-order for low-pass), and re-sampled to 256 Hz. The data was then cleaned in two stages using Independent Component Analysis (ICA) and Denoising Source Separation (DSS). First, continuous EEG data was epoched to 1 second segments; segments with amplitude range exceeding 3 s.d. from the mean by channel were excluded before applying ICA to identify components attributable to eye motion artifacts. These artifact components were removed from the continuous EEG data, and the ICA-cleaned data was epoched to melody trials. DSS was then used to enhance stimulus-locked activity; the top 5 DSS components that were most repeatable across melody trials were kept and projected back to sensor space, thus removing EEG signal not related to auditory stimulation [68].

We used regression to investigate effects of model surprisal on ERP responses based on the framework described in [69, 70]. For each subject, EEG data was further low-pass filtered at 30Hz (6th-order Butterworth) and epoched by tone with the 50-ms window preceding tone onset used for baseline subtraction. Outlier tone trials with amplitude exceeding 3 s.d. from the mean were excluded from the analysis.

We fit the following regression model to single-trial ERPs:

$$y_i(t) = \beta_0(t) + S_L\beta_L(t) + S_H\beta_H(t) + \epsilon_i(t)$$

where surprisal from the LOS model (S_L) and the HOS model (S_H) serve as predictors in the regression for the i^{th} single-trial ERP (y_i). The regression contains an intercept term β_0, which captures the baseline ERP response, and slope terms β_L and β_H, which capture the differential response due to a unit change in S_L and S_H, respectively. Finally, ϵ_i is the residual error for the i-th trial. Note that these terms are indexed by time, so the regression finds the linear relationship between regressors (S_L and S_H) and the single-trial ERPs at each time point, yielding a regression-ERP, or rERP [69]. The regression was applied separately for each subject to EEG data averaged across all 32 electrodes.

We used phase-locking value (PLV) to measure neural phase-locking to tones. PLV is a measure of phase agreement across trials independent of signal power:

$$PLV = \frac{1}{n}\left|\sum_{i=1}^{n} \phi_i/|\phi_i|\right|$$

where the ϕ_i's are complex phasors extracted from the Fourier transform at the frequency of interest (6.25Hz, the tone presentation rate) for the i^{th} trial, and n is the number of trials. PLV was calculated separately for 1120ms (7-tone) epochs before and after the changepoints, and the difference, $\Delta PLV = PLV_{after} - PLV_{before}$, was used to measure the change in phase-locking at the changepoints. Only change trials correctly detected by both listener and model were included in this analysis.

For statistical testing, ΔPLV was compared to 0 (t-test) and to a null distribution (random permutation test) estimated by calculating ΔPLV from randomly sampled changepoints across the melody. The null distribution ensures any observed change in PLV at the changepoints is not simply due to the random variability in phase-locking present across the melody trial.

3.2.5 Model

We use the D-REX model described in Chapter 2 to interpret behavioral and neural data in Experiments 1–3. To collect responses from the model that are comparable to those collected from human listeners, we derived a *change probability*—the probability a change has occurred—from the context beliefs, $\vec{B}_t$, which form the posterior probability over context hypotheses given all observed observations: $\mathbb{P}(c_i|x_{1:t})$. The probability that a change has *not* occurred before time t is equal to the belief that the current context is equal to the length of the entire observed sequence (i.e., $P(c_i = t|x_{1:t})$); the probability that *at least* one change has occurred is then the converse of this, or the sum of beliefs in contexts less than the length of the observed sequence:

$$P(\text{Change}|x_{1:t}) = 1 - P(c_i = t|x_{1:t}) = \sum_{c'<t} P(c_i = c'|x_{1:t})$$

This probability of a change grows over time, representing the accumulation of evidence of a change. We then apply a simple decision rule to get a binary change detection response from the model. At the end of the melody (i.e., post-trial), the model makes a *change decision* by comparing the final change probability to a decision threshold:

$$\text{Change decision} = \begin{cases} \text{Yes,} & \mathbb{P}(\text{Change}|x_{1:T}) \geq \tau \\ \text{No,} & \mathbb{P}(\text{Change}|x_{1:T}) < \tau \end{cases}$$

where T is the full melody length and the threshold τ is an additional parameter of the model. We then define the model *changepoint* as the earliest time at which the change probability exceeds this threshold:

$$\text{Model changepoint} = \arg\min_t \{\mathbb{P}(\text{Change}|x_{1:t}) \geq \tau\}$$

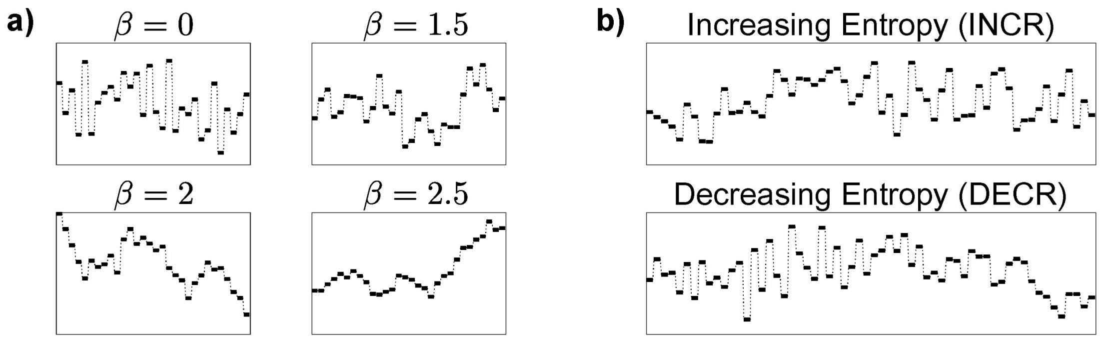

Figure 3-1. Random fractal stimuli. Schematic spectrograms shown with frequency and time along the vertical and horizontal axes, respectively. a) Melodies at four levels of entropy, parameterized by β. Higher β corresponds with lower entropy, and vice versa. b) Change stimuli for each change direction; INCR and DECR stimuli always end and begin, respectively, with the highest level of entropy ($\beta = 0$ or white noise).

3.3 Results

3.3.1 Perceptual experiments

A series of experiments probed listener's ability to detect changes in fractal melodies. Stimuli were constructed from melodies at four levels of *randomness* or *entropy* in pitch (both terms used interchangeably). Melody entropy is parameterized by β, where $\beta = 0$ corresponds to the highest entropy (white noise), and entropy decreases as β increases (see Fig 3-1a for examples of fractal melodies at different levels of β). Lower-order statistics (mean and variance) were normalized across the melody. Half-way through the melody, only the higher-order statistics change (see Fig 3-1b for examples of change stimuli). The task in all experiments was the same: detect a change in entropy of the melody.

Experiment 1

We tested how well listeners could detect changes in the entropy of tone sequences and whether the direction of change affected detection performance; see Fig 3-1b for example stimuli. Listeners ($N = 10$) heard stimuli with three degrees of change in entropy (between $\beta = 0$ and $\beta = 1.5, 2, 2.5$) in both directions (INCR and DECR), with control stimuli containing no change (with $\beta = 0, 1.5, 2, 2.5$). Each melody trial contained 60 tones presented isochronously over 10.5 seconds (175 ms inter-onset interval); there were 150 trials in total, with 15 trials per condition. After each melody trial, listeners responded whether they heard a change and received immediate feedback.

Detection performance as measured by d' is shown in Fig 3-2a; d' comprises both hits and false-alarms (FAs), with higher d' corresponding to better detection performance and $d' = 0$ corresponding to chance performance. Repeated-measures ANOVAs were used in all analyses to account for between-subject variability. An ANOVA with 2 within-subjects factors (3 change degree x 2 direction) showed a strong effect of degree ($F(2, 18) = 31.5$, $p < 0.0001$), no significant effect of direction, and a significant interaction ($F(2, 18) = 9.4$, $p < 0.01$). We investigated this interaction further by applying ANOVAs separately to hit- and FA-rates. The hit-ANOVA showed a strong effect of degree ($F(2, 18) = 21.9$, $p < 0.0001$) but no effect of direction *or* interaction, while the FA-ANOVA showed an effect of entropy level ($F(3, 27) = 4.7$, $p < 0.01$), with FAs increasing with entropy (Note the increase in degrees-of-freedom is due to the 4 levels of β for control stimuli). The significant interaction between degree and direction seen in d' above is therefore only due to the effect of entropy on FAs: all DECR stimuli begin with the same high level of entropy ($\beta = 0$), thus increasing FAs and decreasing d' for DECR compared to INCR stimuli.

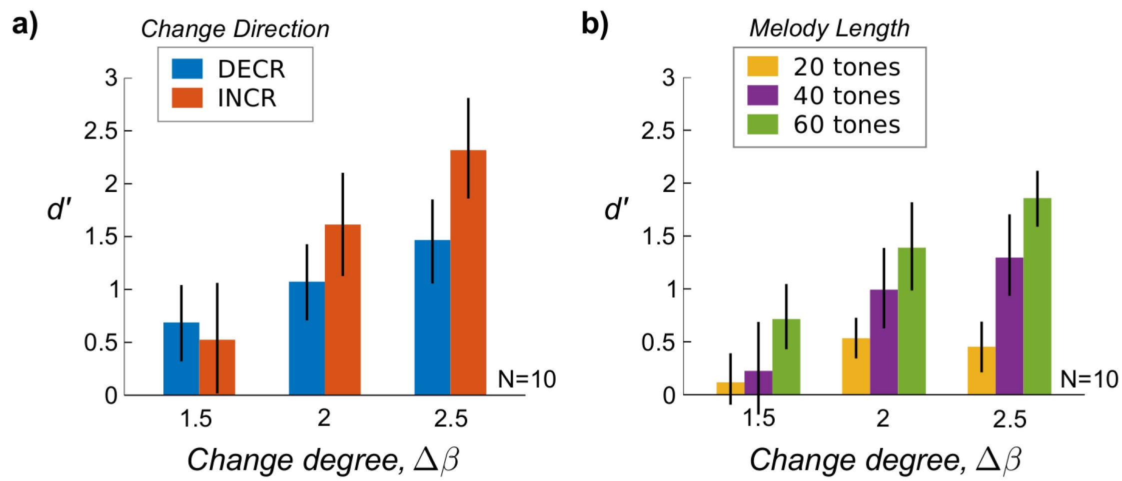

Figure 3-2. Psychophysics results from Experiments 1 and 2. Average change detection performance (d') across subjects is shown by stimulus condition. Error bars indicate 95% bootstrap confidence interval across subjects. a) In Experiment 1 ($N = 10$), melody entropy changed with different degrees ($\Delta\beta$, abscissa) and in both INCR and DECR direction (color). Detection performance increased with $\Delta\beta$ but did not differ by direction, although there was a weak interaction between $\Delta\beta$ and direction due to FAs only. b) In Experiment 2 ($N = 10$), an additional factor of melody length was introduced (color). Detection performance increased with both $\Delta\beta$ and melody length.

It is surprising that there is no effect of change direction on hit-rates. If listeners are relying solely on lower-order statistics, INCR changes should be easier to detect than DECR changes by listening for outliers. We look closely at this effect in a follow-up experiment (Experiment 1b) to contrast response time (RT) to INCR versus DECR changes.

Experiment 1b

In this experiment, listeners ($N = 21$) responded *as soon as* they heard a change during melody presentation; otherwise, the stimuli and procedure were the same as in Experiment 1. To confirm that the difference in task itself had no effect on detection performance, two-sample t-tests of d' for each condition showed no difference across the two experiments ($p > 0.05$ for all tests, using Bonferroni correction for multiple comparisons). In addition, ANOVAs applied to hit- and FA-rates as in Experiment 1 showed the same significant effects.

A repeated-measures ANOVA applied to the RT data averaged within conditions for change-trials (3 change degree x 2 direction) showed a significant main effect of change degree ($F(2, 40) = 14.3$, $p < 0.0001$) but no main effect of direction and no significant interaction, confirming the result from Experiment 1 with no effect of change direction on detection performance.

Experiment 2

Next, we tested the effect of sequence length on change detection performance. In addition to the same change degree and direction manipulations from Experiment 1, listeners ($N = 10$) heard melodies with different lengths (20, 40, and 60 tones), with the change always occurring at the midpoint of the melody. As there was no effect of change direction on performance seen in Experiments 1 and 1b, we pooled results across INCR and DECR trials. As in Experiment 1, listeners responded whether they heard a change after the melody presentation and received immediate feedback.

Detection performance as measured by d' is shown in Fig 3-2b. A repeated-measures ANOVA with 2 factors (3 change degree and 3 melody length) showed significant main effects of both change degree ($F(2, 18) = 23.9$, $p < 0.0001$) and melody length ($F(2, 18) = 17.7$, $p < 0.0001$), with a weak interaction ($F(4, 36) = 2.8$, $p < 0.05$). Post-hoc tests indicated the weak interaction was due to chance performance in the most difficult conditions: $\Delta\beta = 1.5$ with lengths of 20 and 40 tones. In separate ANOVAs for hit- and FA-rates, hit-rates showed both main effects of change degree ($F(2, 18) = 10.2$, $p < 0.01$) and length ($F(2, 18) = 29.6$, $p < 0.0001$) with no significant interaction, while the FA-rates only showed a significant effect of entropy level ($F(2, 18) = 14.6$, $p < 0.001$) and no effect of length or interaction.

3.3.2 Computational Model

In this application of the D-REX model, the generating distribution is assumed to be a D-dimensional multivariate Gaussian with unknown mean and covariance structure, where the dimensionality D specifies the amount of temporal dependence in the model. As new observations come in, the model incrementally collects sufficient statistics whose form depends on D. Here, we ask whether human behavior from Experiments 1–2 can be captured by a model that collects marginal lower-order statistics (D=1, i.e., mean and variance) or if higher-order statistics (D=2, i.e., mean, variance, and covariance) are needed; we refer to these two versions of the model as the *LOS model* and *HOS model*, respectively.

Perceptual parameters and model behavior

We first examined the model detection performance for different sets of model parameters: memory (m), observation noise (n), change-prior (π), and threshold (τ). Using a parameter sweep, we collected model change decision responses to the same stimuli used in Experiments 1–2 and measured model performance for each operating point in the sweep.

Fig 3-3 shows model performance for Experiment 1. Performance is displayed in Receiver Operating Characteristic space (ROC-space); ROC-space is a method for visualizing the trade-off between Hit- and FA-rates in system performance at multiple operating points (i.e., parameter sets); the upper-left corner is perfect performance (Hit=1, FA=0), and the diagonal is chance performance (Hit=FA). Fig 3-3a displays the coverage of model performance in ROC-space for the LOS and HOS model (in blue and red, respectively); for example, at every red-colored coordinate in ROC-space, there is a set of parameters $\{m, n, \tau, \pi\}$ in the HOS model with that performance (i.e.,

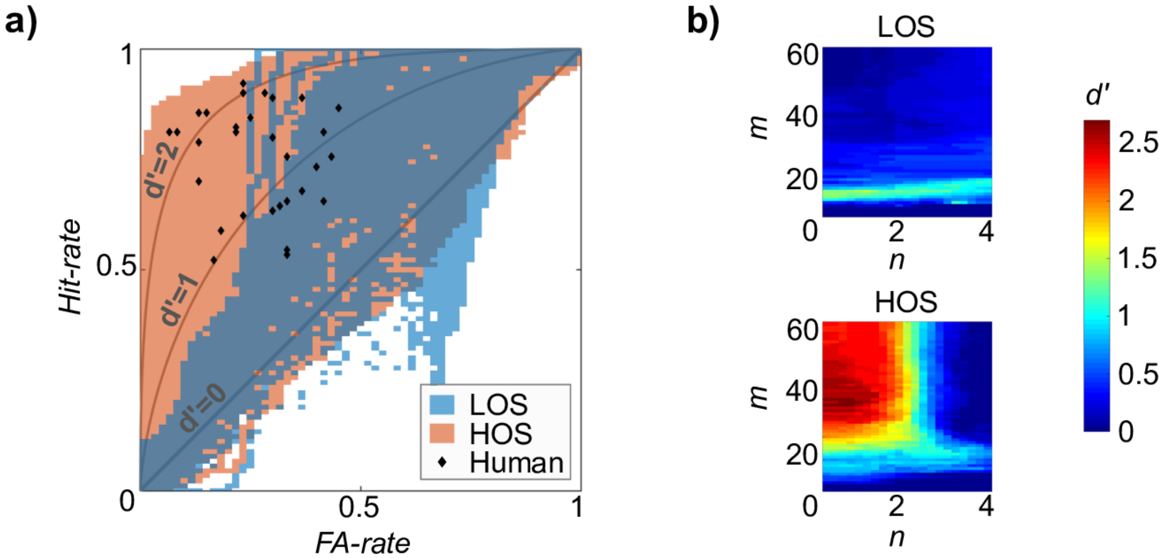

Figure 3-3. Range of model behavior in Experiment 1. Model detection performance measured at different operating points in a parameter sweep. a) Comparison of detection performance for LOS and HOS models displayed in ROC-space across the parameter sweep, with model type denoted by color. Each blue (red) coordinate indicates existence of a parameter set for the LOS (HOS) model yielding that performance. Individual human performance from Experiments 1 and 1b is overlaid, along with equal-d' curves. b) d' surface as a function of memory (m) and observation noise (n) parameters for LOS model (top) and HOS model (bottom). π and τ were held constant at 0.01 and 0.5, respectively.

Hit- and FA-rate). In this manner, we can compare the *range* of performance between the two models across the entire parameter sweep. Individual human performance from Experiments 1 and 1b (with the same stimuli, $N = 31$) and equal-d' curves are overlaid in the same space for comparison. Results from Experiment 2 were similar.

There is a clear contrast in the range of performance in ROC-space between LOS and HOS models, with the HOS model having both wider coverage and higher ceiling performance overall compared to the LOS model. While the LOS model only overlaps with poorer performing subjects ($d' < 1.5$), the HOS model overlaps with all human performance points. Additionally, human performance never exceeds the range of the HOS model, indicating that with unconstrained resources (i.e., infinite memory and zero observation noise) the HOS model can act as an "ideal observer", providing an upper bound for human performance.

Fig 3-3b shows the d' surface for the LOS model (top) and HOS model (bottom) as a function of the two perceptual parameters, allowing us to assess which parameters are responsible for the performance variability seen in Fig 3-3a for each model. With the LOS model, the memory m is largely responsible for performance variability, with only a narrow band around $m = 10$ where the LOS model performs well above chance ($d' = 0$). The HOS model performance, on the other hand, varies jointly with both memory m and observation noise n, with the best performance around $\{n = 0, m = 30\}$.

Fitting the model to subject behavior

We fit the model parameters to each subject from Experiments 1–2. There was very high between-subject variability in performance (e.g., see human performance plotted in ROC-space in Fig 3-3a), so we examined how the parameters from the fitted model explain this variance. Model performance was measured for each set of parameters in the parameter sweep, and the best set of parameters was selected for each subject using minimum Euclidean distance between model and subject performance. Performance was measured using hit- and FA-rate within each change direction, which provided a more stringent criterion for distinguishing between parameters with equal overall hit- and FA-rates.

Fig 3-4 shows results from fitting the model to subjects from Experiments 1–2 ($N = 41$). In Fig 3-4a, subject d' is plotted against model d' for both LOS and HOS models. Using a linear regression with zero-intercept, the HOS model provided a better fit to subject behavior ($r^2 = 0.85$, $p < 0.0001$) compared to the LOS model ($r^2 = 0.23$, $p < 0.0001$), which cannot match the better-performing subjects.

Fig 3-4b shows the fitted perceptual parameters (m and n) plotted against subject d'

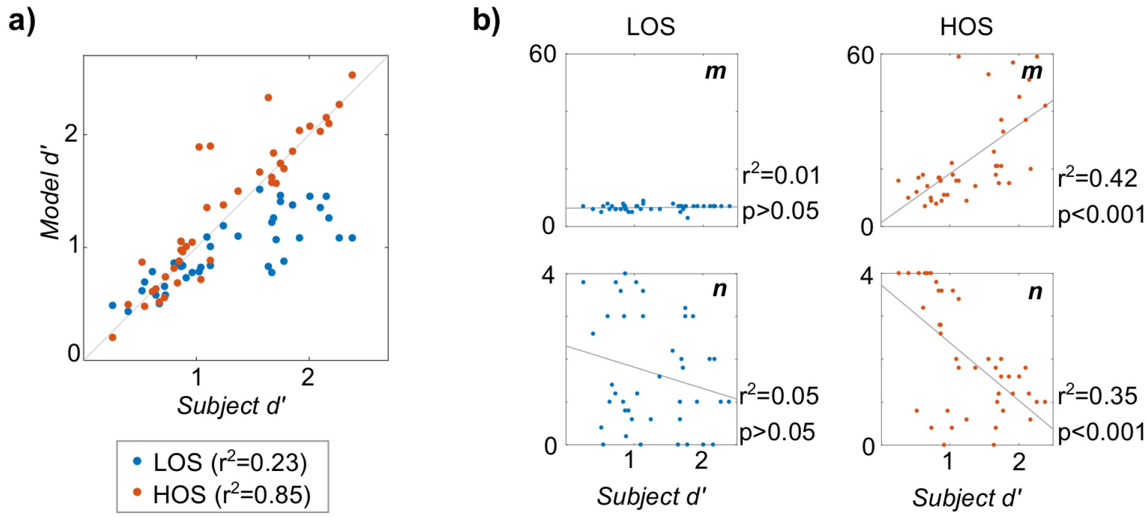

Figure 3-4. Model fit to subject behavior from Experiments 1–2. a) Subject d' plotted against fitted model d' for both LOS and HOS models, denoted by color. Legend shows r^2-value from zero-intercept linear regression. b) Fitted perceptual parameters plotted against subject d' for m (top) and n (bottom), with LOS model on the left and HOS model on the right. r^2 and p-values shown for standard linear regression.

for the LOS and HOS models. With the LOS model (left), neither perceptual parameter has a significant linear relationship with subject d' (m: $r^2 = 0.009$, $F(1, 39) = 0.359$, $p > 0.05$; n: $r^2 = 0.05$, $F(1, 39) = 2.03$, $p > 0.05$). With the HOS model (right), *both* memory and observation noise exhibit significant linear relationships with subject d' (m: $r^2 = 0.423$, $F(1, 39) = 28.6$, $p < 0.0001$; n: $r^2 = 0.352$, $F(1, 39) = 21.1$, $p < 0.0001$), with higher memory and lower observation noise corresponding with better subject performance. Similar analysis with the other model parameters (π and τ) showed no correlation with subject d' for either model.

To determine whether both perceptual parameters are needed to fit the HOS model to subject behavior, we tested a reduced model with only one of the perceptual parameters free. The memory-only HOS model, holding observation noise at $n = 0$, provided a poorer fit compared to the full HOS model shown in Fig 3-4a ($r^2 = 0.60$, $p < 0.001$), as did the observation noise-only HOS model, holding memory at the maximum stimulus length $m = 60$ ($r^2 = -0.29$, $p < 0.001$). Both memory and observation noise are needed as constraints to the model to fit the full range of human behavior.

Additionally, we compared the model changepoints to the RTs collected in Experiment 1b. Using a linear regression, the HOS model showed a significant linear relationship between model changepoint and subject RTs ($r^2 = 0.05$, $F(1, 1512) = 86.9$, $p < 0.0001$), while the LOS model showed no significant relationship. Importantly, the model was fitted using the Yes/No response only and not the RTs themselves.

3.3.3 Electroencephalography

Next, we examined neural underpinnings of higher-order stochastic regularities in the brain. Experiment 3 is structured similarly to Experiments 1 and 2 above: listeners were asked to detect changes in stochastic melodies while EEG was simultaneously recorded from central and frontal locations on the scalp. Stimuli were generated at two levels of entropy (i.e., one change degree) with both INCR and DECR change direction.

Deviance response according to melody entropy

We first examined effects of melody entropy on ERPs to individual tones. Magnitude of frequency deviation (ΔF) is known to affect ERP morphology [15], so to determine any additional effect of entropy on the ERP, we computed average ERPs for both small and large ΔF (ΔF=1 and 4 s.t. or semitones from the previous tone) at each entropy level (LOW and HIGH). Large ΔF tones are more rare in LOW entropy melodies compared to HIGH entropy melodies, so we might expect a deviance response that reflects this difference in relative occurrence (as seen in [15]). $\Delta F = 1$ was chosen because it is the most frequent in both entropy levels, and $\Delta F = 4$ was chosen to maximize frequency deviation magnitude while ensuring an adequate number of trials in the LOW entropy condition. We note that this analysis is more closely aligned with lower-order statistics, where deviance is always proportional to ΔF.

Fig 3-5a (top) shows grand-average ERPs for the four conditions averaged across frontal electrodes, which exhibited the strongest effect (described below). There is a divergence around 150-280 ms post-onset, where the ERP to large ΔF in LOW entropy (purple-dotted line) increases relative to the corresponding ERPs with the same ΔF (gray-dotted line) or the same entropy context (purple-solid line). Fig 3-5a (bottom) shows the mean amplitude in two time windows: ① 90–150ms and ② 170–260ms, corresponding roughly to N1/MMN and P2 time ranges [15]. A repeated-measures ANOVA with 2 factors (entropy and ΔF) applied to the later window showed a main effect of entropy ($F(1,7) = 7.49$, $p < 0.05$) and a trend due to ΔF ($F(1,7) = 4.57$, $p < 0.07$) with no interaction effect. Considering large-ΔF amplitudes only, a post-hoc paired t-test showed a significant difference between LOW and HIGH entropy contexts ($p < 0.05$). We performed the same t-test for each electrode; Fig 3-5a (bottom, far right) shows the p-values by electrode plotted on the scalp, with significant differences at frontal electrodes only. Similar analysis on the earlier window ① showed no effects of frequency deviation or entropy context.

An MMN response is notably absent from the ERPs in Fig 3-5a, even though large frequency deviations are rare in LOW entropy melodies. Assuming an MMN response in the brain to regularity deviations, this indicates a discrepancy between the "regularity" as defined in this analysis and the regularity collected by the brain: the MMN response is not well-differentiated by frequency deviation alone, and therefore it does not show up in this analysis. To see an MMN response, we need the proper definition of regularity in our analysis.

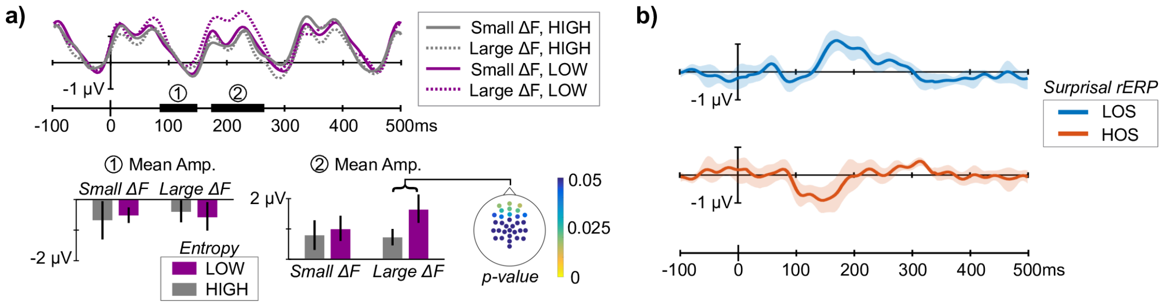

Figure 3-5. Contextual effects on tone ERP. a) Grand-average ERPs (top) for large and small ΔF in LOW and HIGH entropy melodies show a positivity for large ΔF in LOW entropy context around 200ms after tone onset. Mean amplitudes are shown for ① and ② time windows (bottom). Scalp map (right) shows frontal distribution of t-test p-values for large ΔF deflection between entropy contexts. b) Using model surprisal, regression-ERP analysis teases out distinct components depending on the set of statistics used in the model: a positivity 150-230ms after onset with LOS surprisal (similar to a) above) and an MMN-like negativity 100-200ms after onset with HOS surprisal. Error bars show 95% bootstrap confidence interval across subjects.

Deviance response according to model surprisal

The model outputs *surprisal* as a continuous measure of regularity violation, where the regularity is defined by the statistics collected by the model. We used a linear regression analysis to find contributors to the tone-elicited ERPs attributable to surprisal from the LOS and HOS models fit to individual subject behavior [69, 70]. The resulting regression ERPs (or rERPs) give a fitted regression to single-trial ERPs at each time-point for each measure of surprisal, and their interpretation is straightforward: the surprisal rERP shows the change in the baseline ERP for a unit increase in surprisal (see Methods).

Fig 3-5b shows the surprisal rERP for the LOS model (top) and HOS model (bottom). The rERPs show two distinct contributors to the ERP differing both in polarity and latency, with the LOS-rERP containing a positive deflection around 150–250ms post-onset and the HOS-rERP containing a negative deflection around 100–200ms.

To test the significance of these rERP deflections, we applied a linear mixed effects (LME) model to single trial amplitudes in the same two windows as the analysis above: 90-150ms and 170-260ms after tone onset, roughly corresponding to N1/MMN and P2 time windows. LME models are well-suited for testing single-trial effects with unbalanced designs [71], which is the case with surprisal (by definition, there are fewer surprising events than unsurprising events). In the later time window, the LME model showed a significant effect of LOS-surprisal ($p < 0.01$) on mean amplitude and no effect from HOS-surprisal. The same model applied to mean amplitude in the earlier time window showed the opposite: no significant effect from LOS-surprisal and a significant effect from HOS-surprisal ($p < 0.001$). This analysis shows deviance responses in the tone-ERP that differ depending on the statistics, or regularities, collected by the model, and an MMN-like response only to tones surprising according to the higher-order statistics of the preceding melody.

Disruption in phase-locking at model changepoint

We examined neural phase-locking to tone onsets before and after changepoints obtained from the LOS and HOS models. Phase-locking at the tone presentation rate (6.25 Hz) was measured from EEG data averaged across all 32 electrodes using the phase-locking value (PLV). PLV provides a measure of the phase agreement of the stimulus-locked response across trials, independent of power [72]. The difference in PLV before and after the changepoint (ΔPLV) measures the disruption in phase-locking at that time (see Fig 3-6a for illustration of ΔPLV calculation).

ΔPLV was measured at four sets of changepoints: the LOS and HOS model-changepoints, the nominal changepoint, and a control condition. The nominal change-point (i.e., the midpoint) is the time where the generating distributions before and after have the greatest contrast. As a control for this analysis, HOS-changepoints

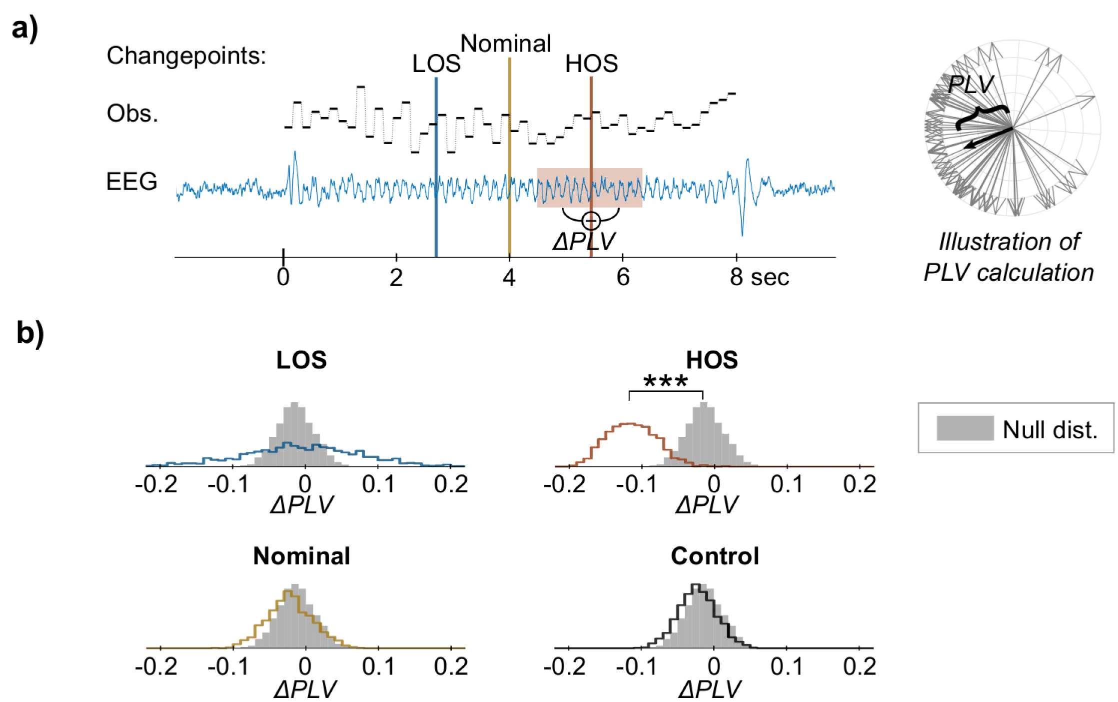

Figure 3-6. Phase-locking analysis at model changepoints. ΔPLV is used to measure disruptions in phase-locking of EEG to the tone presentation rate (6.25 Hz) at the time when the model detects a change in the stimulus (i.e., at the changepoint). a) Illustration of ΔPLV calculation. PLV measures phase agreement across trials independent of power; an example PLV calculation (right) shows the phase of individual EEG trials (in grey)—PLV is the magnitude of the mean of these normalized phasors (in black). ΔPLV is then the difference in PLV within a 7-tone (1-sec) window before and after the changepoint (left, shown at the HOS changepoint in the melody). For each subject, ΔPLV was calculated for three sets of changepoints: the changepoints output from the LOS and HOS models, and the nominal changepoint (i.e., midpoint) used to generate the stimuli. Additionally, as a control, the same HOS changepoints were applied to responses to no-change stimuli. b) Empirical distributions of ΔPLV at the LOS-, HOS-, Nominal-, and Control-changepoints (line) calculated by bootstrap sampling across subjects, along with the null distribution (solid gray) calculated by performing the same analysis with random sampling of the changepoint position. This null distribution estimates variability in ΔPLV present throughout the melody. Significant change from zero *and* from the null distribution is seen in the HOS-changepoint only.

were randomly assigned to control trials to ensure that any difference in PLV was due to the neural response recorded during change trials, and not simply due to the position of the changepoints.

Fig 3-6b shows the bootstrap distributions of the mean ΔPLV for each set of changepoints (lines). A paired t-test shows a significant decrease in PLV at the HOS-changepoints ($p < 0.001$), while there was no significant difference for the other changepoints. We also tested the ΔPLV measured at the changepoints against the variation in phase-locking present throughout the melody by estimating a null distribution, sampling null-changepoints at random positions in the melody and calculating ΔPLV. There was again a significant difference for the HOS-changepoints only ($p < 0.001$). These results together indicate there is a disruption in phase-locking that is specifically related to the changepoints obtained from the fitted HOS model.

3.4 Discussion

How the brain extracts information from stochastic sound sources for auditory scene analysis is not well understood. We investigated stochastic regularity processing using change detection experiments, where listeners detected changes in the entropy of pitches in melodies. Results from Experiments 1–2 confirmed results from previous work showing that listeners represent information about stochastic sounds through statistical estimates [14, 31]. Listeners' detection performance scaled with change degree (Experiments 1, 1b) and with the length of the sequence (Experiment 2), consistent with the use of a sufficient statistic to detect changes: a larger change in the statistic and a larger pool of sensory evidence both improved detection performance.

What statistics are collected by the brain?

We introduced a perceptual model for stochastic regularity extraction and applied this model to the same change detection experiments as our human listeners. We used different sets of statistics in the model to determine which best replicate human behavior: a lower-order statistics (LOS) model that collects the marginal mean and variance of tone pitches or a higher-order statistics (HOS) model that additionally collects the covariance between successive tone pitches. Comparing the performance range for LOS and HOS models to human performance, we showed that higher-order statistics are necessary to capture all human behaviors, while lower-order statistics are insufficient to capture the full range of subject behaviors. This disparity strongly suggests the brain is collecting and using higher-order statistics about the temporal dependencies between incoming sounds. Furthermore, the model revealed effects in EEG that are only discernible using higher-order statistics: ERP evidence showed an MMN response elicited by tones that are surprising according to the higher-order statistics of the preceding melody, and cortical phase-locking was disrupted at the changepoints specified by the HOS model.

Interestingly, both LOS and HOS models were able to replicate behavior from poorer performing subjects ($d' < 1.5$), but the LOS model is unable to mirror behaviors with high hit-rates without also increasing the FA-rate (Fig 3-3a). Intuition states that marginal statistics within the local context (i.e., short memory or small m) might be effective for detecting changes in local variance in the fractal sequences; this notion is supported by the model, where $m = 10$ tones yields the best LOS model performance (Fig 3-3b). Yet this local LOS model, with limited sampling in the statistics collected, is unable to match the performance exhibited by better performing subjects. In other words: if listeners (or the LOS model) rely solely on *marginal statistics*, then their ability to accurately flag changes in random fractal structure is highly constrained. Furthermore, relying on low-order statistics should elicit an effect of the direction of change (from low to high entropy or vice versa) on the hit-rates. Behavioral data shows no such effect of change direction on behavioral hit-rates (Experiments 1 and 1b), which further corroborates that listeners cannot be solely relying on lower-order statistics.

While these results strongly argue for the brain's ability to track higher-order statistics in sound sequences, they do not disagree with previous work demonstrating sensitivity to lower-order statistics [14, 16]. Rather, by designing a task in which higher-order statistics are beneficial, we show that listeners are additionally sensitive to the temporal covariance structure of stochastic sequences. We also do not argue that the statistics collected by the brain are limited to these, but could include longer-range covariances. We performed the same analysis using a $D = 3$ model that collects covariance between non-adjacent sounds, but it did not provide any improvement over the $D = 2$ (HOS) model. This merely means that for our stimuli, there was no additional information to aid in change detection beyond the adjacent covariances. Additional experiments with stimuli that specifically control for this are needed to determine the extent of the temporal range of statistics collected by the brain.

Individual differences revealed by stochastic processing

By their very nature, the stimuli used here exhibit a high degree of irregularity and randomness across individual instances of sequences. For the listener, deciding where the actual change in regularity occurs in a particular stimulus is a noisy process that arises with some level of uncertainty. Perceptually, most trials do not contain an obvious "aha moment" when change is detected; rather, the accumulation of evidence for statistical change emerges as a gradual process. Similarly, from a data analysis point of view, determining the exact point of time when the statistical structure undergoes a notable change is a nontrivial problem, given that the perception of statistical change is not binary but continuous and varies both between trials *and* between listeners. As such, the study of stochastic processing hinges on the use of a model that is well-matched to the computations occurring in the brain, combining the right granularity of statistics with the right scheme for cue integration and decision making. And with the introduction of perceptual parameters to the model, we gain flexibility in the behaviors that can be reproduced by the model with clear interpretation as to the computational constraints leading to these behaviors.

Taking a close look at individual differences through the lens of the model, we were able to inspect underlying roots of this variability. Rather than simply a difference in decision threshold (i.e., “trigger-happiness”), we argue the variability across listeners was due to individual differences in the limitations of the perceptual system. We incorporated these limitations into the model via perceptual parameters. The memory parameter represents differences in working memory capacity [51, 52], and the observation noise parameter represents individual differences in pitch perception fidelity [53]. We should note that these parameters may also be capturing other factors that affect listener performance like task engagement, neural noise, or task understanding, which could be contributing noise to these results. However, preliminary evidence supports the connection between the memory parameter in the model and working memory capacity, as measured by established paradigms (see Appendix II for preliminary results), and future investigation could further strengthen this claim.

By fitting the model to individual listeners through their behavior, we showed correlates between human performance and the perceptual parameters of the model, and we found that neither perceptual parameter alone was adequate to fit all subjects.

Rather than a nuisance, we see the inter-subject variability in these results as a consequence of individual differences in the perceptual system that are amplified by the uncertainty present in stochastic processing.

Neural response depends on statistical context

We found effects of the statistical context on the neural response. First, examining ERP responses to individual tones, we found an enhanced P2 response to large frequency deviations in low-entropy melodies compared to high-entropy melodies and a frontal distribution of this difference consistent with sources in the auditory cortex. This result corresponds with previous work where large frequency deviations that were *less likely* given the previous context showed an enhanced P2 amplitude [15]. Similarly, we interpret this result reflecting a release from adaptation, where the low-entropy melody has a narrow local frequency range. Importantly, we do not see an MMN effect, arguably because frequency deviation alone is too crude to provide an adequate definition of "deviant" with our stochastic stimuli: large frequency deviations do not always violate the regularities in our stimuli, which may explain the lack of an observable MMN in the average differential response.

Using the fitted model, we were able to tease out distinct surprisal effects on the tone ERP that differ both in statistics and in temporal integration window: the LOS surprisal measured how well each tone was predicted by the lower-order statistics of the local context, while the HOS surprisal measured how well each tone was predicted by the higher-order statistics of the longer context, as fit by the model to individual behavior. Because LOS and HOS surprisal are partially (and unavoidably) correlated, both LOS and HOS surprisal were included in a single regression in order to find components in the ERP that correlate with each *independent of the other* [69].

We found an enhanced P2 amplitude with increasing LOS surprisal that is similar in amplitude and latency to the P2 difference discussed above; indeed, LOS surprisal provides a similar definition of regularity to the ERP analysis based on melody entropy above, for large frequency deviations are always "deviants" according to the lower-order statistics. We again attribute this increased P2 to a release from adaptation. Consequently, we can then attribute the MMN response to HOS surprisal as a deviance response according to higher-order statistics *independent from* lower-order adaptation effects.

There has been much discussion on whether the MMN response is truly a deviance response or merely due to adaptation [73, 74]. Many experiments suffer from confounding frequency deviance with regularity deviance, making it difficult to definitively attribute MMN to one or the other. With our stochastic stimuli differing in higher-order statistics, we were able to disentangle the two interpretations. We again stress that this result is not in conflict with previous results showing effects of lower-order statistics on the MMN [14, 16], because deviants in these studies could also be considered deviants according to their higher-order statistics (i.e., the HOS model reduces to the LOS model when the covariance between sounds is zero).

Finally, we found a disruption in the brain's phase-locked response to tone onsets that coincides with HOS model changepoints, where the model detects a change in the higher-order statistics of each stimulus. Contrasting various controls using different estimates of when the change point occurs, we observed a notable phase disruption with changes in higher-order statistics only. The change in phase synchrony across trials could be due to the combined modulation of multiple ERPs to tones following the changepoint, or it could reflect a change in the oscillatory activity of the brain, which has been shown to correspond with both changes in predictive processing and attentional effects [48, 75]. Further experimentation is needed to determine the source of this disruption. Importantly, this analysis takes into account the stochastic nature of the stimuli by interpreting the statistical structure of each stimulus through the model, rather than with the changepoint used to generate the stimuli (i.e., the "nominal" changepoint).

Chapter 4

Statistical inference along multiple dimensions

4.1 Introduction

In everyday environments, the brain sifts through a plethora of sensory inputs, tracking pertinent information along multiple dimensions despite the persistent uncertainty in real-world scenes. While listening to an orchestral performance, the brain tracks variability in pitch and timbre as the music unfolds, just as it can visually track a flock of birds flying overhead despite the high uncertainty in their flight pattern and orientation. Inferring statistical structure in complex environments is a hallmark of perception that facilitates robust representation of sensory objects as they evolve along different perceptual dimensions (or features, used interchangeably). Evidence of statistical inference has been documented in audition [76–79], vision [20, 80], and olfaction [81], as well as across sensory modalities [82, 83], showing it underlies the encoding of sensory scenes in memory. These mnemonic representations then guide the interpretation of future sensory inputs.

In this chapter, we examine the mechanisms behind statistical inference along multiple dimensions. Just as in Chapter 3, we use the D-REX model to guide investigation into the nature of the brain's internal model used for predictive processing. This internal model reflects the statistics of objects in the environment, and as such, must incorporate predictive information across multiple perceptual dimensions (e.g. pitch, timbre, color, shape). The nature of this internal model as it spans multiple dimensions has often been examined by invoking learning of rules and cross-feature associations, or encoding of complex exemplars in memory [8, 77, 84–87]; and there are suggestions that this model can be based on both object- and feature-level representations, depending on whether there are dependencies across features indicating a shared source [88, 89]. Yet, structured regularities embedded in these association-based stimuli tend to over-simplify the dynamics and volatility present in real-world environments. Importantly, they conceal the granularity of the mnemonic representation as it tracks features that may not be so tightly associated even if originating from the same source or object. In the present study, we use stochastic auditory sequences to explore the internal representation of more complex regularities and the integration of statistical predictive information across features.

The oddball paradigm has been used extensively to demonstrate the brain's ability to track regularities along various auditory dimensions such as pitch, loudness, duration, timbre, and spatial location [90–93]. Many neurophysiology studies have shown that the brain makes predictions along multiple features simultaneously, where deviants co-occurring along multiple features elicit a neural response that is the sum of responses to single-feature deviants [64, 94–97]. This parallel tracking likely leverages the topographic organization in auditory cortex along different features [98, 99] (although cortical responses also show complex interactions to sounds varying along multiple dimensions [100–102]). While these studies suggest each dimension is processed independently at the prediction stage, they do not give any indication of how these independent predictions are combined at later stages of processing to give rise to integrated *object-level* percepts. It is clear through behavioral studies (and everyday experience) that listeners integrate across features to represent sound sources wholly as objects [89, 103–106]. What is not clear is the manner in which independently tracked sensory dimensions are joined into a unified statistical representation that reflects the complexity and non-deterministic nature of natural listening scenarios.

To address the limitation of quasi-predictable regularities often employed in previous studies, we again utilize stimuli exhibiting random fractal structure in a change detection paradigm, where listeners are tasked with detecting changes in entropy of sound sequences. However, in this chapter we use fractal stimuli that vary along multiple features—both spectral and spatial—and task listeners with detecting changes in entropy along one or more features. With this paradigm, we probe the ability of the brain to abstract statistical properties across features from complex sound sequences in a manner that has not been addressed by previous work. Importantly, the statistical structure of the sequences used in this study carry no particular coupling or correlation across features, hence restricting the brain's ability to leverage this correspondence in line with previously reported feature fusion mechanisms observed within and between visual, somatosensory, vestibular, and auditory sensory modalities [107–111].

In this chapter, we extend the D-REX model to multidimensional inputs in order to make inferences about the underlying computational mechanisms behind multidimensional predictive coding in the brain. We use this model as a framework to ask targeted questions about statistical integration in complex listening environments: Which statistics are tracked along each feature? When does integration across features occur? Are features combined linearly or through some other function? The model is used to formulate alternative hypotheses addressing these questions and compare them using the proposed behavioral paradigm. In addition, we use the output of the model as an anchor for time-locking analysis of neural responses, combating the temporal uncertainty that invariably creeps into the analysis of stochastic responses to stochastic stimuli.

4.2 Methods

We conducted four experiments: two psychophysics experiments (experiments SP and TP) and two similarly structured electroencephalography (EEG) experiments (experiments nSP and nTP, with 'n' denoting neural). In experiments SP and nSP, stimuli varied in spatial location (S) and pitch (P), as denoted by the naming convention; in experiments TP and nTP, stimuli varied in timbre (T) and pitch (P).

4.2.1 Participants

In experiment SP, sixteen participants (8 Female) were recruited from the general population (mean age: 25.1 years); one participant was excluded from further analysis because their task performance was near chance ($d' < 0.05$). In experiment TP, eighteen participants (12 Female) were recruited (mean age: 21.5 years); three participants were excluded due to chance performance. In experiment nSP, twenty participants (9 Female) were recruited (mean age: 23.4); two participants were excluded due to chance performance. In experiment nTP, twenty-two participants (13 Female) were recruited (mean age: 22.5); four participants were excluded due to chance performance.

All participants reported no history of hearing loss or neurological problems. Participants gave informed consent prior to the experiment and were paid for their

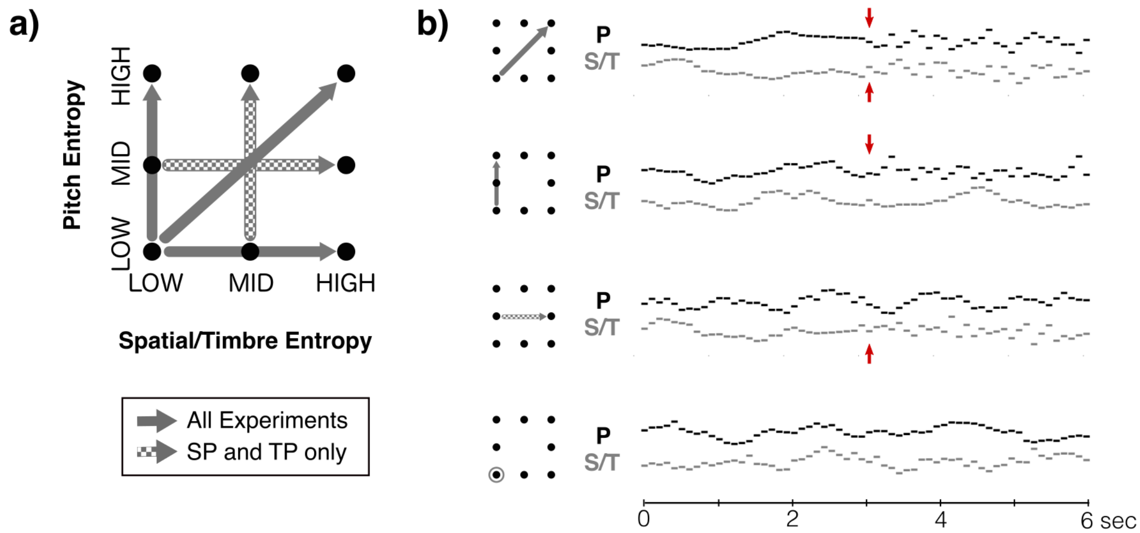

Figure 4-1. Multidimensional fractal stimuli. a) Stimuli were melodies comprised of tones varying according to fractal structure along two dimensions simultaneously: Pitch & Spatial location (in experiment SP and nSP) or Pitch & Timbre (in experiment TP and nTP). At the midpoint of the melody, one or both features increased in entropy (non-diagonal and diagonal arrows, respectively), while the non-changing feature remained at low-entropy. For psychophysics experiments (SP and TP), the non-changing feature could also have mid-level entropy (checkered arrows). b) Four example stimulus sequences with condition indicated by small schematic on left. Red arrows indicate change in each feature, when present. The bottom example is a control trial with no change. (See Supplementary Materials for audio examples.)

participation. All experimental procedures were approved by the Johns Hopkins IRB.

4.2.2 Stimuli

Stimuli in all experiments were melodies comprised of a sequence of complex tones varying along two perceptual features. Stimuli in experiments SP and nSP varied in pitch and spatial location; stimuli in experiments TP and nTP varied in pitch and timbre. Each feature followed the contour of a random fractal at different levels of entropy, or randomness.

Random fractals are stochastic processes with spectrum inversely proportional to frequency with log-slope β (i.e., $1/f^{\beta}$), where β parameterizes the entropy of the sequence. Fractals at three levels of entropy were used as seed sequences to generate the stimuli: low ($\beta = 2.5$), mid ($\beta = 2$), and high ($\beta = 0$, white noise). In all experiments stimuli began with both features at lower entropy, and halfway through the melody, one or both features increased to high entropy. In the psychophysics experiments (SP and TP) for conditions with a single feature changing, the non-changing feature could have either low or mid entropy. In the EEG experiments (nSP and nTP), the non-changing feature always had low entropy. Control conditions contained stimuli with no entropy change in either feature. See Fig 4-1 for an illustration of the different stimulus conditions in each experiment.

Each complex tone in the melody sequence was synthesized from harmonic stack of sinusoids with frequencies at integer multiples of the fundamental frequency, then high- and low-pass filtered at the same cutoff frequency using fourth-order Butterworth filters. Pitch was manipulated through the fundamental frequency of the complex tone, and timbre was manipulated through the cutoff frequencies of the high- and low-pass filters (i.e., the spectral centroid) [102]. Spatial location was simulated by convolving the resulting tone with interpolated head-related impulse functions for the left and right ear at the desired azimuthal position [112]. Seed fractals were generated independently for each feature and each stimulus, standardized (i.e., zero mean and unit variance), and then mapped to feature space as follows:

$$\begin{aligned} F_0[t] &= 350 * 2^{\,3x[t]/12} \\ S[t] &= 15y[t] \\ T[t] &= 1200 * 2^{\,3z[t]/12} \end{aligned}$$

where $F_0[t]$, $S[t]$ and $T[t]$ are pitch (fundamental frequency in Hz), spatial location (azimuth in degrees), and timbre (spectral centroid in Hz) sequences indexed by time

t. $x[t]$, $y[t]$, and $z[t]$ are their respective seed fractals. Fundamental frequency ranged from 208 to 589 Hz, spatial location ranged from $-45°$ to $45°$ azimuth at $0°$ elevation, and spectral centroid (timbre) ranged from 714 to 2018 Hz.

In experiments SP and TP, melody stimuli were comprised of 60 complex tones, each 100 ms in duration with 20 ms onset/offset ramps presented isochronously at a rate of 10 Hz. 200 stimuli were generated, 25 for each condition (5 change, 3 no-change). In experiments nSP and nTP, melody stimuli were comprised of 60 complex tones, each 100 ms in duration with 20 ms onset/offset ramps presented isochronously at a rate of 8.6 Hz. 200 stimuli were generated, 50 for each condition (3 change, 1 no-change).

4.2.3 Procedure

Stimuli were presented in randomized order in four blocks with self-paced breaks between blocks. During each trial, participants were instructed to listen for a change in the melody. After the melody finished, participants responded via keyboard whether or not they heard a change. Immediate feedback was given after each response.

Listeners were not given explicit instructions about what to listen for, learning the task implicitly in a training block prior to testing. Incorrect responses in the training block resulted in the same stimulus being re-played with feedback (including, in the case of missed detections, a visual indication of change during playback).

Stimuli were synthesized on-the-fly at 44.1 kHz sampling rate and presented at a comfortable listening level using PsychToolbox (psychtoolbox.org) and custom scripts in MATLAB (The Mathworks, Natick, MA). Participants were seated in an anechoic chamber in front of the presentation screen.

In experiments SP and TP, stimuli were presented via over-ear headphones

(Sennheiser HD 595) and participants responded via keyboard. The experiment duration was approximately 50 minutes. In experiments nSP and nTP, stimuli were presented via in-ear headphones (Etymotic ER-2) and participants responded via response box. Additionally, before each melody trial, a fixation cross appeared on the screen to reduce eye movement during EEG acquisition. The experiment duration, including EEG setup, was approximately 120 minutes.

4.2.4 EEG data recording and analysis

EEG data in experiments nSP and nTP was recorded using a BioSemi ActiveTwo system (BioSemi, Amsterdam, Netherlands) with 64 electrodes placed on the scalp according to the international 10-20 system, along with two additional electrodes specified by the BioSemi system used as online reference for common-mode rejection. Data was recorded at a sampling rate of 2,048 Hz.

For each subject, EEG data were preprocessed with custom scripts in MATLAB using the Fieldtrip toolbox (www.fieldtriptoolbox.org) [113]. Bad channels were identified by eye and removed before proceeding with pre-processing. Continuous EEG was filtered to 0.3–100 Hz (two-pass 4th-order Butterworth for high-pass and 6th-order Butterworth for low-pass) and re-sampled to 256 Hz. Data was then cleaned in three stages: the Sparse Time Artifact Removal algorithm (STAR) was used to remove channel-specific artifacts [114], Independent Component Analysis (ICA) was used to remove artifacts due to eye movement and heartbeat, and missing channels were interpolated using spline interpolation. The cleaned data was then epoched by melody trial (-1 sec to 8 sec, relative to melody onset), re-referenced to the average of all 64 scalp electrodes, and baseline corrected to the 1 sec window preceding melody onset. Epochs with power exceeding 2 s.d. from the mean were removed from further analysis (on average, 3.8% of trials excluded in nSP, 5% in nTP).

We examined neural responses time-locked to outputs from the `Late_D22_MAX` model by further epoching neural response around events of interest (-0.1 to 0.3 sec relative to tone onset).

In the oddball analysis, the EEG response was averaged over nine fronto-central electrodes (Fz, F1, F2, FCz, FC1, FC2, Cz, C1, C2) to maximize auditory-related responses. High and low surprisal events were defined as tones with overall surprisal above the 95^{th} and below the 5^{th} percentile, respectively. Tone-epochs within each bin were averaged, and the high-surprisal response was subtracted from the low-surprisal response to yield a difference wave.

To examine the linear relationship between the EEG response magnitude and surprisal, tone-epochs across all stimuli were split into 40 bins according to overall surprisal, and tone-epochs with power exceeding 2 s.d. from the mean were excluded from analysis (average bin size per subject: 185 epochs). The average response across tone-epochs within each bin was calculated, and the cumulative response magnitude was computed over in the window 80–150 ms after tone onset and plotted against the average surprisal within each bin. A similar analysis was performed using the individual surprisal along each feature using 128 bins (average bin size per subject: 66 epochs), where the bins were determined by bifurcating the 2-D surprisal space across all tones.

We examined the neural response time-locked to high surprisal and to maximal belief change in two time windows: 80–150 ms and 300–800 ms. In each window, 10 channels with largest amplitude in the grand average (5 positive, 5 negative polarity) were selected for statistical analysis. For each subject, response magnitude was measured as the dB RMS amplitude across channels averaged over the time window relative to a baseline window (-152 to -82 ms and -630 to -130 ms for the early and late windows, respectively).

4.2.5 Model

The D-REX model was extended to multidimensional predictive processing with multiple potential implementations. These models differed in the statistics collected along each feature, in the integration stage, and in the integration operator.

The statistics collected by the model were specified separately for each feature by the dimensionality D. This parameter took two values: with $D = 1$ the model assumed inputs were statistically independent, collecting only lower-order statistics (mean and variance); with $D = 2$ the model assumed temporal dependence in the input sequence, and collected higher-order statistics (i.e., covariance between adjacent inputs).

Upon observing a new input x_{t+1}, all models produce independent predictive probabilities for each feature and for each context hypothesis: for example, p_i^S and p_i^P, where $m \in \{1, \ldots, M\}$ denotes the context hypothesis and the superscript denotes the feature. The integration stage and integration operator specified where and how information was integrated across features. With early-stage integration, predictions within each context hypothesis were combined before updating shared context beliefs B_t and outputting a shared change signal. With late-stage integration, the context was inferred separately for each feature with distinct context beliefs (e.g., B_t^P and B_t^S) and change signals, and integration occurred across change signals. In early and late integration, four integration operators were used: average, weighted average, minimum, and maximum. For the weighted average, weights between 0 and 1 in steps of 0.1 were used for convex weighting of the two features, and the weight yielding the best fit for each subject was selected for comparison (more details on model fitting below).

In total, there were 32 variants of the model (2 D x 2 D x 2 stages x 4 operators).

To fit the models to individual listeners in experiments nSP and nTP, a grid search with 95,000 iterations was used to find parameters M, N, and τ (memory, observation noise, and decision threshold, respectively) that best replicated listener behavior for each model variant. The model detection rate (i.e., percentage of trials wherein a change was detected) in each condition was collected for each iteration in the search procedure, and the parameters resulting in the least mean squared error in detection rate across conditions between model and listener behavior was selected. A modified hinge loss was then used to compute goodness-of-fit for each model: this loss function penalized both incorrect model responses and correct responses close to threshold (i.e., correct with low certainty), thus rewarding models with decision signals far from threshold (i.e., correct with high certainty). Note that "correct" in this case is the response from the individual subject being fit.

4.3 Results

4.3.1 Perceptual experiments

We conducted four experiments to probe the mechanisms behind predictive processing along multiple dimensions in auditory perception: two psychophysics experiments (experiments SP and TP) and two similarly structured electroencephalography (EEG) experiments (experiments nSP and nTP, with 'n' denoting neural). Listeners were asked to detect changes in the statistical properties of a sequence of complex sounds varying along two perceptual features: in experiments SP and nSP, stimuli varied in spatial location (S) and pitch (P), as denoted by the naming convention; in experiments

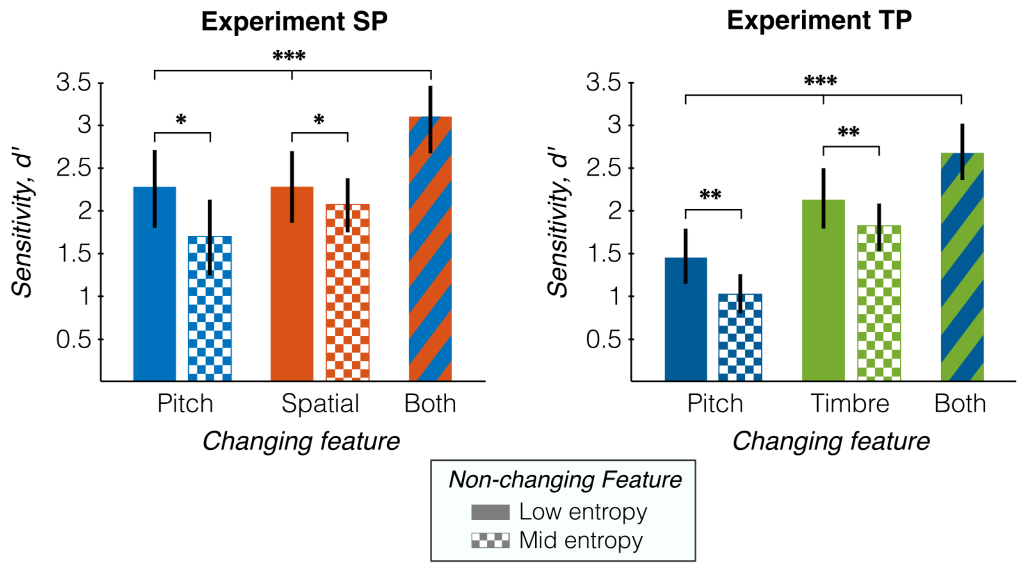

Figure 4-2. Behavioral results for experiments SP and TP. Average change detection performance (d') is shown by changing feature (abscissa) and entropy of non-changing feature (fill pattern). Error bars indicate 95% bootstrap confidence interval across subjects (N=15 for both experiments).

TP and nTP, stimuli varied in timbre (T) and pitch (P). Changes could occur in one, both, or none of the features (see Fig 4-1). Conditions were randomized, so listeners did not know *a priori* at the beginning of each trial which feature was informative for the task.

Detection performance improves with feature conjunction

Fig 4-2 shows detection performance in psychophysics experiments SP (left) and TP (right). To establish whether listeners integrated information across features to perform the change detection task, we compared single- and both-change conditions, with the non-changing feature at low-entropy (excluding mid-entropy conditions, checkered bars in Fig 4-2).

In experiment SP, an ANOVA with 1 within-subject factor (3 conditions) showed strong significant differences between conditions ($F(2, 28) = 12.07$, $p = 0.0002$), with post-hoc paired t-tests confirming the effect between *Both* and each single-change condition (*Both* vs. *Pitch*, $t(14) = 6.12$, $p < 0.0001$; *Both* vs. *Spatial*, $t(14) = 4.64$, $p = 0.0004$). In addition, a more stringent test showed that for each subject, performance in the *Both* condition was significantly better than the highest of the two single-change conditions (*Both* vs. max(*Pitch*, *Spatial*), $t(14) = 3.70$, $p = 0.0024$).

We found the same effects in experiment TP. The ANOVA showed strong differences between change conditions ($F(2, 28) = 23.74$, $p < 0.0001$), with post-hoc paired t-tests confirming the effect between *Both* and each single-change condition (*Both* vs. *Pitch*, $t(14) = 7.77$, $p < 0.0001$; *Both* vs. *Timbre*, $t(14) = 3.35$, $p = 0.0047$). The more stringent test also showed that each subject performed significantly better in the *Both* condition compared to the maximum of the single-change conditions (*Both* vs. max(*Pitch*, *Timbre*), $t(14) = 3.01$, $p = 0.0093$).

We replicated the same analysis for behavioral responses in the EEG experiments nSP and nTP (not shown in figure). Listeners performed the same change-detection task, with the only difference being the exclusion of the mid-entropy conditions (checkered bars in Fig 4-1). We observed the same behavioral effects as above in the EEG experiments: detection performance increased in the *Both* condition relative to each of the single-change conditions (nSP: *Both* vs. max(*Spatial*, *Pitch*), $t(17) = 4.86$, $p = 0.00015$; nTP: *Both* vs. max(*Timbre*, *Pitch*), $t(17) = 3.29$, $p = 0.0043$).

If listeners were processing each feature completely independently, we would expect performance in the *Both* condition to be, at most, the maximum of the two single-change conditions. Instead, the apparent increase in detection performance suggests that listeners can flexibly integrate predictive information when corroborative evidence across features is available.

Higher entropy in uninformative feature increases false alarms but not missed detections

In a second analysis of experiments SP and TP, we looked at whether the uninformative (i.e., non-changing) feature could disrupt change detection in the informative (i.e., changing) feature. We compared performance in the single-change conditions when the non-changing feature was low- vs. mid-entropy (excluding the *Both* condition, striped bars in Fig 4-2).

In experiment SP, an ANOVA with 2 within-subject factors (2 changing feature x 2 entropy of non-changing feature) showed a significant main effect of entropy ($F(1,42) = 5.01$, $p = 0.031$), and no effect of changing feature ($F(1,42) = 1.15$, $p = 0.29$) or interaction ($F(1,42) = 1.12$, $p = 0.30$). Interestingly, post-hoc t-tests showed that the decrease in performance was due to an increase in false alarms (FAs) (*Pitch*/*Spatial* entropy: Low/Low vs. Low/Mid, $t(14) = -7.44$, $p < 0.0001$); Low/Low vs. Mid/Low, $t(14) = -2.48$, $p = 0.013$) and not a decrease in hit-rates (same ANOVA as above applied to hit-rates: Entropy $F(1,42) = 2.82$, $p = 0.10$, Feature $F(1,42) = 0.44$, $p = 0.51$, Interaction $F(1,42) = 0.55$, $p = 0.46$).

We found similar effects in experiment TP. The ANOVA showed a significant main effect of entropy ($F(1,42) = 8.00$, $p = 0.0071$) and no interaction effect ($F(1,42) = 0.28$, $p = 0.60$), but it did show a main effect of changing feature ($F(1,42) = 32.03$, $p < 0.0001$). This difference between the *Pitch* and *Timbre* conditions likely reflects a difference in task difficulty due to stimulus design, rather than a persistent effect due to the features themselves or an interaction between the two. As for the main effect of non-changing entropy, post-hoc t-tests again showed the decrease in detection performance was due to an increase in FAs (Pitch/Timbre entropy: Low/Low vs. Low/Mid, $t(14) = -5.91$, $p < 0.0001$); Low/Low vs. Mid/Low, $t(14) = -3.93$,

$p = 0.00075$) and not a decrease in hit-rates with higher entropy (same ANOVA as above applied to hit-rates: Entropy $F(1, 42) = 3.5$, $p = 0.068$, Feature $F(1, 42) = 29.48$, $p < 0.0001$, Interaction $F(1, 42) = 1.75$, $p = 0.19$).

The uninformative feature did in fact affect overall detection performance, where higher entropy led to increased FAs. However, as hit-rates did not decrease as well, listeners' ability to track statistics in the informative feature was not disrupted by the uninformative feature, even when the identity of informative and uninformative feature changed from trial to trial. This result suggests that statistics are collected *independently* along each feature, and integration across features occurs *after* statistical estimates have been formed.

4.3.2 Computational model

Behavioral results so far demonstrate that listeners collect statistics independently along multiple features and then integrate across features at some later processing stage, begging the question of *how* this combination occurs. To answer this, we formulate a model for multidimensional predictive processing to appraise different hypotheses for the underlying computational mechanism that could lead to listener behavior.

As a starting point, we use the Dynamic Regularity Extraction (D-REX) model described in Chapter 2, which was initially formulated for statistical inference along a single feature [49]. To make specific hypotheses for how predictive processing operates along multiple features, we constructed model variants that differ in: (i) the statistics collected along each feature, (ii) the processing stage at which integration occurs, and (iii) the function or operator used to combine across features. Source code is available at: https://engineering.jhu.edu/lcap/downloads.

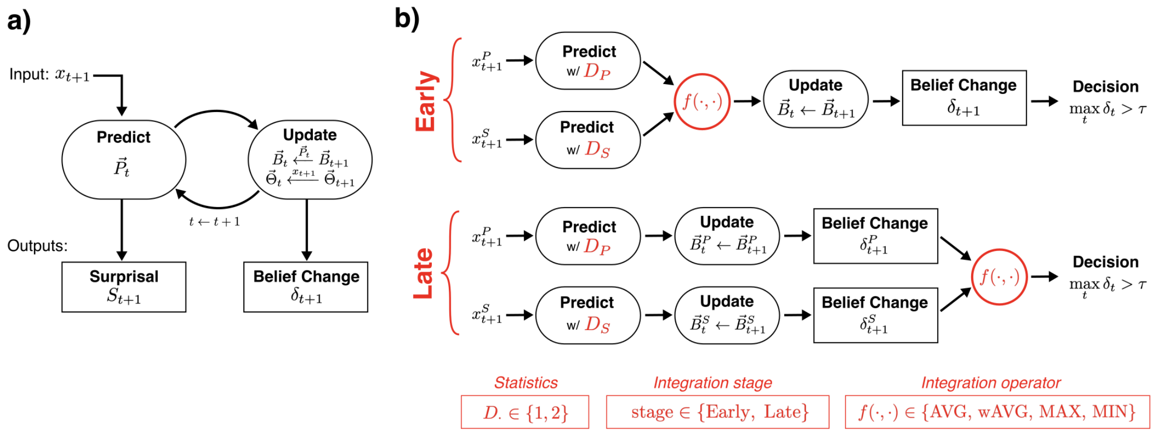

Figure 4-3. Multidimensional model schematic. a) Building blocks of the model for predictive processing along a single dimension. b) Illustration of potential variants of the model for statistical inference along multiple dimensions. Red indicates aspects of the model that differed by variant: statistics collected along each dimension ($D \in \{1, 2\}$), early- vs. late-stage integration, and the operator used in integration (MAX, MIN, AVG, wAVG). Summary of model variants in red boxes at bottom.

In the next section, we give a brief description of how the D-REX model was used to formulate hypotheses for the computational mechanisms behind predictive processing of multi-feature sounds.

Building blocks of statistical inference

The D-REX model makes sequential predictions of the next input x_{t+1} given all previously observed inputs $x_1, x_2, \ldots, x_t$. In the present study, the input $\{x_t\}_{t\in\mathbb{Z}^+}$ is a sequence of pitches, spatial locations, or spectral centroids (timbre). This sensory input is assumed to be successively drawn from a multivariate Gaussian distribution with unknown parameters, as this structure fits a wide range of natural and experimental phenomena [14, 16, 45, 115–117]. Over time, the model collects sufficient statistics $\hat{\theta}$ from observed inputs to estimate the unknown distribution parameters [43].

The D-REX model has two main processing stages: a prediction stage and an update stage. Fig 4-3a illustrates these main processing stages for a single time-step.

Upon observing the new input x_{t+1}, the model first computes the set of predictions $\vec{P}_t$ using the collected statistics $\vec{\Theta}_t$ across context hypotheses (see "Predict"). The model then incrementally updates two quantities (see "Update"): the beliefs $\vec{B}_t$ are updated with new evidence from $\vec{P}_t$ based on how well x_{t+1} was predicted under each context hypothesis, and the set of statistics $\vec{\Theta}_t$ are updated with the newly observed input x_{t+1}. These are in turn used for predicting the subsequent input at time $t+2$, and so on.

In this work, we consider two outputs from the model that reflect different levels of uncertainty and dynamics in the input:

- *Surprisal* is a local measure of probabilistic mismatch between the model prediction and the just-observed input:

  $$S_{t+1} = -\log \mathbb{P}(x_{t+1}|x_{1:t})$$

  where S_{t+1} is the surprisal at time $t+1$, based on the predictive probability of x_{t+1}..

- *Belief Change* is a global measure of statistical change in the input sequence derived from the context beliefs. If the new input x_{t+1} is no longer well predicted using the beliefs $\vec{B}_t$ (e.g., after a change in underlying statistics), the updated beliefs $\vec{B}_{t+1}$ shift to reflect the change in context inferred by the model. The belief change δ_t measures the distance between these two posterior distributions before and after x_{t+1} is observed:

  $$\delta_t = D_{KL}(\vec{B}_t \mid\mid \vec{B}_{t+1})$$

  where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence. This measure ultimately reflects dynamics in the global statistics of the observed sequence. In contrast to the change probability defined in Chapter 3, this measure of change does not assume a single change in the input observations.

We derived a change detection response from the model that is analogous to listener behavioral responses by applying a detection threshold τ to the maximal belief change δ_t:

$$\text{Model Response} = \max_t(\delta_t) \geq \tau$$

We use this response to compare the model to listeners' behavioral responses. In addition, we use the moment *when* this maximal belief change occurs, along with surprisal, to examine the neural response related to different dynamics in the stimuli.

Modeling statistical inference along multiple dimensions

Now, let the input sequence x_t be multidimensional with two components along separate dimensions, e.g., pitch and spatial location: $x_t = \{x_t^P, x_t^S\}$. The extension of the D-REX model to multidimensional inputs is not trivial. In this study, we use the D-REX model as a springboard to entertain multiple hypotheses about how statistical inference operates across multiple dimensions. Fig 4-3b illustrates three attributes of the model we explore (indicated in red):

- *Statistics D.* Listeners potentially collect different statistics along different dimensions. In the model, sufficient statistics are specified by the D parameter, the dimensionality of the Gaussian distribution, or the temporal dependence, assumed by the model. In the proposed multidimensional model, there are two D parameters, one for each feature (see "Predict" in Fig 4-3b). We examine model variants with $D = 1$ (no temporal dependence) and $D = 2$ to test what statistics are tracked along each feature.

- *Integration stage.* Building on previous neural evidence for independent predictions along different dimensions, the model generates predictions separately along each feature. We examine two possible stages for combining across dimensions after the prediction: Early-stage integration (Fig 4-3-top), where predictions are combined across features before updating context beliefs, and Late-stage integration (Fig 4-3-bottom), where the belief change δ_t is computed separately for each feature and combined before the final decision. These two alternatives represent whether the context window for estimating statistics is inferred jointly across features (Early) or independently for each feature (Late).
- *Integration operator* $f(\cdot,\cdot)$. We test four different operators for how predictive information is combined across features: two linear operators, average (AVG) and weighted average (wAVG), where the relative weighting between features is adapted to each listener; and two non-linear operators, minimum (MIN) and maximum (MAX). These operators are applied at the processing level specified by the integration stage.

We examine models with each permutation of these attributes, yielding 32 variants of the model (2 D x 2 D x 2 stage x 4 operator). In the following section, we examine which variant best replicates human behavior.

Model comparison to listener behavior

We fit parameters of each model to individual listener behavior in Experiments nSP and nTP. In addition to the decision threshold τ, there are two parameters of the model that reflect neural constraints individual to each listener: the memory parameter M sets an upper bound on the context window (and the number of context hypotheses), and the observation noise parameter N sets a lower bound on prediction uncertainty, adding independent Gaussian noise with variance N to the predictions. These parameters represent plausible constraints on perception known to vary across individuals: the former representing working memory capacity [51, 52] and the latter, perceptual fidelity [53, 54].

Models with early-stage integration have a single memory parameter, due to shared context beliefs across features; models with late-stage integration have two memory parameters (one for each feature). All models have two observation noise parameters and a single decision threshold. For each model and listener, these parameters were fit using a grid search of the parameter space, where change detection responses from the model were compared against the same responses from listeners, and a loss function was used to determine the goodness-of-fit of each model (see SI-Fig S9 for examples of belief change outputs across model variants). Note that in this comparison, ground truth is not whether there *was* a change in the stimulus itself, but whether the individual listener *detected* a change.

Fig 4-4 shows the loss by model (rows) and subject (columns) after the fitting procedure for Experiments nSP and nTP. For each experiment, models are ordered by decreasing average loss (top row, minimum average loss) and subjects are ordered by increasing detection performance d' (right column, highest d'). Model variants are labeled according to the configuration illustrated in Fig 4-3b: `stage_DXX_operator`, where `XX` specifies the statistics (1 or 2) used for each feature. For example, in Experiment nSP the `Early_D12_MAX` model uses early-stage integration, $D = 1$ for Pitch, $D = 2$ for Spatial, and the MAX operator for integration.

The column to the right of each fit matrix in Fig 4-4 shows the average loss across all subjects. The model labels reveal high agreement in the top-performing models fit across Experiment nSP and nTP—in fact, the ordering of the top 11 models is

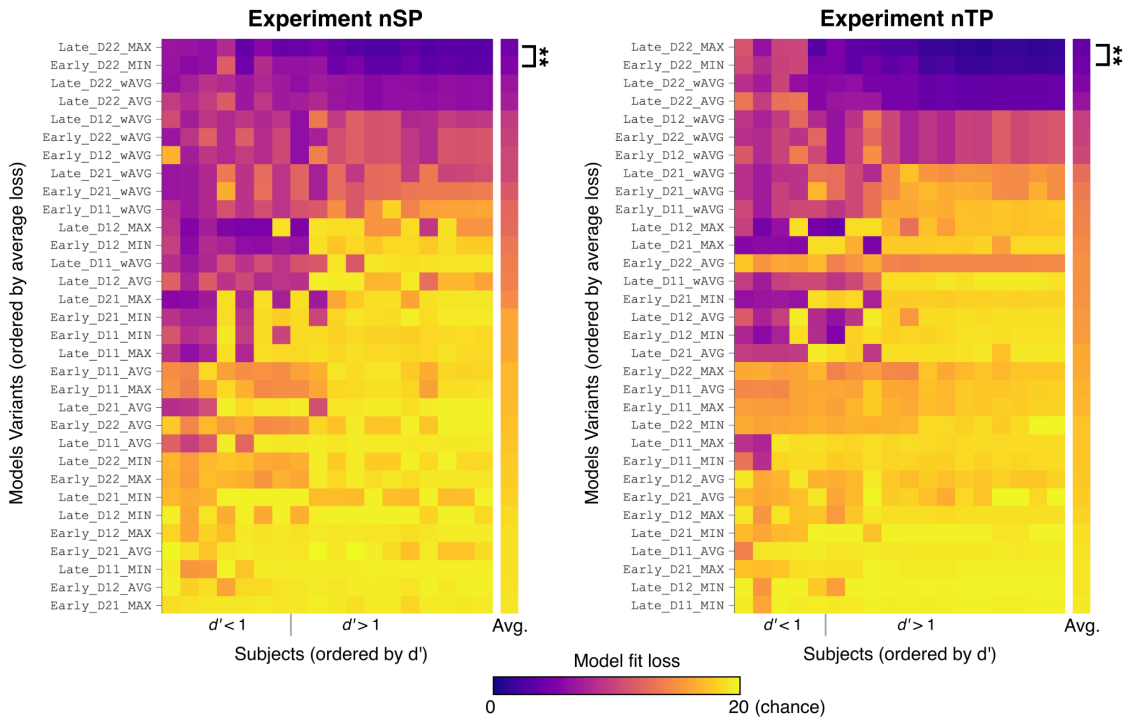

Figure 4-4. Model comparison for Experiments nSP (left) and nTP (right). Each model variant was fit to individual subjects and the resulting loss is displayed by color. Each row is a model variant (ordered by average loss) and each column is a subject (ordered by d'). Model names to the left of each image indicate integration stage, statistics (D) collected for each feature, and integration operator. The two best models, `Late_D22_MAX` and `Early_D22_MAX`, were compared using a t-test. (N=18 in both experiments)

identical across experiments. Notably, model Late_D22_MAX yields the best fit *on average* across all subjects for both experiments. Late_D22_MAX has a better fit than *all* other variants of the model. Specifically, Late_D22_MAX has a significantly lower loss (i.e., better fit) across subjects when compared to the next best model, Early_D22_MIN, in both experiments (nSP: $t_{17} = -3.82$, $p = 0.0014$; nTP: $t_{17} = -3.63$, $p = 0.0021$).

With the poorer fitting models in the lower half of Fig 4-4, model variants with Early&MAX or Late&MIN have a fit loss near chance. This is not surprising given that both are less sensitive to changes: the Early&MAX models only detect changes when *both* features violate prediction, and similarly the Late&MIN models require the change signal of *both* features to cross threshold. Neither of these types of models fit listener behavior well. Additionally, models with lower-order statistics (i.e., $D = 1$) in one or both features tend to have poorer fits (and higher loss).

Together, these results suggest that with both spectral and spatial features, listeners track higher-order statistics separately along each feature and integrate at a later stage, making a non-linear change decision based on the feature with the most evidence for change. In later analyses, we use this fitted model to guide analysis of neural responses.

Model interpretation of individual differences

Looking closer at variability in model loss across individuals in Fig 4-4, some patterns emerge across experiments nSP and nTP. For better-performing subjects ($d' > 1$, right side of each image), there is high agreement in loss across all model variants. For poorer-performing subjects (left side of each image), there is more variability in model fit across subjects, with some model variants with higher overall loss fitting individual subjects quite well. For example, in Experiment nSP (Fig 4-4-left) the Late_D12_MAX model has loss near chance for subjects with $d' > 1$, but for subjects with $d' < 1$, loss is near zero. This suggests that variability in task performance across subjects could be due to different listening strategies—these could relate to inherent ability for tracking statistics of sound sequences or differences in task understanding.

We can also examine how individual differences are explained by the model parameters fit to each subject. Using the `Late_D22_MAX` model, the "best" overall model, we tested for correspondence between the four perceptual parameters (memory and observation noise for each feature) and detection performance across listeners. In experiment nSP, a multiple linear regression explained 82% of the variance in d' and showed strongly significant correlation between both memory parameters and detection performance (M_S:, $p = 0.0070$, M_P: $p = 0.0004$) and no significant correlation between the observation noise parameter and performance in either feature (N_S:, $p = 0.82$, N_P: $p = 0.33$). We see similar results in experiment nTP, with the perceptual parameters accounting for 81% of the variance in d' and significant correlation between the spatial memory parameter with weaker significance in the pitch memory parameter (M_T: $p = 0.0009$, M_P: $p = 0.0975$, N_T: $p = 0.87$, N_P: $p = 0.54$). Fig 4-5 shows the fitted memory parameters for each feature plotted against overall d' for experiment nSP (left) and nTP (right), along with the multiple linear regression. This result suggests that the differences in behavior across listeners in experiment nSP and nTP could be due to differences in memory capacity rather than difference in perceptual fidelity (as represented by observation noise), where better-performing subjects use higher memory capacity for statistical estimation in each feature.

We additionally tested for correlations *between* memory parameters across feature. Linear regression showed significant correlations in memory across features in both experiments (nSP: $\rho = 0.53$, $p = 0.0232$; nTP: $\rho = 0.61$, $p = 0.0076$). This result holds implications for the independence of neural resources used in statistical predictive

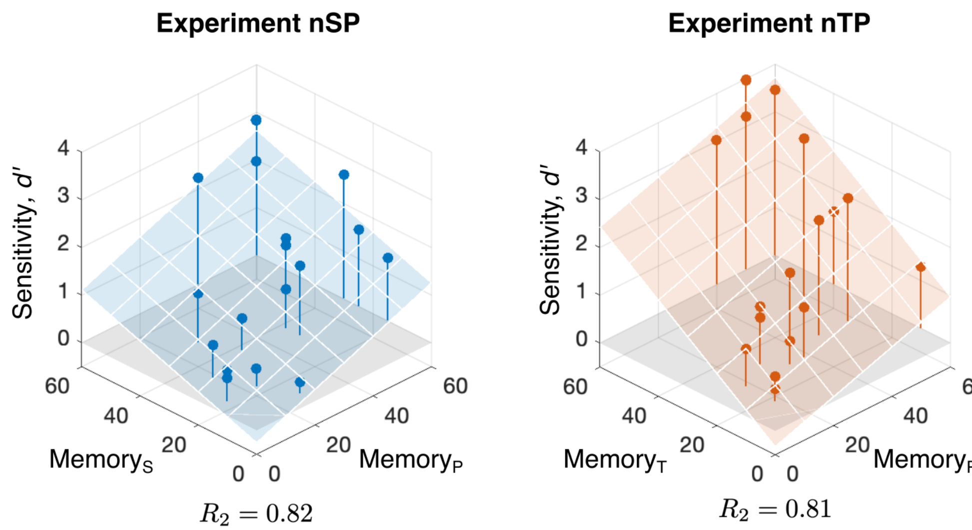

Figure 4-5. Memory parameters of the `Late_D22_MAX` model fit to individual subjects in Experiments nSP (left) and nTP (right). Fitted memory parameters plotted against overall detection performance d', along with multiple linear regression fit (R^2 at bottom of each plot). Observation noise parameters (not shown) did not have significant correlation with d'.

processing: While predictions occur separately across features, this suggests that the working memory capacity for statistical estimation is linked across features.

4.3.3 Electroencephalography

The model simulates predictive processing moment-by-moment, giving a window into the underlying processes that cannot be observed through behavior. In this section, we use the `Late_D22_MAX` model to guide analysis of neural responses in experiments nSP and nTP.

Two model outputs were used to specify epochs for trial-averaging: surprisal, the local measure of deviance between each observation and its prediction; and maximal belief change, the global measure of melody-level statistics when the largest change in beliefs occurs in each trial. Note that there are distinct surprisal responses for each feature, e.g., each tone in the melody elicits a surprisal in pitch and a surprisal in spatial location from the model. In comparison, the maximal belief change occurs once in each trial and reflects more global statistical processing of the stimulus sequence.

Neural response magnitude increases with local surprisal

We used model surprisal to perform an oddball-like analysis of neural responses. While this type of analysis typically relies on deterministic patterns to define "deviant" and "standard" events, without such structure we use surprisal from the model to guide identification of tones that fit predictions well and those that do not. First, we use an overall measure of surprisal to define "deviant" and "standard" by summing surprisal across features, e.g., $S_t = S_t^P + S_t^S$, where S_t^P and S_t^S are the surprisal from pitch and spatial location, respectively. We compared the neural response time-locked to high-surprisal tones to the response time-locked to low-surprisal tones, where high and low were defined as the top and bottom 5%, respectively, for each subject. In this analysis, we averaged the EEG response across fronto-central electrodes typically used in auditory analyses (according to 10/20 system: Cz, C1, C2, FCz, FC1, FC1, Fz, F1, F2).

Fig 4-6a shows the grand-average response to high- and low-surprisal tones along with their difference wave for experiments nSP and nTP. High-surprisal tones elicit a larger magnitude response relative to low-surprisal tones, as can be seen in deviations in the difference wave from 0 μV at typical N1 and P2 time windows. Topography in Fig 4-6a shows amplitude of differential response in the 80–150 ms window after tone onset, along with channels used in this analysis. Note the oscillations in the grand-average response are entrained to tone onsets (every 116 ms) – the response to high surprisal tones augments this obligatory onset response.

To determine if there is a linear relationship between overall surprisal S_t and the neural response, we took advantage of surprisal as a *continuous* measure of probabilistic deviance to bin tones across all trials into 40 equal-sized bins by overall surprisal. We then averaged the neural response within each bin across subjects and across tone epochs, and extracted the neural response magnitude 80-150 ms after tone onset (corresponding to typical N1/MMN time window, overlaid on difference wave in Fig 4-6a). Fig 4-6b shows EEG magnitude plotted against surprisal in each bin. Linear regression showed a strongly significant increase in EEG magnitude with increasing surprisal in both experiments with high levels of explained variance (nSP: $R^2 = 0.62$, $p < 0.0001$; nTP: $R^2 = 0.54$, $p < 0.0001$), showing that the neural response not only increases in magnitude at the *most* surprising moments, but increases proportionally with the level of surprisal.

We examined this linear relationship further in a similar analysis using the feature-specific surprisal (e.g., S_t^P and S_t^S). For each subject, tone epochs were binned into 128 equal-sized bins in the 2-D space spanned by surprisal along each feature, and the neural response was averaged within each bin over epochs and subjects. Fig 4-6c displays EEG magnitude for each bin at the average surprisal along each feature. Multiple linear regression shows a strongly significant correlation between EEG magnitude and surprisal in both experiments (nSP: $R^2 = 0.41$, $p < 0.0001$; nTP: $R^2 = 0.38$, $p < 0.0001$) with EEG magnitude significantly increasing with surprisal along both features (nSP: Pitch surprisal $p = 0.0124$, Spatial surprisal $p < 0.0001$; nTP: Pitch surprisal $p = 0.0272$, Timbre surprisal $p < 0.0001$).

Going beyond previous work showing linear superposition of deviance responses in oddball paradigms (such as in [94]), these results show that the neural response magnitude increases proportionally with the level of surprisal along *each* feature, which then combine linearly in the EEG response recorded at the scalp. This effect cannot be measured from stimulus properties alone nor by behavior, requiring a model to

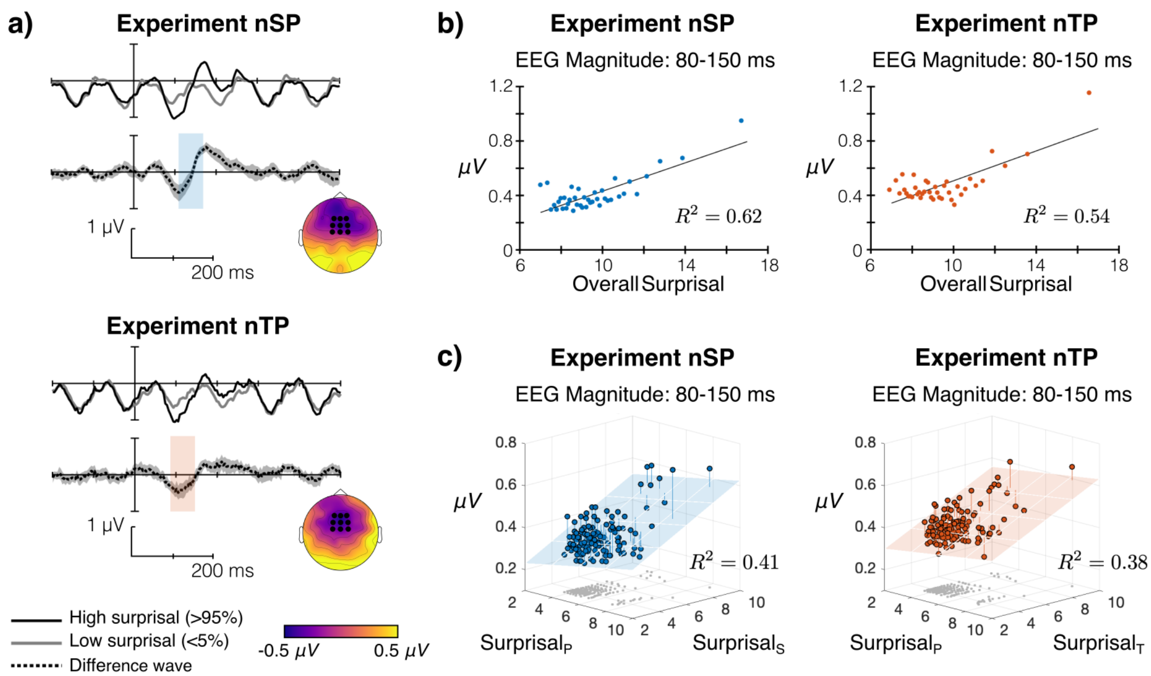

Figure 4-6. Surprisal response in experiment nSP and nTP. a) Oddball-like analysis contrasting neural response to high-surprisal tones (top 5%) with response to low-surprisal tones (bottom 5%), where overall surprisal is summed across features (e.g., $S_t^P + S_t^T$). Difference wave (high-low) shows 95% confidence interval across subjects. b) EEG magnitude (80–150 ms) in sub-averages of tone epochs binned by overall surprisal (abscissa). R^2 from linear regression. c) EEG magnitude (80–150 ms) binned by feature-specific surprisal in both features (horizontal axes). Gray points on horizontal axis show position of each point in surprisal-space. R^2 from multiple linear regression.

estimate the local surprisal of each tone along each feature given its context.

Distinct responses to local surprisal and global statistical change

We next examined neural responses aligned to high surprisal events alongside responses aligned to the maximal belief change, where the former represents local prediction mismatch and the latter represents global statistical change in the stimulus. High surprisal is again defined as tones with overall surprisal (e.g., $S_t = S_t^P + S_t^S$) in the top 5%. Maximal belief change is the moment when the belief change (δ_t) reaches its maximum across the melody trial.

Fig 4-7a (top) shows an illustration of this analysis with an example stimulus and its model outputs, surprisal S_t and belief change δ_t. Dotted lines show moments used to align epochs for each type of event. High surprisal events can occur at multiple points within the same melody stimulus, while there is only one maximal belief change. Note that when an epoch qualified as *both* high surprisal and maximal belief change, it was excluded from the high surprisal events to keep the epochs in each average response distinct. For each subject, the neural response was averaged for each aligning event (i.e., high surprisal and maximal belief change) across epochs from all melody trials.

Below the illustration, Fig 4-7a shows the grand-average neural response across subjects for all 64 channels time-locked to the two aligning events, high surprisal (left) and maximal belief change (right), in experiments nSP (top) and nTP (bottom). Topography to the right of each grand average show two responses that emerge in the highlighted time-windows after alignment: an early fronto-central negativity (FCN) with a latency of 80–150 ms (the same surprisal response examined above), and a later (and much slower) centro-parietal positivity (CPP) with a latency of 300–800 ms.

To determine whether the neural response is significantly larger in these two time windows, we compared the cumulative RMS amplitude of the neural response to baseline amplitudes in windows at the same cyclic position relative to neural entrainment (-152 to -82 ms and -630 to -130 ms for the early and late windows, respectively). In each time window, 10 channels with the largest magnitude in the grand average (5 with positive polarity, 5 with negative polarity) were selected for within-subjects analysis; selected channels for each response are highlighted in the topography in Fig 4-7a. Fig 4-7b shows dB amplitude in experiments nSP (left) and nTP (right). In both experiments, the neural response amplitude increased significantly in the early window after high surprisal tones (nSP: $t_{17} = 3.88$, $p = 0.0012$; nTP: $t_{17} = 2.45$, $p = 0.0253$) *and* after the maximal belief change (nSP: $t_{17} = 2.93$, $p = 0.0093$; nTP: $t_{17} = 4.86$, $p = 0.0001$). Note that maximal belief change often coincides with high surprisal (as illustrated in the top of Fig 4-7), so this result is not altogether "surprising". However, in the later window, the neural response only significantly increased after maximal belief change (nSP: $t_{17} = 3.02$, $p = 0.0076$; nTP: $t_{17} = 4.98$, $p = 0.0001$), with no significant increase in amplitude after other high surprisal moments in both experiments (nSP: $t_{17} = 1.05$, $p = 0.31$; nTP: $t_{17} = -0.43$, $p = 0.67$).

Finally, we examined the relationship between these effects and behavioral performance in the change detection task in experiments nSP and nTP. Fig 4-7c shows the overall d' for each subject (vertical axis) plotted against the neural response amplitude (horizontal axis) in each time window (by row) at each aligning event (by column). Linear regression analysis showed no significant correlation between neural responses and behavior in the early time window at either aligning event. At the maximal belief change, however, correlations between the neural response amplitude in the late time window (i.e., the CPP response) and behavior is significant in experiment nSP ($R^2 = 0.2$, $p = 0.036$) and marginally significant in experiment nTP ($R^2 = 0.12$, $p = 0.086$).

Together, these results suggest distinct underlying neural computations leading to the FCN and CPP effects. The FCN effect is elicited by *any* high surprisal event. Moments of maximal belief change are a subset of these events, where incoming observations no longer fit with current statistical estimates, resulting in poor predictions and higher surprisal. The surprisal response, as shown in the previous analysis, is elicited independently along each feature and combines linearly for multidimensional sounds. The CPP effect, on the other hand, occurs only at the maximal belief change, suggesting this response relates to global contextual processing after integrating non-linearly across features. Additionally, this CPP effect is weaker for poorer performing subjects, possibly reflecting individual differences in integration strategies or memory capacity for statistical estimation.

4.4 Discussion

Sound sources in natural environments vary along multiple acoustic dimensions, yet how the brain integrates these features into a coherent auditory object is an open question. Our approach combined psychophysics, computational modeling, and EEG to probe the mechanisms behind feature integration in predictive processing. Importantly, we used a stochastic change detection paradigm to approximate the challenges and uncertainty encountered in natural environments, where regularities emerge at unknown times and along unknown perceptual dimensions.

Through behavioral results, we demonstrated that listeners have access to a joint representation to perform the stochastic change detection task, flexibly combining evidence for statistical change across multiple features. To illuminate how this joint representation is constructed, we employed a computational model grounded in Bayesian accounts of statistical predictive coding in the brain [23, 37, 38, 41, 64]. This model embodies several theoretical principles of predictive processing: that the brain maps sensory inputs onto compact summary statistics [21], that the brain entertains multiple hypotheses or interpretations of sensory information [118], and that the brain incrementally updates its predictions over time based on evidence from new inputs [119]. The D-REX model and its multifeature extension presented above represent a computational instantiation of these theoretical principles which can be used to

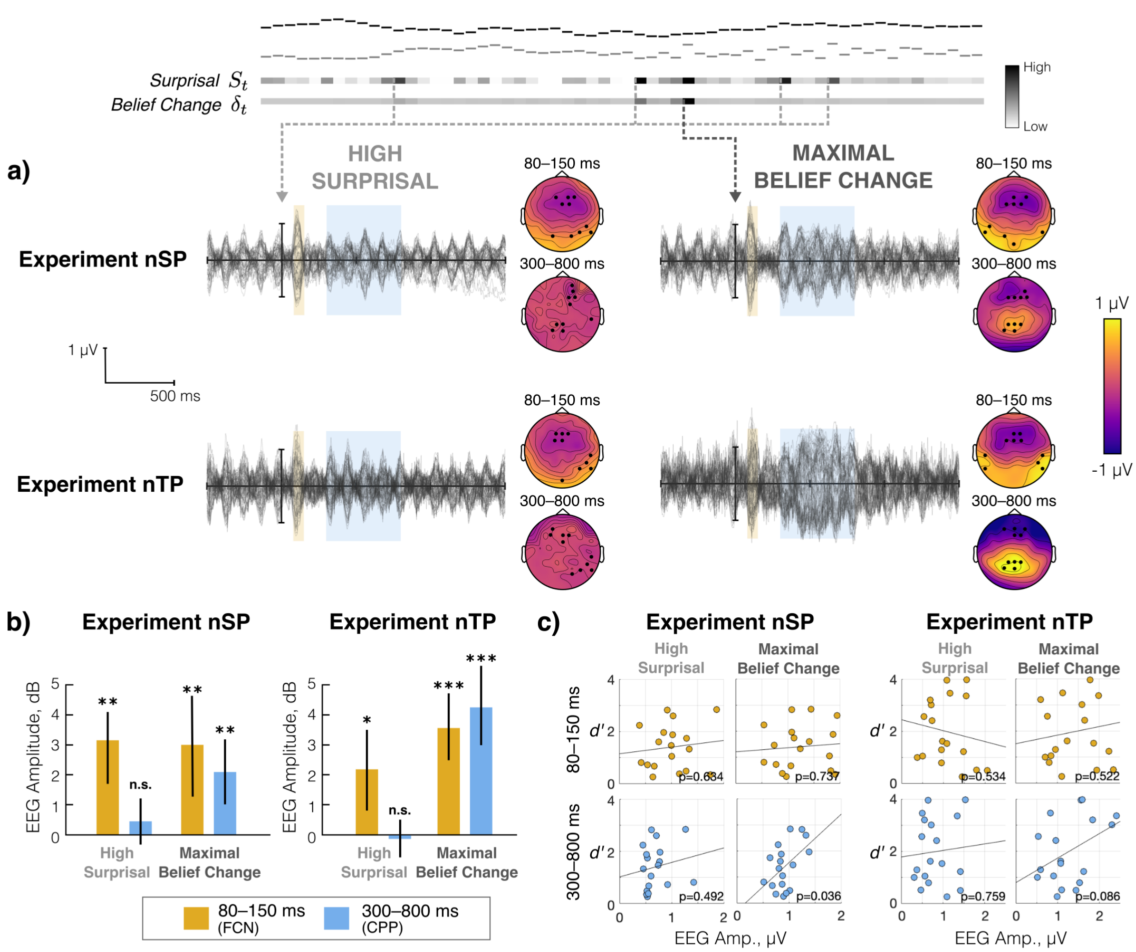

Figure 4-7. Multiplexed neural responses aligned to model outputs. Illustration of example stimulus with model outputs above: moments of high surprisal and maximal surprisal (black=high) used to align epochs for time-averaging. a) Grand-average responses for experiments nSP (top) and nTP (bottom). Shaded regions indicate two time windows of interest, with topography to the right showing average response amplitude within each time window at each channel relative to baseline. Highlighted channels used in b) for statistical analysis. b) RMS amplitude in dB relative to baseline in each time window (color) at each aligning event (horizontal axis). Error bars indicate 95% bootstrap confidence interval across subjects. c) Response amplitude in each time window at each aligning event plotted against detection performance (d') across subjects.

interpret experimental results.

We formulated multiple possible implementations for statistical prediction and integration. Using experimental data to fit these model variants to each subject, our analysis suggests that listeners independently collect higher-order statistics and infer context along multiple dimensions, integrating across dimensions at a later stage. We additionally used this "best" model to interpret variability in behavior across listeners, where detection performance ranged from near chance to near ceiling. A high degree of variability in listener behavior could be explained by the memory parameter of the model, which represents working memory capacity used to estimate statistics along each feature known to vary from person to person [51–53]. Interestingly, the fitted memory parameters correlated across features, suggesting that listeners are estimating statistics under the same neural resource constraints across dimensions. Preliminary results relating established measures of working memory capacity to statistical inference support this claim of shared memory across features (see Appendix II).

An alternative interpretation from our approach is that variability in behavior across participants is due to differing listening strategies or statistical representations ($D = 1$ or 2), particularly for lower performing subjects. Worth noting is that the same lower performing subjects ($d' < 1$) also reveal weaker centro-parietal late activity in response to maximal belief change of the melody which may underlie limited predictive tracking or sluggish cross-feature integration of statistical beliefs. The lack of any correlation between surprisal brain responses and perceptual performance (Fig. 4-7c) argues against weaker deviance tracking at the level of individual features for weaker performing subjects. In future work, these experiments could be more tailored to tease apart the source of these individual differences using the model.

It is clear from the neural responses that the brain does multiplex two types of responses that can be defined in terms of predictive processing. The fronto-central negativity (FCN) is an MMN-like response, having similar characteristics to the response to deviants in oddball experiments[14, 91, 120]. Borrowing terminology from the oddball paradigm, in our analysis we used the model to define "deviant" events in our stochastic stimuli. These high surprisal events were followed by the FCN response (changepoint or not), signifying a local, tone-level response due to mismatch between the immediate sensory input and internal predictions. Furthermore, we found that the response magnitude was proportional with surprisal in *each* feature, agreeing with similar results in the literature using less stochastic stimuli [90, 91, 95], and show evidence for linear combination across features in this early prediction-level response.

The centro-parietal positivity (CPP), on the other hand, is later, having similar latency and topography to the P3b response, which has been linked to context updating in working memory due to expectation violations [119, 121–123]. Additionally, in contrast to the MMN response, the P3b is associated with changes in global regularities encompassing higher-order statistics [124–126] and more complex stimuli [89]. Our interpretation agrees with these previous results: the CPP effect follows maximal changes in the context beliefs, the equivalent of context updating within the terminology of our model, and these shifts reflect broader changes in the statistics of the melody after integrating across features, rather than a response to a single tone or a single feature.

Finally, all of our results, from behavior to modeling to EEG, were consistent across two sets of experiments, each using a different combination of features. Where in one set of experiments (SP and nSP) the features were spectral and spatial, the second set (TP and nTP) used features that were both spectral in nature, countering the argument that these results were due to distinct what/where pathways in the brain [127]. Instead, these results support a domain-general statistical predictive coding machinery in the brain that operates in parallel along multiple perceptual features to tackle the uncertainty present in complex environments.

Chapter 5

Conclusion

Faced with the uncertainty inherent in our ever-changing surroundings, the listening brain effortlessly abstracts predictive structures embedded in sensory inputs, building an internal representation of contextual information for efficiently processing future inputs. Previous work has focused on the brain's remarkable ability to extract patterns from sounds over time, however such "template-matching" abilities have limited benefit in the dynamics of real-world environments. In this dissertation, we investigated the statistical inference processes employed in auditory perception in order to form a more complete picture of the predictive mechanisms in the brain. We use perceptual experiments to assess the statistical inference facility of human listeners employing a paradigm that mimics the complexity of real-world listening. In combination, we developed a computational model that provides a framework for understanding the intervening processing stages that connect stochastic sensory inputs to listener behavior.

Several main takeaways emerge from the behavioral results in the perceptual experiments. First, it is clear that the brain collects higher-order statistics from sounds as they evolve over time, capturing the temporal dependence between sequential sounds. In other words, the brain not only tracks the position and precision of sound sources, but also their velocity as they move through feature-space. Additionally, we confirmed this result with multiple features, where tracking occurs simultaneously along multiple dimensions. Second, the brain flexibly integrates predictive information across dimensions only when corroborative evidence exists; otherwise, the predictive processes operate independently within each dimension. This is an important skill for interpreting real-world environments, where dependencies between dimensions can change over time. Third, the integration across features occurs in posterior beliefs, rather than with the predictions themselves, aligning our results with previous findings in the literature showing independent predictive processing across features.

The model adds additional interpretive heft to our experimental results, going beyond what can be deduced from behavior alone. In all of our behavioral results, we see high variability across listeners. This variability is explained by the perceptual parameters of the model, suggesting behavioral differences can be traced to differences in the underlying perceptual fidelity and/or memory capacity of each listener.

In addition to accounting for variability across listeners, the model counters the trial-to-trial variability that unavoidably pops up in experimental paradigms involving uncertainty. Rather than using properties of the stimuli themselves for time-locked analysis of the neural response, the model provides a temporal anchor for aligning trials in terms of the underlying predictive processes. This reveals responses in the neural response that would otherwise be temporally smeared without the model. We observe multiplexed neural responses reflecting different levels of predictive processing: a local deviance response that scales with model surprisal and is elicited independently along each feature, and a global response to statistical change corresponding to belief updating in the model. These two distinct responses shed light on the internal predictive processes involved in making sense from complex, dynamic sounds.

The computational model that forms the backbone of experimental results presented in this dissertation is by no means designed to apply to these experiments alone. The D-REX model provides a general tool for studying predictive processing in audition. We demonstrate its broad applicability to modeling the statistical structures present in real-world sounds, and we showed that the same statistical processes can account for a wide range of existing results in the predictive coding literature, providing a necessary link between the controlled listening scenarios employed in perceptual research and the messy real-world scenarios they represent.

Future work

This dissertation offers a first step in understanding how the brain robustly interprets the acoustic environment, paving the way for many interesting avenues of further study.

One obvious question raised by the perceptual experiments presented here is whether attention is required for such statistical representations to form in the brain. Using the computational model and similar electroencephalography experiments with distracted listeners, we could see if the same neural signatures of statistical processing outlined above persist without attention. This would determine whether statistical representations automatically arise from bottom-up processes in the auditory hierarchy, or if they require the spotlight of attention to form a more granular representation.

The model offers a starting point to explore the role of experience in perception. As a simulated listener, the model can be used to investigate trial-to-trial learning within an experiment, with individual differences in learning rates represented by parameters in the model. The model could also be used to investigate the effects of long-term experience, such as musical experience, on statistical inference. For example, do musicians collect more complex statistical representations compared to non-musicians? Or does long-term experience modify prior expectations? Because of the modularity and generality of the model, it can be extended under the same framework to form new hypotheses for how experience is represented in long-term memory.

Finally, all experimental results presented in this dissertation involved normal-hearing listeners, but the same schemes could be used to investigate statistical inference in hearing impaired listeners or listeners with other sensory processing difficulties. This could lead to several clinical applications of this research: in diagnostics to assess the statistical inference abilities of individual listeners, in therapies to improve these abilities, or in signal processing algorithms to bootstrap the inference computations in the brain, for example, by emphasizing surprising event for hearing impaired listeners or, conversely, by dampening surprising events for listeners with sensory integration difficulties, such as individuals with Autism Spectrum Disorder.

The road to studying how perception operates “in the wild” is long, but this dissertation provides a step towards understanding the computations behind the human brain’s ability to unravel the complexity of real-world acoustic environments, and it lays the groundwork for future investigation in the perception of complex scenes.

References

1. Friston, K. J. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360,** 815–836 (2005).

2. Seriès, P. & Seitz, A. R. *Learning what to expect (in visual perception)* 2013.

3. Heilbron, M. & Chait, M. Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience* **389,** 54–73 (2018).

4. Bendixen, A. Predictability effects in auditory scene analysis: a review, 1–16 (2014).

5. Winkler, I., Denham, S. L. & Nelken, I. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences* **13,** 532–540 (2009).

6. Sams, M., Paavilainen, P., Alho, K. & Näätänen, R. Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials* **62,** 437–448 (1985).

7. Näätänen, R., Paavilainen, P., Rinne, T. & Alho, K. The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology* **118,** 2544–2590 (2007).

8. Bendixen, A., Schröger, E., Ritter, W. & Winkler, I. Regularity extraction from non-adjacent sounds. *Frontiers in Psychology* **3** (2012).

9. Saarinen, J, Paavilainen, P, Schöger, E, Tervaniemi, M & Näätänen, R. *Representation of abstract attributes of auditory stimuli in the human brain.* 1992.

10. Paavilainen, P., Degerman, A., Takegata, R. & Winkler, I. Spectral and temporal stimulus characteristics in the processing of abstract auditory features. *Neuroreport* **14,** 715–8 (2003).

11. Overath, T. *et al.* An Information Theoretic Characterisation of Auditory Encoding. *PLoS Biology* **5,** e288 (2007).

12. Yu, Y., Romero, R. & Lee, T. S. Preference of sensory neural coding for 1/f signals. *Physical Review Letters* **94,** 108103 (2005).

13. Schmuckler, M. A. & Gilden, D. L. Auditory Perception of Fractal Contours. English. *Journal of Experimental Psychology: Human Perception and Performance* **19,** 641–660 (1993).

14. Garrido, M. I., Sahani, M. & Dolan, R. J. Outlier Responses Reflect Sensitivity to Statistical Structure in the Human Brain. *PLoS Computational Biology* **9** (ed Sporns, O.) e1002999 (2013).

15. Herrmann, B., Henry, M. J., Fromboluti, E. K., McAuley, J. D. & Obleser, J. Statistical context shapes stimulus-specific adaptation in human auditory cortex. *Journal of Neurophysiology* **113,** 2582–2591 (2015).

16. Winkler, I. *et al.* The Effect of Small Variation of the Frequent Auditory Stimulus on the Event-Related Brain Potential to the Infrequent Stimulus. *Psychophysiology* **27,** 228–235 (1990).

17. Furl, N *et al.* Neural prediction of higher-order auditory sequence statistics. *NeuroImage* **54,** 2267–2277 (2011).

18. Skoe, E., Krizman, J., Spitzer, E. & Kraus, N. Prior experience biases subcortical sensitivity to sound patterns. *Journal of Cognitive Neuroscience* **27,** 124–140 (2015).

19. Agres, K., Abdallah, S. & Pearce, M. Information-Theoretic Properties of Auditory Sequences Dynamically Influence Expectation and Memory. *Cognitive Science* (2018).

20. Brady, T. F., Konkle, T. & Alvarez, G. A. Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations. *Journal of Experimental Psychology: General* (2009).

21. McDermott, J. H., Schemitsch, M. & Simoncelli, E. P. Summary statistics in auditory perception. *Nature Neuroscience* **16,** 493–498 (2013).

22. Winkler, I. & Schröger, E. Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and Language* **148,** 1–22 (2015).

23. Daunizeau, J. *et al.* Observing the Observer (II): Deciding When to Decide. *PLOS ONE* **5,** e15555–19 (2010).

24. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2009).

25. Levitin, D. J., Chordia, P & Menon, V. Musical rhythm spectra from Bach to Joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences* **109,** 3716–3720 (2012).

26. Pickover, C. A. & Khorasani, A. Fractal characterization of speech waveform graphs. *Computers and Graphics* **10,** 51–61 (1986).

27. Lieder, F, Daunizeau, J, Garrido, M. I., Friston, K. J. & Stephan, K. E. Modelling Trial-by-Trial Changes in the Mismatch Negativity. *PLoS computational biology* **9,** e1002911 (2013).

28. Wacongne, C., Changeux, J. P. & Dehaene, S. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *Journal of Neuroscience* **32,** 3665–3678 (2012).

29. Mill, R. W., Bohm, T. M., Bendixen, A, Winkler, I. & Denham, S. L. Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS computational biology* **9,** e1002925 (2013).

30. Denham, S. *et al.* Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Frontiers in Neuroscience* (2014).

31. Boubenec, Y., Lawlor, J., Górska, U., Shamma, S. & Englitz, B. Detecting changes in dynamic and complex acoustic environments. *eLife* (2017).

32. Pearce, M. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition* PhD thesis (2005), 267.

33. Hansen, N. C. & Pearce, M. T. Predictive uncertainty in auditory sequence processing. *Frontiers in psychology* **5,** 1052 (2014).

34. Di Liberto, G. M. *et al.* Cortical encoding of melodic expectations in human temporal cortex. *eLife* **9** (2020).

35. Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J. & Chait, M. Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences* **113,** E616–E625 (2016).

36. Grossberg, S., Govindarajan, K. K., Wyse, L. L. & Cohen, M. A. ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks* (2004).

37. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences* **27,** 712–719 (2004).

38. Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* **10,** 309–318 (2006).

39. Daunizeau, J. *et al.* Observing the observer (I): Meta-bayesian models of learning and decision-making. *PLoS ONE* **5,** e15554–10 (2010).

40. Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience* **30,** 12366–12378. arXiv: NIHMS150003 (2010).

41. Wilson, R. C., Nassar, M. R. & Gold, J. I. A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems. *PLoS Computational Biology* **9** (2013).

42. Adams, R. P. & MacKay, D. J. C. *Bayesian Online Changepoint Detection* tech. rep. (University of Cambridge, Cambridge, UK, 2007). arXiv: 0710.3742.

43. Murphy, K. P. Conjugate Bayesian Analysis of the Gaussian Distribution. *Def* **1,** 1–29 (2007).

44. Samson, E. Fundamental Natural Concepts of Information Theory. *ETC: A Review of General Semantics* **10,** 283–297 (1953).

45. Skerritt-Davis, B. & Elhilali, M. A Model for Statistical Regularity Extraction from Dynamic Sounds. *Acta Acustica united with Acustica* **105,** 1–4 (2019).

46. Sedley, W. *et al.* Neural signatures of perceptual inference. *eLife* (2016).

47. Kumar, S. *et al.* Resource allocation and prioritization in auditory working memory. *Cognitive Neuroscience* **4,** 12–20 (2013).

48. Arnal, L. H. & Giraud, A.-L. Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences* **16,** 390–398 (2012).

49. Skerritt-Davis, B. & Elhilali, M. Detecting change in stochastic sound sequences. *PLOS Computational Biology* **14** (ed Einhäuser, W.) e1006162 (2018).

50. Wilson, R. C. & Niv, Y. Inferring relevance in a changing world. *Frontiers in Human Neuroscience* (2012).

51. Conway, A. R. A., Cowan, N. & Bunting, M. F. The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review* **8,** 331–335 (2001).

52. Just, M. A. & Carpenter, P. A. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* **99,** 122–149 (1992).

53. Kidd, G. R., Watson, C. S. & Gygi, B. Individual differences in auditory abilities. *The Journal of the Acoustical Society of America* **122,** 418–435 (2007).

54. Wightman, F. L. & Kistler, D. J. Individual differences in human sound localization behavior. *The Journal of the Acoustical Society of America* (1996).

55. Haenschel, C. Event-Related Brain Potential Correlates of Human Auditory Sensory Memory-Trace Formation. *Journal of Neuroscience* **25,** 10494–10501 (2005).

56. Miller, T., Chen, S., Lee, W. W. & Sussman, E. S. Multitasking: Effects of processing multiple auditory feature patterns. *Psychophysiology* **52,** 1140–1148 (2015).

57. Ruhnau, P., Schröger, E. & Sussman, E. S. Implicit expectations influence target detection in children and adults. *Developmental Science* **20** (2017).

58. Naatanen, R, Gaillard, A. W. & Mantysalo, S. Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica* **42,** 313–329 (1978).

59. Sussman, E., Winkler, I., Huotilainen, M., Ritter, W. & Näätänen, R. Top-down effects can modify the initially stimulus-driven auditory organization. *Cognitive Brain Research* **13,** 393–405 (2002).

60. Friston, K. J. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11,** 127–138 (2010).

61. Bizley, J. K., Walker, K. M. M., Nodal, F. R., King, A. J. & Schnupp, J. W. H. Auditory cortex represents both pitch judgments and the corresponding acoustic cues. *Current biology : CB* **23,** 620–625 (2013).

62. Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A. & Whitney, D. Humans Use Summary Statistics to Perceive Auditory Sequences. *Psychological Science* (2013).

63. Dahmen, J. C., Keating, P., Nodal, F. R., Schulz, A. L. & King, A. J. Adaptation to stimulus statistics in the perception and neural representation of auditory space. *Neuron* **66,** 937–948 (2010).

64. Pieszek, M., Widmann, A., Gruber, T. & Schröger, E. The Human Brain Maintains Contradictory and Redundant Auditory Sensory Predictions. *PLoS ONE* **8,** e53634 (2013).

65. Lau, B., Monteiro, T. & Paton, J. J. *The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia* 2017.

66. Escera, C., Leung, S. & Grimm, S. Deviance detection based on regularity encoding along the auditory hierarchy: Electrophysiological evidence in humans. English. *Brain Topography* **27,** 527–538 (2014).

67. Garrido, M. I., Kilner, J. M., Stephan, K. E. & Friston, K. J. The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology* **120,** 453 (2009).

68. De Cheveigné, A. & Parra, L. C. Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage* **98,** 487–505 (2014).

69. Smith, N. J. & Kutas, M. Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology* **52,** 157–168 (2015).

70. Smith, N. J. & Kutas, M. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology* **52,** 169–181 (2015).

71. Tibon, R. & Levy, D. A. Striking a balance: Analyzing unbalanced event-related potential data. *Frontiers in Psychology* **6,** 1–4 (2015).

72. Lachaux, J.-P., Rodriguez, E., Martinerie, J. & Varela, F. J. Measuring phase synchrony in brains signals. *Hum Brain Mapping* **8,** 194–208 (1999).

73. Horváth, J. *et al.* MMN or no MMN: No magnitude of deviance effect on the MMN amplitude. *Psychophysiology* **45,** 60–69 (2008).

74. Symonds, R. M. *et al.* Distinguishing Neural Adaptation and Predictive Coding Hypotheses in Auditory Change Detection. *Brain Topography* **30,** 136–148 (2017).

75. Herrmann, B., Henry, M. J., Haegens, S. & Obleser, J. Temporal expectations and neural amplitude fluctuations in auditory cortex interactively influence perception. *NeuroImage* **124,** 487–497 (2016).

76. Creel, S. C., Newport, E. L. & Aslin, R. N. Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning Memory and Cognition* (2004).

77. Agus, T. R., Thorpe, S. J. & Pressnitzer, D. Rapid formation of robust auditory memories: insights from noise. *Neuron* **66,** 610–618 (2010).

78. Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A. & Bhattacharya, J. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage* (2010).

79. Krishnan, S., Carey, D., Dick, F. & Pearce, M. Effects of statistical learning in passive and active contexts on reproduction and recognition of auditory sequences. *PsyArXiv* (2019).

80. Fiser, J. & Aslin, R. N. Statistical Learning of Higher-Order Temporal Structure from Visual Shape Sequences. *Journal of Experimental Psychology: Learning Memory and Cognition* (2002).

81. Degel, J. Implicit Learning and Implicit Memory for Odors: the Influence of Odor Identification and Retention Time. *Chemical Senses* (2001).

82. Frost, R., Armstrong, B. C., Siegelman, N. & Christiansen, M. H. *Domain generality versus modality specificity: The paradox of statistical learning* 2015.

83. Conway, C. M. & Christiansen, M. H. Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology-Learning Memory and Cognition* **31,** 24–38 (2005).

84. Fiser, J. & Aslin, R. N. *Encoding multielement scenes: Statistical learning of visual feature hierarchies* 2005.

85. Gao, Z., Gao, Q., Tang, N., Shui, R. & Shen, M. Organization principles in visual working memory: Evidence from sequential stimulus display. *Cognition* (2016).

86. Baker, C. I., Olson, C. R. & Behrmann, M. Role of attention and perceptual grouping in visual statistical learning. *Psychological Science* (2004).

87. Glicksohn, A. & Cohen, A. The role of Gestalt grouping principles in visual statistical learning. *Attention, Perception, and Psychophysics* (2011).

88. Turk-Browne, N. B., Isola, P. J., Scholl, B. J. & Treat, T. A. Multidimensional Visual Statistical Learning. *Journal of Experimental Psychology: Learning Memory and Cognition* (2008).

89. Chernyshev, B. V., Bryzgalov, D. V., Lazarev, I. E. & Chernysheva, E. G. Distributed feature binding in the auditory modality. *NeuroReport* **27,** 837–842 (2016).

90. Vuust, P., Liikala, L., Näätänen, R., Brattico, P. & Brattico, E. Comprehensive auditory discrimination profiles recorded with a fast parametric musical multi-feature mismatch negativity paradigm. *Clinical Neurophysiology* **127,** 2065–2077 (2016).

91. Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M. & Näätänen, R. Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clinical Neurophysiology* **118,** 177–185 (2007).

92. Schröger, E. & Wolff, C. Mismatch response of the human brain to changes in sound location. *NeuroReport* **7,** 3005–3008 (1996).

93. Takegata, R. *et al.* Preattentive representation of feature conjunctions for concurrent spatially distributed auditory objects. *Cognitive Brain Research* **25,** 169–179 (2005).

94. Takegata, R., Paavilainen, P., Näätänen, R. & Winkler, I. Independent processing of changes in auditory single features and feature conjunctions in humans as indexed by the mismatch negativity. *Neuroscience Letters* **266,** 109–112 (1999).

95. Paavilainen, P, Valppu, S & Näätänen, R. The additivity of the auditory feature analysis in the human brain as indexed by the mismatch negativity: 1+1 approximately 2 but 1+1+1. *Neuroscience letters* **301,** 179–82 (2001).

96. Du, Y. *et al.* Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. *Cerebral Cortex* **21,** 698–707 (2011).

97. Caclin, A. *et al.* Separate Neural Processing of Timbre Dimensions in Auditory Sensory Memory. *Journal of Cognitive Neuroscience* **18,** 1959–1972 (2006).

98. Schreiner, C. E. Functional organization of the auditory cortex: maps and mechanisms. *Current Opinion in Neurobiology* **2,** 516–521 (1992).

99. Read, H. L., Winer, J. A. & Schreiner, C. E. *Functional architecture of auditory cortex* 2002.

100. Bizley, J. K., Walker, K. M. M., Silverman, B. W., King, A. J. & Schnupp, J. W. H. Interdependent Encoding of Pitch, Timbre, and Spatial Location in Auditory Cortex. *Journal of Neuroscience* **29,** 2064–2075 (2009).

101. Walker, K. M. M., Bizley, J. K., King, A. J. & Schnupp, J. W. H. Multiplexed and Robust Representations of Sound Features in Auditory Cortex. *Journal of Neuroscience* **31,** 14565–14576 (2011).

102. Allen, E. J., Burton, P. C., Olman, C. A. & Oxenham, A. J. Representations of Pitch and Timbre Variation in Human Auditory Cortex. *The Journal of Neuroscience* **37,** 1284–1293 (2017).

103. Shinn-Cunningham, B. G. *et al.* Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 13174–13178 (2008).

104. Dyson, B. J. & Ishfaq, F. Auditory memory can be object based. *Psychonomic Bulletin and Review* **15,** 409–412 (2008).

105. Thompson, W. F. & Sinclair, D. Pitch pattern, durational pattern, and timbre: A study of the perceptual integration of auditory qualities. *Psychomusicology: A Journal of Research in Music Cognition* **12,** 3–21 (1993).

106. Melara, R. D. & Marks, L. E. Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics* **48,** 169–178 (1990).

107. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cognitive psychology* **12,** 97 (1980).

108. Fetsch, C. R., Deangelis, G. C. & Angelaki, D. E. *Visual-vestibular cue integration for heading perception: Applications of optimal cue integration theory* 2010.

109. Angelaki, D. E., Gu, Y. & DeAngelis, G. C. *Multisensory integration: psychophysics, neurophysiology, and computation* 2009.

110. Parise, C. V., Spence, C. & Ernst, M. O. When correlation implies causation in multisensory integration. *Current Biology* (2012).

111. Ernst, M. O., Harrar, V., Parise, C. V. & Spence, C. *Cross-correlation between auditory and visual signals promotes multisensory integration* in *Multisensory Research* (2013).

112. Algazi, V., Duda, R., Thompson, D. & Avendano, C. The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575),* 99–102 (2001).

113. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J. M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience* **2011,** 156869. arXiv: 156869 (2011).

114. De Cheveigné, A. Sparse Time Artifact Removal. *Journal of Neuroscience Methods* **262,** 14–20 (2016).

115. Nelken, I., Rotman, Y. & Yosef, O. B. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **397,** 154–157 (1999).

116. Attias, H & Schreiner, C. E. *Temporal low-order statistics of natural sounds* in *Adv. Neural Inf. Proc. Sys. (NIPS)* (MIT Press: Cambridge, MA, 1997), 27–33.

117. Daikhin, L. & Ahissar, M. Responses to deviants are modulated by subthreshold variability of the standard. *Psychophysiology* **49,** 31–42. arXiv: NIHMS150003 (2012).

118. Mumford, D. On the computational architecture of the neocortex - I. The role of the thalamo-cortical loop. *Biological Cybernetics* (1991).

119. Darriba, A. & Waszak, F. Predictions through evidence accumulation over time. *Scientific Reports* **8,** 1–15 (2018).

120. Takegata, R., Huotilainen, M., Rinne, T., Näätänen, R. & Winkler, I. Changes in acoustic features and their conjunctions are processed by separate neuronal populations. *NeuroReport* **12,** 525–529 (2001).

121. Donchin, E. & Coles, M. G. Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences* (1988).

122. Polich, J. *Updating P300: An integrative theory of P3a and P3b* 2007. arXiv: arXiv:1011.1669v3.

123. Romero-Rivas, C. *et al.* Seeing music: The perception of melodic ‘ups and downs’ modulates the spatial processing of visual stimuli. *Neuropsychologia* (2018).

124. Chennu, S *et al.* Expectation and Attention in Hierarchical Auditory Prediction. *Journal of Neuroscience* **33,** 11194–11205 (2013).

125. Bekinschtein, T. A. *et al.* Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 1672–1677 (2009).

126. Wacongne, C. *et al.* Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* **108,** 20754–20759 (2011).

127. Murray, M. M. & Spierer, L. Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans. *Hearing Research* **258,** 121–133 (2009).

128. De Leeuw, J. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* **47,** 1–12 (2015).

Appendix I

Related publications & abstracts

Journal publications & conference proceedings

- Skerritt-Davis B, Elhilali M (under review). Neural encoding of auditory statistics.
- Skerritt-Davis B, Elhilali M (under review). Computational framework for investigating predictive processing in auditory perception.
- Kothinti SR, Skerritt-Davis B, Nair A, Elhilali M (2020). Synthesizing engaging music using dynamic models of statistical surprisal. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).
- Skerritt-Davis B, Elhilali M (2019). A model for statistical regularity extraction from dynamic sounds. Acta Acustica United with Acustica, 105(1).
- Skerritt-Davis B, Elhilali M (2018). Detecting change in stochastic sound sequences. PLoS Computational Biology, 14(5):e1006162.

Presentations & abstracts

- Skerritt-Davis B, Elhilali M (2020). Neural responses to statistical change across multiple auditory dimensions. Association for Research in Otolaryngology Mid-Winter meeting. San Jose, CA.
- Skerritt-Davis B, Elhilali M (2019). Theoretical underpinnings of statistical processing of complex sounds. International Congress on Acoustics. Aachen, Germany.
- Skerritt-Davis B, Elhilali M (2019). Modeling multi-feature regularity extraction from dynamic sounds. Association for Research in Otolaryngology Mid-Winter meeting. Baltimore, MD.
- Skerritt-Davis B, Elhilali M (2018). A model for statistical regularity extraction from dynamic sounds. International Symposium for Hearing. Snekkersten, Denmark.
- Skerritt-Davis B, Elhilali M (2018). Multi-feature regularity extraction from stochastic sequences. Association for Research in Otolaryngology Mid-Winter meeting. San Diego, CA.
- Skerritt-Davis B, Elhilali M (2017). Modeling individual differences in tracking changes in complex sound sequences. Association for Research in Otolaryngology Mid-Winter meeting. Baltimore, MD.
- Skerritt-Davis B, Elhilali M (2016). Hearing statistical regularities in fractal melodies: an EEG study. Association for Research in Otolaryngology Mid-Winter meeting. San Diego, CA.
- Skerritt-Davis B, Elhilali M (2015). Hearing statistical regularities in fractal tone sequences. Association for Research in Otolaryngology Mid-Winter meeting. Baltimore, MD.

Appendix II

Statistical inference & working memory

Introduction

In preliminary results from four experiments, we explored the relationship between statistical inference (SI) and working memory (WM) in audition. SI ability was measured with the same fractal change detection paradigm described in Chapters 3 and 4, wherein listeners are tasked with detecting statistical changes along one or two dimensions. Listeners additionally performed one of two tasks to measure WM capacity: an N-back task or a precision task. We then compared performance between the SI task and WM task within the same subjects.

In experiments 1a and 1b, listeners performed SI and WM tasks along the pitch dimension. In experiment 1a, listeners performed the N-back task to measure WM; in experiment 1b, listeners performed the precision task. In experiments 2a and 2b, listeners performed SI and WM tasks along both pitch and spatial dimensions. In experiment 2a, listeners performed the N-back task to measure WM; in experiment 2b, listeners performed the precision task.

Data for experiments 1b and 2b were collected in-person; data for experiments 1a and 2a were collected remotely over Mechanical Turk due to closures from COVID-19.

Methods

Participants

In experiment 1a, there were 166 participants; 63 participants were excluded from analysis because their performance was at or below chance ($d' <= 0$ in either task). In experiment 1b, there were 34 participants; 3 participants were excluded from analysis because of chance performance and 1 participant was excluded because of technical error in data collection. In experiment 2a, there were 104 participants; 37 participants were excluded from analysis because they had chance performance in tasks along at least one feature. In experiment 2b, there were 35 participants; no subjects were excluded from analysis.

All participants reported no history of hearing loss. Participants gave informed consent prior to the experiment and were paid for their participants. All experimental procedures were approved by the Johns Hopkins IRB.

Stimuli

SI task

Stimuli were fractal sequences of complex tones varying in pitch (experiment 1a, b) or in pitch and spatial location (experiment 2a, b). Fractal sequences were sampled power-law noise, with the power ß parameterizing entropy: as β decreases, entropy increase, with maximum entropy at $\beta = 0$ (i.e., white noise).

Each complex tone was synthesized from a harmonic stack of sinusoids with frequencies at integer multiples of the fundamental frequency, then high- and low-pass filtered at the same cutoff frequency (1200 Hz) using fourth-order Butterworth filters. Pitch was manipulated by changing the fundamental frequency [102]. For experiments 2a and b, spatial location was simulated by convolving the complex tone with interpolated head-related impulse functions for the left and right ears at the desired azimuthal position [112]. Pitches ranged from 208 to 588 Hz; spatial locations ranged from -45 to 45 degrees azimuth relative to front-center of the head.

For each stimulus sequence, the entropy changed at the midpoint of the sequence. In experiments 1a and b, the entropy changed in pitch. In experiment 2a and b, the entropy changed in pitch, in spatial location, or in both features. In all experiments, corresponding control conditions were included with no change in entropy.

In experiment 1a, entropy always began with low entropy ($\beta = 2.5$) and increased at the midpoint to one of three different levels ($\beta = 1.5$, 1, 0). The spatial location was held constant throughout at 0 degrees azimuth.

In experiment 1b, pitch entropy increased ($\beta = 2.5$ to $\beta = 0$) or decreased ($\beta = 0$ to $\beta = 2.5$) at the midpoint, with corresponding control conditions at high ($\beta = 0$) and low ($\beta = 2.5$) entropy. The spatial location was held constant throughout at 0 degrees azimuth.

In experiment 2a, entropy increased in either pitch or spatial location ($\beta = 2.5$ to $\beta = 0$), with control conditions having constant entropy at ß=2.5. In the feature that was not changing, to add small variations to the uninformative feature, the entropy was $\beta = 2.5$ but had a range equal to half of the range used in the informative feature.

In expeirment 2b, entropy increased in pitch, in spatial location, or in both features simultaneously ($\beta = 2.5$ to $\beta = 0$). In the control condition, both features had $\beta = 2.5$ for the entirety of the stimulus.

All stimulus sequences were composed of 60 complex tones with total duration of 7 seconds. Each tone had a duration of 100 ms with 10ms onset and offset ramps, and tones were presented at 8.6 Hz (116 ms inter-onset interval).

WM task

In the WM task, stimuli were comprised of complex tones synthesized similarly to the fractal experiment. In the N-back task, sequences of 30 tones were presented with 3 second inter-onset intervals. In the precision task, sequences of 1 to 3 tones were presented with 500 ms inter-onset intervals. Tones were 100 ms in duration, and pitches spanned an octave from 247.5 to 495 Hz.

Procedure

In experiments 1a and 2a, data was collected remotely via Mechanical Turk using the jsPsych javascript library [128] and custom HTML scripts. Listeners participated through their personal computers using a web browser displayed in full-screen mode, and they were instructed to use headphones. Audio playback loudness was calibrated to a comfortable level using a test stimulus, and audibility was tested prior to the beginning of the experiment by asking listeners to type a spoken number in a text box. Stimuli were synthesized at 44.1 kHz sampling and converted to MP3 format for playback.

In experiments 1b and 2b, data was collected in-person in an anechoic chamber, where listeners were seated in front of the presentation screen. Stimuli were synthesized on-the-fly at 44.1 kHz sampling rate and presented at a comfortable listening level via over-ear headphones (Sennheiser HD 595) using PsychToolbox (`psychtoolbox.org`) and custom scripts in MATLAB (The Mathworks, Natick, MA).

All experiments were split into two sections for SI and WM tasks. The order of SI and WM tasks was counterbalanced across subjects. All experiments were under 1 hour in duration. In experiments 2a and 2b, listeners performed the WM task separately for pitch and for spatial location.

SI task

In the SI task, listeners were presented with tone sequences and asked after each trial: "Did you hear a change?". Subjects responded via keyboard with "Y" and "N" keys. Prior to testing, listeners completed a series of training trials, where feedback was given after each trial. In the testing blocks, feedback was not given. Conditions were randomized in experiments 1a, 1b, and 2b. In experiment 2a, separate testing blocks were used for each feature to test detection performance in pitch and in spatial location (i.e., within each block, the change in statistics occurred in a single feature). In experiment 2b, listeners performed a single SI task, wherein the change in statistics could occur in one or both features.

WM task

In experiments 1a and 2a, the WM task was an N-back task with 1-back and 2-back blocks. In both types of blocks, listeners were presented with a sequence of 30 complex tones: in the 1-back blocks, listeners were instructed to hit the "space-bar" on the keyboard when a tone matched the previous tone; in the 2-back blocks (i.e., the task with higher load on working memory), listeners were instructed to hit the "space-bar" on the keyboard when a tone matched the tone before the previous tone. Listeners performed 3 blocks of each type interleaved, with the starting block (1-back or 2-back) counterbalanced across subjects.

In experiment 1a, listeners performed the N-back WM task in pitch. In experiment 2a, listeners performed the N-back WM task separately for complex tones varying in pitch and for noise bursts varying in spatial location.

In experiments 1b and 2b, the WM was a precision task based on [47]. In this task, listeners heard a sequence of 1 to 3 tones, and then they were asked to replicate one of the previously heard tones using a slider. The working memory load was higher for longer sequences, as listeners have to maintain all tones in the sequence in memory to successfully perform the task.

In experiment 1a, listeners performed the precision WM task in pitch. In experiment 2b, listeners performed the precision WM task separately for complex tones varying in pitch and for noise bursts varying in spatial location.

Data analysis

To determine the relationship between SI tasks and WM tasks, Spearman correlation was used to test for statistical significance in monotonicity between overall task performance. In the N-back (WM) and fractal change detection (SI) tasks, overall d' was used to measure performance, which incorporates both hit rates and false alarm rates across all conditions to measure listeners' sensitivity. In the precision (WM) task, the overall mean standard error between the target tone and the response tone was used as a measure of task performance. In experiment 2b, because the fractal change detection task was collected jointly across features, the SI task performance in the single-change conditions is used for each feature (e.g., in pitch, WM task performance for pitch is compared to SI task performance in the pitch-only change condition).

Results

Figure II-1 shows the results from experiments 1a (left) and 1b (right), where SI and WM task performance was measured in pitch. Each point corresponds to a single listener, with horizontal position indicating SI task performance and vertical position indicating WM task performance. In both experiments, there is a statistical significant correlation between both the N-back and the precision WM tasks and the fractal change detection SI task, suggesting that the SI and WM tasks are measuring the same neural mechanisms.

Figure II-2 shows the results from experiments 2a (left) and 2b (right), where SI and WM performance was measured in pitch and in spatial location. Again, each point corresponds to a single listener's performance in the SI and WM tasks. Performance is shown separately for each feature, with the top plots showing performance when pitch is varying, and the bottom plots showing performance when spatial location is varying. In both experiment and in both features, correlations in overall performance across tasks are statistical significant, again suggesting that the SI and WM tasks are measuring the same neural mechanisms.

Finally, Figure II-3 examines the relationship between features within each task in experiments 2a (left) and 2b (left). Top plots show SI task performance in both features and bottom plots show WM task performance in both features. Each point corresponds to a single subject, where the horizontal axis indicates performance in the task testing pitch, and the vertical indicates performance in the task testing spatial location. Note that in the SI task in experiment 2b (Fig II-3b, top), hit-rates are displayed for each feature, because d' was not independently measures in each feature in the joint SI task. Correlations across features suggest shared, domain-general neural resources were used to perform each task.

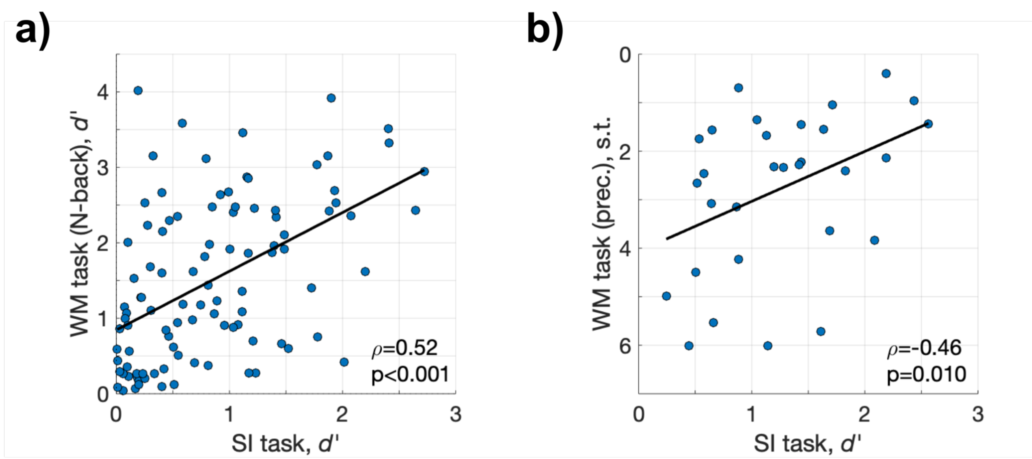

Figure II-1. Results from experiments 1a (a) and 1b (b) comparing SI (x-axis) and WM (y-axis) task performance in pitch. Spearman correlations displayed in lower right.

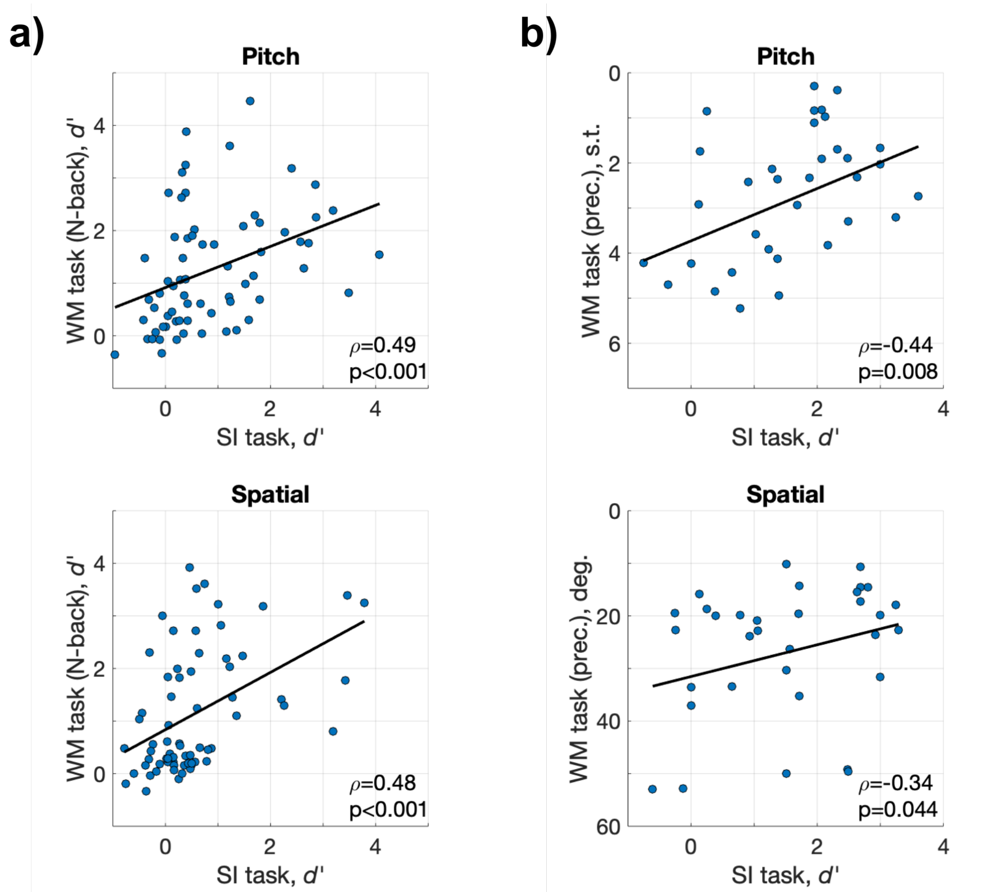

Figure II-2. Results from experiments 2a (a) and 2b (b) comparing SI (x-axis) and WM (y-axis) task performance in pitch (top) and spatial location (bottom).

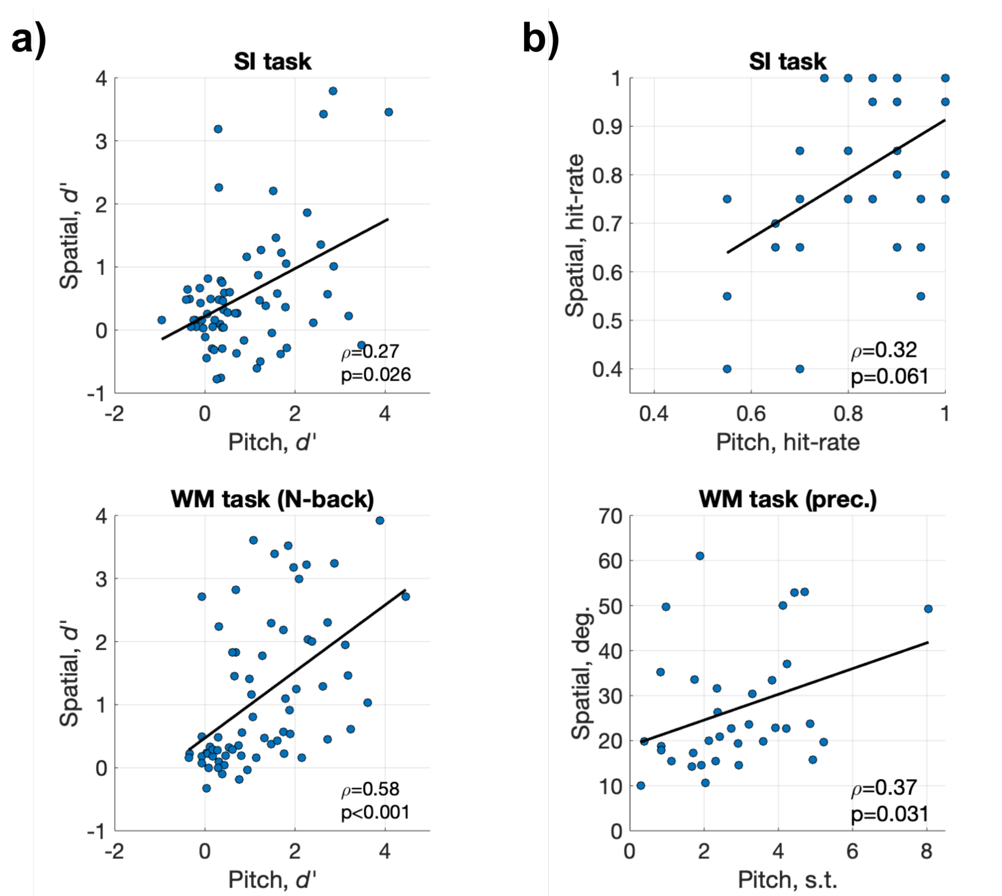

Figure II-3. Results from experiments 2a (a) and 2b (b) comparing task performance across features. Overall task performance shown for pitch (x-axis) and spatial location (y-axis). Top plots show SI task performance, bottom plots show WM task performance

Appendix III

Computer code

D-REX Model computer code downloaded from https://github.com/JHU-LCAP/DREX-model on August 11, 2020. (GitHub commit: 742a1341947cda73d6de49c2c01c1452bd13711e)

```
function [out] = run_DREX_model(x, params)
% Usage: [out] = run_DREX_model(x, params)
%
% D-REX model for Dynamic statistical REgularity eXtraction

% Assumes observations come from an underlying probabilitity distribution
% (specified in params) with unknown parameters, builds robust predictions
% by collecting sufficient statistics and calculating beliefs across
% multiple context windows causally. Distributions currently supported:
% Gaussian, Log-Normal, Gaussian Mixture Model (GMM), Poisson. Gaussian and
% Log-Normal have options temporal dependence between inputs, GMM and
% Poisson assume independent inputs.
%
% NOTE: If input has multiple features (i.e., size(x,2)>1), predictions
% along each feature are multiplied before updating beliefs.
%
% ===INPUT===
%   x        input sequence of observations (dim: time x feature)
%   params   structure with model parameters (see below for more info)
%
% ===OUTPUT===
%   out      output structure with sequential model results (see below for more info)
%
%
% * Params structure
%   distribution Distribution choice: 'gaussian','lognormal','gmm', or 'poisoon' (default='gaussian')
%   D           temporal dependence (or interval size for Poisson), integer (default=1, 50 for Poisson),
%   prior       structure with priors for sufficient statistics (see below)
%   hazard      prior probability of change, scalar (constant) or vector (time-varying) (default=0.01)
%   obsnz       observation noise for each feature, vector (default=0.0)
%   memory      maximum number of context hypotheses, integer (default=inf)
%   maxhyp      maximum number of simultaneous context hypotheses, integer (default=inf)
%
```

```
% * Priors structure, depends on distribution choice, for example for 'gaussian':
% Each field is a cell array with a cell for each feature
%   mu{f}    prior mean (size: D x 1)
%   ss{f}    prior sum of squares (size: D x D)
%   n{f}     prior observation count (size: 1 x 1)
% Note: same structure as output of function 'estimate_suffstat.m'
%
% * Output structure
%   surprisal      surprisal due to each observation in bits (dim: time x feature)
%   context_beliefs posterior beliefs for context hypotheses (dim: context-boundary x time)
%   prediction_params parameters of predictive distribution at each time (dim: time x feature)
%
% v3
% Benjamin Skerritt-Davis
% bsd@jhu.edu

[ntime, nfeature] = size(x);

if isfield(params,'changeprior')
   error('changeprior -> hazard in params')
end

% Parameters
if ~isfield(params,'distribution'), distribution = 'gaussian'; else, distribution = params.distribution; end
if ~isfield(params,'prior'), error('set prior');   else, prior = params.prior; end
if ~isfield(params,'hazard'), hazard = 0.01;       else, hazard = params.hazard; end
if ~isfield(params,'D'),    D = 1;                 else, D = params.D; end
if ~isfield(params,'obsnz'), obsnz = zeros(nfeature,1); else, obsnz = params.obsnz; end
if ~isfield(params,'memory'), memory = inf;        else, memory = params.memory; end
if ~isfield(params,'maxhyp'), maxhyp = inf;        else, maxhyp = params.maxhyp; end
if ~isfield(params,'predscale'), predscale = 1e-3; else, predscale = params.predscale; end

% check input and parameters match
if size(x,2) > size(x,1); error('input should be time x feature'); end
if size(x,1)==0 || numel(x)==0; error('input has zero length'); end
if nfeature ~= length(obsnz); error('obsnz and nfeature mismatch'); end
if ~strcmp(distribution, 'poisson') && any([prior.n{:}] < D); error('prior n''s must all be >= D'); end
if isinf(memory) || memory > ntime+1; memory = ntime+1; end
if memory < 2; error('memory must be greater than 1'); end


% Distribution-specific parameters and parameter checks
switch distribution
   case 'gmm'
      % max number of components
```

```
      if ~isfield(params,'max_ncomp'), max_ncomp = 10; else, max_ncomp = params.max_ncomp; end
      % Thresh for creating new comp. Lower threshold means new inputs
      % are more likely to be incorporated into existing components.
      if ~isfield(params,'beta'), beta = 0.001; else, beta = params.beta; end
      if D ~= 1
         error('Temporal dependence not supported. Set D=1 for GMM distribution.');
      end
   case 'poisson'
      % For Poisson distribution, D is the temporal interval into the past for counting events
      if ~isfield(params,'D'), D = 50; else, D = params.D; end
end

% If hazard rate is scalar (constant), vectorize
if numel(hazard)==1
   hazard = hazard*ones(size(x,1),1);
end


%=== INITIALIZE ==========================================

% Initialize conditioning observations for D>1
cond_obs = nan(D-1,nfeature);

% Initialize output arrays
surprisal = zeros(ntime,nfeature); % Surprisal at each time for each feature
B = zeros(memory, ntime+1); % Beliefs, or context posterior, at each time (dim: context_hypothesis x time)
B(1,1) = 1;                 % context_length=0 at time=0 (i.e., sequence begins at first observation)
prediction_theta = cell(ntime,1);


% Initialize sufficient statistics with priors
suffstat = [];
for f = 1:nfeature
   switch distribution
      case 'gaussian'
         try
            % Initialize with NaNs
            suffstat.n{f} = nan(memory,1); % obs count
            suffstat.mu{f} = nan(D,memory); % mean
            suffstat.ss{f} = nan(D,D,memory); % sum of squared deviations

            % Initialize first hypothesis with prior
            suffstat.n{f}(1) = prior.n{f};
            suffstat.mu{f}(:,1) = prior.mu{f};
            suffstat.ss{f}(:,:,1) = prior.ss{f};
         catch err
            getReport(err)
```

```
            error('Issue with prior and Gaussian sufficient statistics');
         end
      case 'lognormal'
         try
            % Initialize with NaNs
            suffstat.n{f} = nan(memory,1); % obs count
            suffstat.mu{f} = nan(D,memory); % mean
            suffstat.ss{f} = nan(D,D,memory); % sum of squared deviations

            % Initialize first hypothesis with prior
            suffstat.n{f}(1) = prior.n{f};
            suffstat.mu{f}(:,1) = prior.mu{f};
            suffstat.ss{f}(:,:,1) = prior.ss{f};
         catch err
            getReport(err)
            error('Issue with prior and Log-normal sufficient statistics');
         end
      case 'gmm'
         try
            % Initialize with NaNs
            suffstat.k{f} = nan(memory, 1); % num of components
            suffstat.n{f} = nan(memory, max_ncomp); % obs count
            suffstat.mu{f} = nan(memory, max_ncomp); % mean
            suffstat.sigma{f} = nan(memory, max_ncomp); % sum of squared deviations
            suffstat.pi{f} = zeros(memory, max_ncomp); % component weight
            suffstat.sp{f} = nan(memory, max_ncomp); % component likelihood

            % Initialize first hyp with prior
            suffstat.k{f}(1) = prior.k{f};
            suffstat.n{f}(1,:) = prior.n{f};
            suffstat.mu{f}(1,:) = prior.mu{f};
            suffstat.sigma{f}(1,:) = prior.sigma{f};
            suffstat.pi{f}(1,:) = prior.pi{f};
            suffstat.sp{f}(1,:) = prior.sp{f};
         catch err
            getReport(err)
            keyboard;
            error('Issues with prior and GMM sufficient statistics');
         end
      case 'poisson'
         try
            % Initialize with NaNs
            suffstat.n{f} = nan(memory,1); % obs count
            suffstat.lambda{f} = nan(memory,1); % mean

            % Initialize first hypothesis with prior
            suffstat.n{f}(1) = prior.n{f};
            suffstat.lambda{f}(1) = prior.lambda{f};
         catch err
            getReport(err)
            error('Issues with prior and Poisson sufficient statistics');
```

```
            end
         otherwise
            error(['Unsupported distribution: ' distribution]);
      end
   end


% ================
%    MAIN LOOP
% =================
for t = 1:ntime

   % ==== OBSERVE: new input ======================================================
   obs = x(t,:);

   % ==== PREDICT: compute context-specific predictive probs of new input =========
   switch distribution
      case 'gaussian'
         pred = predict_GAUSSIAN(obs, cond_obs, suffstat, B(:,t), D, obsnz, predscale);
      case 'lognormal'
         pred = predict_LOGNORMAL(obs, cond_obs, suffstat, B(:,t), D, obsnz, predscale);
      case 'gmm'
         pred = predict_GMM(obs, suffstat, B(:,t), obsnz, predscale);
      case 'poisson'
         pred = predict_POISSON(obs, cond_obs, suffstat, B(:,t), predscale);
      otherwise
         error(['Unsupported distribution: ' distribution]);
   end

   % Extra prediction info: expected value, error, predictive distribution
   % params (for computing full predictive distribution, \Psi)
   if isempty(pred)
      prediction_theta{t} = prediction_theta{t-1};
      pflds = fields(prediction_theta{t});
      for f = 1:length(pflds)
         prediction_theta{t}.(pflds{f})(end+1,:) = prediction_theta{t}.(pflds{f})(end,:);
      end
   else
      prediction_theta{t} = pred.ss;
   end

   % Calculate Surprisal
   if isnan(obs) % no input, no surprisal
      surprisal(t,:) = nan;
   else
      surprisal(t,:) = -1*log2(pred.prob'*B(1:min(t,memory),t));
   end

```

```
   % ==== UPDATE context-beliefs with predictive probabilities ===========
   % Combine prediction across features (i.e., probabilistic-AND across
   % features) to update context beliefs
   pp = [];
   if ~isempty(pred)
      pp = prod(pred.prob,2);
   end
   B = update_context_posterior(B, pp, hazard(t), t, maxhyp);


   % ==== UPDATE sufficient statistics with new observation ==============
   switch distribution
      case 'gaussian'
         [cond_obs, suffstat] = update_GAUSSIAN(obs, cond_obs, D, suffstat, B(:,t),
             prior, obsnz);
      case 'lognormal'
         [cond_obs, suffstat] = update_LOGNORMAL(obs, cond_obs, D, suffstat, B(:,t),
              prior, obsnz);
      case 'gmm'
         try
         suffstat = update_GMM(obs, suffstat, pred, prior, obsnz, beta*predscale);
         catch err
            getReport(err)
            keyboard;
         end
      case 'poisson'
         [cond_obs, suffstat] = update_POISSON(obs, cond_obs, suffstat, B(:,t),
             prior);
      otherwise
         error(['Unsupported distribution: ' distribution]);
   end

end


% ========= OUTPUT ==========
out.distribution = distribution;
out.surprisal = surprisal;
out.context_beliefs = B;
out.prediction_params = prediction_theta;

end


%%  *****************************************
%   |         SUB-FUNCTIONS       |
%   *****************************************

function R = update_context_posterior(R, pp, hazard, t, maxhyp)
% Update beliefs with predictive probabilities
% pp: predictive probabilities
```

```
% hazard: hazard rate
% t: current time

memory = size(R,1);

% If no prediction, change prob is 0.
if isempty(pp)
   R(1:min(t,memory-1), t+1) = R(1:min(t,memory-1), t); % Change prob
   R(min(t+1,memory),t+1) = 0; % Growth prob
   return;
end

try
   if memory <= t
      % Growth prob: P(c_t=1:t, x_t:t)
      R(1:(memory-1),t+1) = pp(2:end) .* (1-hazard) .* R(2:memory,t);
      R(1,t+1) = R(1,t+1) + pp(1) .* (1-hazard) .* R(1,t);
      % Change prob: P(c_t=0, x_1:t)
      R(memory,t+1) = sum(pp(1:end) .* hazard .* R(1:memory,t));
   else
      % Growth prob: P(c_t=1:t, x_t:t)
      R(1:t,t+1) = pp .* (1-hazard) .* R(1:t,t);
      % Change prob: P(c_t=0, x_1:t)
      R(t+1,t+1) = sum(pp .* hazard .* R(1:t,t));
   end
catch
   keyboard;
end
% Check context posterior
if any(R(:) < 0)
   disp('ERROR with context posterior');
   keyboard;
end

% prune lowest prob context hypothesis if exdeeded maxhyp
hypidx = find(R(1:min(t,memory-1),t+1) > 0);
if maxhyp < inf && length(hypidx) >= maxhyp
   [~,worsthypidx] = min(R(hypidx,t+1));
   R(hypidx(worsthypidx),t+1) = 0;
end

% Normalize posterior to sum to 1
R(:,t+1) = R(:,t+1) / sum(R(:,t+1));

end

% =====================================================================
%              DISTRIBUTION: GAUSSIAN
% =====================================================================

% ==== PREDICT for each context hypothesis ============================
```

```
function p = predict_GAUSSIAN(obs, cond_obs, suffstat, beliefs, D, obsnz, scale)
% pred: vector of predictive probabilities
% condSS: conditional sufficient statistics

% Skip prediction for any hyps with belief=0
keephyp = find(beliefs > 0);

% If silent/missing observation, no prediction to make
if any(isnan(obs) | isempty(obs))
   % NOTE: assumes observation silent/missing simultaneously for all
   % features
   p = [];
   return;
end

nhyp = sum(~isnan(suffstat.n{1})); % number of hypotheses incl. ones with belief=0
nkeephyp = length(keephyp); % number of hypotheses with belief>0
nfeature = length(suffstat.n);
pred = zeros(nkeephyp,nfeature); % predictive probabilities of new observation

% sufficient statistics
muT = suffstat.mu;
ssT = suffstat.ss;
nT = suffstat.n;

% Loop over features, calc cond distribution and predictions for each context
    hypotheses
nCond = zeros(nkeephyp,nfeature); % conditional count
muCond = zeros(nkeephyp,nfeature); % conditional mean
covCond = zeros(nkeephyp,nfeature); % conditional (co)variance

for f = 1:nfeature
   % condition current observation on past d-1 observations
   if D>1 && sum(isnan(cond_obs)) < length(cond_obs)
      for hh = 1:nkeephyp
         h = keephyp(hh);
         sigmaJoint = ssT{f}(:,:,h)*(nT{f}(h)+1)/(nT{f}(h)*(nT{f}(h)-D+1));
         muJoint = muT{f}(:,h);
         nuJoint = nT{f}(h)-D+1;

         devFromMean = cond_obs(:,f) - muJoint(1:D-1);
         % Replace NaNs with 0 to marginalize over missing context
         devFromMean(isnan(devFromMean)) = 0;

         nCond(hh,f) = nuJoint+D-1;
         z = sigmaJoint(D,1:D-1)/sigmaJoint(1:D-1,1:D-1);
         muCond(hh,f) = muJoint(D) + z*devFromMean;
         covCond(hh,f) = ((nuJoint + devFromMean'/sigmaJoint(1:D-1,1:D-1)*
             devFromMean)/nCond(hh,f))*...
            (sigmaJoint(D,D) - z*sigmaJoint(1:D-1,D));

```

```
         if any(~isreal(covCond) | ~isreal(muCond))
            warning('ERROR with predictive probabilities')
            keyboard;
         end
      end

   else % D=1, no conditioning
      for hh = 1:nkeephyp
         h = keephyp(hh);
         covCond(hh,f) = ssT{f}(1,1,h)*(nT{f}(h)+1)/(nT{f}(h)*(nT{f}(h)));
         muCond(hh,f) = muT{f}(h);
         nCond(hh,f) = nT{f}(h);
      end
   end
   % Calculate predictive probability of new observation given each hypothesis
   pred(:,f) = studentpdf(obs(f), muCond(:,f), covCond(:,f) + obsnz(f)^2, nCond(:,f)
       )*scale;
end

% Put predictions back into array with prediction=0 for belief=0 hypotheses
condSS.mu = zeros(nhyp,nfeature);
condSS.mu(keephyp,:) = muCond;
condSS.cov = zeros(nhyp,nfeature);
condSS.cov(keephyp,:) = covCond;
condSS.n = zeros(nhyp,nfeature);
condSS.n(keephyp,:) = nCond;
tmp = pred;
pred = zeros(nhyp,nfeature);
pred(keephyp,:) = tmp;

% Prob ceiling at 1 (in case of variance << 1)
if any(pred > 1)
   error('Predictive prob greater than one. Decrease predscale to combat this.');
end

% Check predictive probabilities
if any(isnan(pred) | ~isreal(pred))
   warning('ERROR with predictive probabilities')
   keyboard;
end

p = [];
p.prob = pred;
% beliefs = beliefs(1:length(condSS.mu))';
% p.expected = beliefs * condSS.mu;
% p.error = abs(p.expected - obs);
p.ss = condSS;

end

% ==== UPDATE sufficient statistics with new observation ==============
```

```
function [cond_obs, suffstat] = update_GAUSSIAN(obs, cond_obs, D, suffstat, beliefs,
     prior, obsnz)
% If prior==[], only update statistics.

nfeature = length(suffstat.n);
nhyp = sum(~isnan(suffstat.n{1}));
memory = length(suffstat.n{1});

% Skip update for any hyps with belief=0
keephyp = find(beliefs > 0);
nkeephyp = length(keephyp);

% Replace NaNs with 0s to marginalize over missing context
obs_w_context = [cond_obs; obs];
obs_w_context(isnan(obs_w_context)) = 0;

for f = 1:nfeature

   % Update statistics, unless input obs is empty/missing
   if ~any(isnan(obs) | isempty(obs))
      n_update = suffstat.n{f}(keephyp) + 1;
      mu_update = (repmat(suffstat.n{f}(keephyp),1,D)'.*suffstat.mu{f}(:,keephyp) +
          repmat(obs_w_context(:,f),1,nkeephyp))./repmat(n_update,1,D)';

      tmpcov = zeros(D,D,nkeephyp);
      for hh = 1:nkeephyp
         h = keephyp(hh);
         tmpcov(:,:,hh) = ((obs_w_context(:,f)-suffstat.mu{f}(:,h))*(obs_w_context
             (:,f)-suffstat.mu{f}(:,h))' + eye(D)*obsnz(f)^2);
      end

      suffstat.ss{f}(:,:,keephyp) = suffstat.ss{f}(:,:,keephyp) + tmpcov.*repmat(
          shiftdim(suffstat.n{f}(keephyp)./n_update,-2),D,D,1);
      suffstat.mu{f}(:,keephyp) = mu_update;
      suffstat.n{f}(keephyp) = n_update;

      % clear suffstats for hyps with beliefs=0
      suffstat.ss{f}(:,:,~ismember(1:nhyp,keephyp)) = 0;
      suffstat.mu{f}(:,~ismember(1:nhyp,keephyp)) = 0;
      suffstat.n{f}(~ismember(1:nhyp,keephyp)) = 0;
   end

   % Concatenating new hypothesis
   if ~isempty(prior)
      if nhyp < memory
         % add prior as newest hypothesis
         suffstat.n{f}(nhyp+1) = prior.n{f};
         suffstat.mu{f}(:,nhyp+1) = prior.mu{f};
         suffstat.ss{f}(:,:,nhyp+1) = prior.ss{f};
      else
         % remove oldest hypothesis and add prior as newest hypothesis
```

```
         suffstat.n{f} = cat(1,suffstat.n{f}(2:end),prior.n{f});
         suffstat.mu{f} = cat(2,suffstat.mu{f}(:,2:end), prior.mu{f});
         suffstat.ss{f} = cat(3,suffstat.ss{f}(:,:,2:end), prior.ss{f});
      end
   end
end


% increment conditioning observations to include new observation
cond_obs = [cond_obs; obs];
cond_obs(1,:) = [];

end


% =====================================================================
%               DISTRIBUTION: LOG-NORMAL
% =====================================================================

% ==== PREDICT for each context hypothesis ============================
function p = predict_LOGNORMAL(obs, cond_obs, suffstat, beliefs, D, obsnz, scale)
% pred: vector of predictive probabilities
% condSS: conditional sufficient statistics

% Take log of new observation and context
obs = log(obs);
cond_obs = log(cond_obs);

% Skip prediction for any hyps with belief=0
keephyp = find(beliefs > 0);

% If silent/missing observation, no prediction to make
if any(isnan(obs) | isempty(obs))
   % NOTE: assumes observation silent/missing simultaneously for all
   % features
   p = [];
   return;
end

nhyp = sum(~isnan(suffstat.n{1})); % number of hypotheses incl. ones with belief=0
nkeephyp = length(keephyp); % number of hypotheses with belief>0
nfeature = length(suffstat.n);
predprobs = zeros(nkeephyp,nfeature); % predictive probabilities of new observation

% sufficient statistics
muT = suffstat.mu;
ssT = suffstat.ss;
nT = suffstat.n;

% Loop over features, calc cond distribution and predictions for each context
    hypotheses
```

```
nCond = zeros(nkeephyp,nfeature); % conditional count
muCond = zeros(nkeephyp,nfeature); % conditional mean
covCond = zeros(nkeephyp,nfeature); % conditional (co)variance

for f = 1:nfeature
   % condition current observation on past d-1 observations
   if D>1 && sum(isnan(cond_obs)) < length(cond_obs)
      for hh = 1:nkeephyp
         h = keephyp(hh);
         sigmaJoint = ssT{f}(:,:,h)*(nT{f}(h)+1)/(nT{f}(h)*(nT{f}(h)-D+1));
         muJoint = muT{f}(:,h);
         nuJoint = nT{f}(h)-D+1;

         devFromMean = cond_obs(:,f) - muJoint(1:D-1);
         % Replace NaNs with 0 to marginalize over missing context
         devFromMean(isnan(devFromMean)) = 0;

         nCond(hh,f) = nuJoint+D-1;
         z = sigmaJoint(D,1:D-1)/sigmaJoint(1:D-1,1:D-1);
         muCond(hh,f) = muJoint(D) + z*devFromMean;
         covCond(hh,f) = ((nuJoint + devFromMean'/sigmaJoint(1:D-1,1:D-1)*
             devFromMean)/nCond(hh,f))*...
            (sigmaJoint(D,D) - z*sigmaJoint(1:D-1,D));
      end

   else % D=1, no conditioning
      for hh = 1:nkeephyp
         h = keephyp(hh);
         covCond(hh,f) = ssT{f}(1,1,h)*(nT{f}(h)+1)/(nT{f}(h)*(nT{f}(h)));
         muCond(hh,f) = muT{f}(h);
         nCond(hh,f) = nT{f}(h);
      end
   end
   % Calculate predictive probability of new observation given each hypothesis
   predprobs(:,f) = studentpdf(obs(f), muCond(:,f), covCond(:,f) + obsnz(f)^2, nCond
       (:,f)) * scale;
end

% Put predictions back into array with prediction=0 for belief=0 hypotheses
condSS.mu = zeros(nhyp,nfeature);
condSS.mu(keephyp,:) = muCond;
condSS.cov = zeros(nhyp,nfeature);
condSS.cov(keephyp,:) = covCond;
condSS.n = zeros(nhyp,nfeature);
condSS.n(keephyp,:) = nCond;
tmp = predprobs;
predprobs = zeros(nhyp,nfeature);
predprobs(keephyp,:) = tmp;


% Prob ceiling at 1 (in case of variance << 1)
```

```
if any(predprobs > 1)
   error('Predictive prob greater than one. Decrease predscale to combat this.');
end

% Check predictive probabilities
if any(isnan(predprobs) | ~isreal(predprobs))
   warning('ERROR with predictive probabilities')
   keyboard;
end

p = [];
p.prob = predprobs;
% p.expected = beliefs(1:length(condSS.mu))' * exp(condSS.mu+0.5*condSS.cov);
% p.error = abs(exp(p.expected) - exp(obs));
p.ss = condSS;
end

% ==== UPDATE sufficient statistics with new observation ==============
function [cond_obs, suffstat] = update_LOGNORMAL(obs, cond_obs, D, suffstat, beliefs,
     prior, obsnz)
% If prior==[], only update statistics.

% Take log of new observation and context
origobs = obs;
origcontext = cond_obs;
obs = log(obs);
cond_obs = log(cond_obs);


nfeature = length(suffstat.n);
nhyp = sum(~isnan(suffstat.n{1}));
memory = length(suffstat.n{1});

% Skip update for any hyps with belief=0
keephyp = find(beliefs > 0);
nkeephyp = length(keephyp);

% Replace NaNs with 0s to marginalize over missing context
obs_w_context = [cond_obs; obs];
obs_w_context(isnan(obs_w_context)) = 0;

for f = 1:nfeature

   % Update statistics, unless input obs is empty/missing
   if ~any(isnan(obs) | isempty(obs))
      n_update = suffstat.n{f}(keephyp) + 1;
      mu_update = (repmat(suffstat.n{f}(keephyp),1,D)'.*suffstat.mu{f}(:,keephyp) +
          repmat(obs_w_context(:,f),1,nkeephyp))./repmat(n_update,1,D)';

      tmpcov = zeros(D,D,nkeephyp);
      for hh = 1:nkeephyp
```

```
            h = keephyp(hh);
            tmpcov(:,:,hh) = ((obs_w_context(:,f)−suffstat.mu{f}(:,h))∗(obs_w_context
                (:,f)−suffstat.mu{f}(:,h))' + eye(D)∗obsnz(f)^2);
         end

         suffstat.ss{f}(:,:,keephyp) = suffstat.ss{f}(:,:,keephyp) + tmpcov.∗repmat(
             shiftdim(suffstat.n{f}(keephyp)./n_update,−2),D,D,1);
         suffstat.mu{f}(:,keephyp) = mu_update;
         suffstat.n{f}(keephyp) = n_update;

         % clear suffstats for hyps with beliefs=0
         suffstat.ss{f}(:,:,~ismember(1:nhyp,keephyp)) = 0;
         suffstat.mu{f}(:,~ismember(1:nhyp,keephyp)) = 0;
         suffstat.n{f}(~ismember(1:nhyp,keephyp)) = 0;
      end

      % Concatenating new hypothesis
      if ~isempty(prior)
         if nhyp < memory
            % add prior as newest hypothesis
            suffstat.n{f}(nhyp+1) = prior.n{f};
            suffstat.mu{f}(:,nhyp+1) = prior.mu{f};
            suffstat.ss{f}(:,:,nhyp+1) = prior.ss{f};
         else
            % remove oldest hypothesis and add prior as newest hypothesis
            suffstat.n{f} = cat(1,suffstat.n{f}(2:end),prior.n{f});
            suffstat.mu{f} = cat(2,suffstat.mu{f}(:,2:end), prior.mu{f});
            suffstat.ss{f} = cat(3,suffstat.ss{f}(:,:,2:end), prior.ss{f});
         end
      end
   end

   % increment context to include new observation
   cond_obs = [origcontext; origobs];
   cond_obs(1,:) = [];

   end

   % ===================================================================
   %           DISTRIBUTION: GAUSSIAN MIXTURE MODEL (GMM)
   % ===================================================================

   % ==== PREDICT for each context hypothesis ============================
   function p = predict_GMM(obs, suffstat, beliefs, obsnz, scale)
   % pred: vector of predictive probabilities
   % condSS: conditional sufficient statistics

   % Skip prediction for any hyps with belief=0
   keephyp = find(beliefs > 0);

   % If silent/missing observation, no prediction to make
```

```
if any(isnan(obs) | isempty(obs))
   % NOTE: assumes observation silent/missing simultaneously for all
   % features
   p = [];
   return;
end

nhyp = sum(~isnan(suffstat.k{1})); % number of hypotheses incl. ones with belief=0
nkeephyp = length(keephyp); % number of hypotheses with belief>0
nfeature = length(suffstat.n);
component_probs = cell(nfeature,1);
predprobs = zeros(nkeephyp,nfeature); % predictive probabilities of new observation

% sufficient statistics
muT = suffstat.mu;
sigmaT = suffstat.sigma;
spT = suffstat.sp;
piT = suffstat.pi;

for f = 1:nfeature
   component_probs{f} = studentpdf(obs(f), muT{f}(keephyp,:), sigmaT{f}(keephyp,:)+
       obsnz(f)^2, spT{f}(keephyp,:)) * scale; % dim: hypothesis x component
   predprobs(:,f) = sum(component_probs{f} .* piT{f}(keephyp,:),2,'omitnan');
end

% Put predictions back into array with prediction=0 for belief=0 hypotheses
tmp = predprobs;
predprobs = zeros(nhyp,nfeature);
predprobs(keephyp,:) = tmp;

tmp = component_probs;
for f = 1:nfeature
   component_probs{f} = zeros(nhyp,size(tmp{f},2));
   component_probs{f}(keephyp,:) = tmp{f};
end

% Prob ceiling at 1 (in case of variance << 1)
if any(predprobs > 1)
   error('A predictive prob is greater than one. Decrease predscale to combat this.'
       );
end

% Check predictive probabilities
if any(isnan(predprobs) | ~isreal(predprobs))
   warning('ERROR with predictive probabilities')
   keyboard;
end

p = [];
p.prob = predprobs;
p.component_probs = component_probs;
```

```
% p.expected = 0; %beliefs(1:length(condSS.mu))' * condSS.mu;
% p.error = 0; %abs(p.expected - obs);
p.ss = [];
flds = fields(suffstat);
for f = 1:nfeature
   for fld = 1:length(flds)
      p.ss.(flds{fld}){f} = suffstat.(flds{fld}){f}(1:nhyp,:);
   end
end

end

% ==== UPDATE sufficient statistics with new observation ==============
function suffstat = update_GMM(obs, suffstat, pred, prior, obsnz, beta)
% If prior==[], only update statistics.


nfeature = length(suffstat.n);
memory = length(suffstat.n{1});

max_comp = size(suffstat.mu{1},2);

% TODO: Replace NaNs with 0s to marginalize over missing context

for f = 1:nfeature

   % Update statistics, unless input obs is empty/missing
   if ~any(isnan(obs) | isempty(obs))

      % Create new component
      nhyp = size(pred.prob,1);
      try
         create_comp = (max(pred.component_probs{f},[],2,'omitnan') < beta) & (
             suffstat.k{f}(1:nhyp) < max_comp);
      catch
         keyboard;
      end
      % Update existing component components
      % Calculate component likelihood given current observation
      lik = suffstat.pi{f}(1:nhyp,:) .* pred.component_probs{f};
      lik = lik ./ repmat(sum(lik,2,'omitnan'),1,size(lik,2));
      for h = 1:nhyp
         kh = suffstat.k{f}(h); % num of comps for current hypothesis
         if create_comp(h)
            % obs comes from new component with prob 1
            lik(h,:) = 0;
            lik(h,kh+1) = 1;
            suffstat.sp{f}(h,kh+1) = 0;
            suffstat.n{f}(h,kh+1) = 0;
            suffstat.mu{f}(h,kh+1) = obs(f);
            suffstat.sigma{f}(h,kh+1) = prior.sigma{f}(1);
```

```
        end
    end

    % Update likelihood accumulatos and priors
    sp_update = suffstat.sp{f}(1:nhyp,:) + lik;
    w = lik ./ sp_update; % updated weights for each component

    % Update component means
    mu_update = suffstat.mu{f}(1:nhyp,:) + w.*(obs(f) - suffstat.mu{f}(1:nhyp,:));

    % Update component variance
    sigma_update = suffstat.sigma{f}(1:nhyp,:) + w.*((obs(f) - suffstat.mu{f}(1:nhyp,:)).*(obs(f)-mu_update) + obsnz(f)^2 - suffstat.sigma{f}(1:nhyp,:));

    % Update component obs count
    n_update = suffstat.n{f}(1:nhyp,:) + 1;

    % Reset suff stats for new components
    k_update = suffstat.k{f}(1:nhyp)+create_comp;
    mu_update(create_comp, k_update(create_comp)) = obs(f);
    sigma_update(create_comp, k_update(create_comp)) = prior.sigma{f}(1);

    % Update component priors
    pi_update = sp_update ./ repmat(sum(sp_update,2,'omitnan'),1,size(sp_update,2));


    suffstat.k{f}(1:nhyp) = k_update;
    suffstat.n{f}(1:nhyp,:) = n_update;
    suffstat.mu{f}(1:nhyp,:) = mu_update;
    suffstat.sigma{f}(1:nhyp,:) = sigma_update;
    suffstat.pi{f}(1:nhyp,:) = pi_update;
    suffstat.sp{f}(1:nhyp,:) = sp_update;


    % Concatenating new hypothesis
    if ~isempty(prior)

        if nhyp == memory
            % remove oldest hypothesis
            suffstat.k{f} = suffstat.k{f}(2:end);
            suffstat.n{f} = suffstat.n{f}(2:end,:);
            suffstat.mu{f} = suffstat.mu{f}(2:end,:);
            suffstat.sigma{f} = suffstat.sigma{f}(2:end,:);
            suffstat.pi{f} = suffstat.pi{f}(2:end,:);
            suffstat.sp{f} = suffstat.sp{f}(2:end,:);

            nhyp = memory - 1;
        end


```

```
        % add prior as newest hypothesis
        suffstat.k{f}(nhyp+1) = prior.k{f};
        suffstat.n{f}(nhyp+1,:) = prior.n{f};
        suffstat.mu{f}(nhyp+1,:) = prior.mu{f};
        suffstat.sigma{f}(nhyp+1,:) = prior.sigma{f};
        suffstat.pi{f}(nhyp+1,:) = prior.pi{f};
        suffstat.sp{f}(nhyp+1,:) = prior.sp{f};

      end
   end
end

end


% =====================================================================
%           DISTRIBUTION: POISSON
% =====================================================================

% ==== PREDICT for each context hypothesis ============================
function p = predict_POISSON(obs, cond_obs, suffstat, beliefs, scale)
% pred: vector of predictive probabilities
% condSS: conditional sufficient statistics

% Skip prediction for any hyps with belief=0
keephyp = find(beliefs > 0);

% If silent/missing observation, no prediction to make
if any(isnan(obs) | isempty(obs))
   % NOTE: assumes observation silent/missing simultaneously for all
   % features
   p = [];
   return;
end

input = sum([cond_obs; obs],'omitnan');

nhyp = sum(~isnan(suffstat.n{1})); % number of hypotheses incl. ones with belief=0
nkeephyp = length(keephyp); % number of hypotheses with belief>0
nfeature = length(suffstat.n);
pred = zeros(nkeephyp,nfeature); % predictive probabilities of new observation

% sufficient statistics
lambdaT = suffstat.lambda;
nT = suffstat.n;

% Loop over features, calc cond distribution and predictions for each context
    hypotheses
nCond = zeros(nkeephyp,nfeature); % conditional count
lambdaCond = zeros(nkeephyp,nfeature); % conditional mean

```

```

% Calculate predictive probability of new observation given each hypothesis
for f = 1:nfeature
   for hh = 1:nkeephyp
      h = keephyp(hh);
      lambdaCond(hh,f) = lambdaT{f}(h);
      nCond(hh,f) = nT{f}(h);
   end

   pred(:,f) = poissonpdf(input(f), lambdaCond(:,f))*scale;
end

% Put predictions back into array with prediction=0 for belief=0 hypotheses
condSS.lambda = zeros(nhyp,nfeature);
condSS.lambda(keephyp,:) = lambdaCond;
condSS.n = zeros(nhyp,nfeature);
condSS.n(keephyp,:) = nCond;
tmp = pred;
pred = zeros(nhyp,nfeature);
pred(keephyp,:) = tmp;

% Prob ceiling at 1 (in case of variance << 1)
if any(pred > 1)
   error('A predictive prob is greater than one. Decrease predscale to combat this.'
       );
end

% Check predictive probabilities
if any(isnan(pred) | ~isreal(pred))
   warning('ERROR with predictive probabilities')
   keyboard;
end

p = [];
p.prob = pred;
beliefs = beliefs(1:length(condSS.lambda))';
% p.expected = beliefs * condSS.lambda;
% p.error = abs(p.expected - obs);
p.ss = condSS;

end


% ==== UPDATE sufficient statistics with new observation ==============
function [cond_obs, suffstat] = update_POISSON(obs, cond_obs, suffstat, beliefs,
    prior)
% If prior==[], only update statistics.

nfeature = length(suffstat.n);
nhyp = sum(~isnan(suffstat.n{1}));
memory = length(suffstat.n{1});
```

```
% Skip update for any hyps with belief=0
keephyp = find(beliefs > 0);
nkeephyp = length(keephyp);

% Replace NaNs with 0s to marginalize over missing context
obs_w_context = [cond_obs; obs];
obs_w_context(isnan(obs_w_context)) = 0;

for f = 1:nfeature

   % Update statistics, unless input obs is empty/missing
   if ~any(isnan(obs) | isempty(obs))

      new_lambda = sum(obs_w_context(:,f));

      n_update = suffstat.n{f}(keephyp) + 1;
      lambda_update = (suffstat.n{f}(keephyp).*suffstat.lambda{f}(keephyp) + repmat(
          new_lambda,nkeephyp,1))./n_update;

      suffstat.lambda{f}(keephyp) = lambda_update;
      suffstat.n{f}(keephyp) = n_update;

      % clear suffstats for hyps with beliefs=0
      suffstat.lambda{f}(~ismember(1:nhyp,keephyp)) = 0;
      suffstat.n{f}(~ismember(1:nhyp,keephyp)) = 0;
   end

   % Concatenating new hypothesis
   if ~isempty(prior)
      if nhyp < memory
         % add prior as newest hypothesis
         suffstat.n{f}(nhyp+1) = prior.n{f};
         suffstat.lambda{f}(nhyp+1) = prior.lambda{f};
      else
         % remove oldest hypothesis and add prior as newest hypothesis
         suffstat.n{f} = cat(1,suffstat.n{f}(2:end),prior.n{f});
         suffstat.lambda{f} = cat(2,suffstat.lambda{f}(2:end), prior.lambda{f});
      end
   end
end

% increment context to include new observation
cond_obs = [cond_obs; obs];
cond_obs(1,:) = [];

end


% ====== PDF functions ============================================
function p = studentpdf(x, mu, var, n)
```

```
c = exp(gammaln(n/2 + 0.5) - gammaln(n/2)) .* (n.*pi.*var).^(-0.5);
p = c .* (1 + (1./(n.*var)).*(x-mu).^2).^(-(n+1)/2);
end

function p = poissonpdf(x, lambda)
if abs(x - round(x)) > 1e-1
   error('Poisson PDF input x must be an integer.');
else
   x = round(x);
end
p = ((lambda.^x) / factorial(x)) .* exp(-lambda);
end
```

Vita

Benjamin M. Skerritt-Davis received a bachelors degree from Brown University in Math–Physics in 2009. He enrolled in the PhD program in the Department of Electrical and Computer Engineering at Johns Hopkins University in 2013, and received a Master of Science in Engineering degree from Johns Hopkins University in Electrical and Computer Engineering in 2015. His research uses experimental and computational techniques to investigate how human auditory perception operates in the presence of uncertainty.