Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational framework for investigating predictive processing in auditory perception

Benjamin Skerritt-Davis, Mounya Elhilali *

Johns Hopkins University, 3400 N Charles St, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Predictive coding
Statistical inference
Bayesian inference
uncertainty
neural decoding

ABSTRACT

Background: The brain tracks sound sources as they evolve in time, collecting contextual information to predict future sensory inputs. Previous work in predictive coding typically focuses on the perception of predictable stimuli, leaving the implementation of these same neural processes in more complex, real-world environments containing randomness and uncertainty up for debate.

New Method: To facilitate investigation into the perception of less tightly-controlled listening scenarios, we present a computational model as a tool to ask targeted questions about the underlying predictive processes that connect complex sensory inputs to listener behavior and neural responses. In the modeling framework, observed sound features (e.g. pitch) are tracked sequentially using Bayesian inference. Sufficient statistics are inferred from past observations at multiple time scales and used to make predictions about future observation while tracking the statistical structure of the sensory input.

Results: Facets of the model are discussed in terms of their application to perceptual research, and examples taken from real-world audio demonstrate the model's flexibility to capture a variety of statistical structures along various perceptual dimensions.

Comparison with Existing Methods: Previous models are often targeted toward interpreting a particular experimental paradigm (e.g., oddball paradigm), perceptual dimension (e.g., pitch processing), or task (e.g., speech segregation), thus limiting their ability to generalize to other domains. The presented model is designed as a flexible and practical tool for broad application.

Conclusion: The model is presented as a general framework for generating new hypotheses and guiding investigation into the neural processes underlying predictive coding of complex scenes.

1. Introduction

Sound is by nature a temporal signal, unfolding as a series of acoustic events: the patter of footsteps on a city street, the sequence of phonemes in speech, the progression of individual notes or chords in music. Predictive coding theory offers an explanation for how the brain processes such sequential inputs. 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 (Karl and Friston, 2005; Seriès and Seitz, 2013; Heilbron and Chait, 2018). Despite its widespread adoption, there remain many long-standing open questions about how predictive coding is implemented, such as the nature of the brain's internal representation and how it balances stability with flexibility in the face of change and uncertainty (Denham and Winkler, 2020; Clark, 2013; Grossberg, 1980). These questions become particularly salient when considering how predictive coding operates in complex, real-world environments. Here, we propose a computational model that can serve as a tool to guide future investigation into how predictive coding theory manifests in the perception of everyday scenes.

Computational modeling has been used previously to expand the realm of investigation in predictive coding in the brain. It has facilitated the interpretation of trial-by-trial variability in listener responses (Lieder et al., 2013), the link between individual spiking neurons and neural responses to deviance measured at the scalp (Wacongne et al., 2012), and the recasting of various listening phenomena, such as streaming and object perception, in terms of predictive coding (Denham et al., 2014; Winkler and Schröger, 2015). While some models focus on the perception of deterministic sensory inputs (McDermott et al., 2011; Mill et al., 2013), computational modeling is particularly useful for studying statistical processing in the brain, where stimulus-driven analyses are often

* Corresponding author.
E-mail address: mounya@jhu.edu (M. Elhilali).

https://doi.org/10.1016/j.jneumeth.2021.109177
Received 10 August 2020; Received in revised form 7 March 2021; Accepted 25 March 2021
Available online 9 April 2021

constrained by the uncertainty in stochastic stimuli and their elicited responses (Garrido et al., 2013; Herrmann et al., 2015; Boubenec et al., 2017). However, a common limitation of these models is that they are often designed for a particular experimental paradigm (e.g., the oddball paradigm) (Lieder et al., 2013; Mill et al., 2013; Barniv and Nelken, 2015), a particular perceptual dimension (e.g., pitch) (Balaguer-Ballester et al., 2009; Tabas et al., 2019), or a particular perceptual task (e.g., speech segregation) (Nix and Hohmann, 2007), thus limiting their ability to generalize to other domains. Some notable exceptions are the IDyOM model, initially formulated for musical expectation (Pearce, 2005), which has been used to decode neural responses to music stimuli (Hansen and Pearce, 2014; Di Liberto et al., 2020) as well as describe statistical learning of sound sequences in general (Agres et al., 2018; Barascud et al., 2016). Additionally, the ARTSTREAM model, based on Gestalt principles of perception, incorporates predictive coding into a broader framework for auditory scene analysis (Grossberg et al., 2004). 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 work, the computational model put forth provides a potential algorithmic solution for the predictive processes employed in everyday listening. It makes minimal assumptions on the sensory input, instead offering a framework to compare different internal representations in the brain. This model is grounded in theoretical accounts of predictive coding based in Bayesian inference (Knill and Pouget, 2004; Tenenbaum et al., 2006; Daunizeau et al., 2010), and it incorporates key principles of statistical tracking (e.g. predict, observe, update) within a compact formulation. The same mathematical underpinnings have previously been explored in predictive-inference tasks using sequences of numbers (Nassar et al., 2010; Wilson et al., 2013). In lieu of modeling neural mechanisms directly (such as in Wacongne et al., 2012; Balaguer-Ballester et al., 2009; Tabas et al., 2019), 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 behavioral responses in a broad range of experimental studies and realistic listening environments.

We present this modeling framework in its general form for practical application in the study of statistical inference in predictive processing in audition. Previously, we have shown how a specific implementation of this model can replicate various results from controlled psychoacoustic experiments in predictive processing of pitch under a single statistical assumption (Skerritt-Davis and Elhilali, 2019). Here, we demonstrate the flexibility of the model for predictive processing of natural sounds using different statistics along a variety of input dimensions. In contrast to existing models with limited application to real-world sounds, this model can provide a deeper understanding of the computational mechanisms behind predictive tracking of rich, dynamic sounds by guiding interpretation of experimental results under a unified framework, generating new hypotheses and predictions for future investigation, and pushing the boundary of what is considered feasible for study in the laboratory towards the complexity that we encounter in everyday listening.

2. D-REX Model

The Dynamic Regularity Extraction (D-REX) model is a computational model for predictive processing of sequential sounds. Source code is available at: http://www.github.com/jhu-lcap/DREX-model.

2.1. Model assumptions

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

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

where the input observations $\{x_t\}_{t\in\mathbb{Z}^+}$ are continuous-valued and sampled discretely in time, and the notation $x_{1:t}$ refers to the observed sequence of observations from time 1 to time t: $x_{1:t} = \{x_1, x_2, \ldots, x_t\}$. The observed inputs $\{x_t\}$ 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. To maintain generality in this section, input observations x_t are presented with arbitary units at discrete times equivalent to their sequential indices (i.e., $t = 1, 2, 3, \ldots$). In Section 3, we will present specific examples of x_t from real-world sounds sampled in continuous time along various acoustic and perceptual dimensions.

The input sequence is assumed to be stochastic, drawn from a parametric 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. 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. 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 number of successive observations assumed to be statistically dependent in the input sequence. When the input observations have a constant sampling rate, D can equivalently be described as the temporal extent of 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 $D = 2$ is sensitive to dependence (i.e., non-zero covariance) between adjacent observations, while with $D = 1$, observations are assumed to be statistically independent. As D increases, the model can capture temporal dependence across wider spans of the input observations if it exists.

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.

2.2. Robust prediction of dynamic observations

Under these assumptions, the challenge for the model is to make predictions that are robust both to the unknown dynamics in the underlying generating distribution and to the uncertainty stemming from stochastic inputs.

2.2.1. Sufficient statistics $\widehat{\theta}$

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

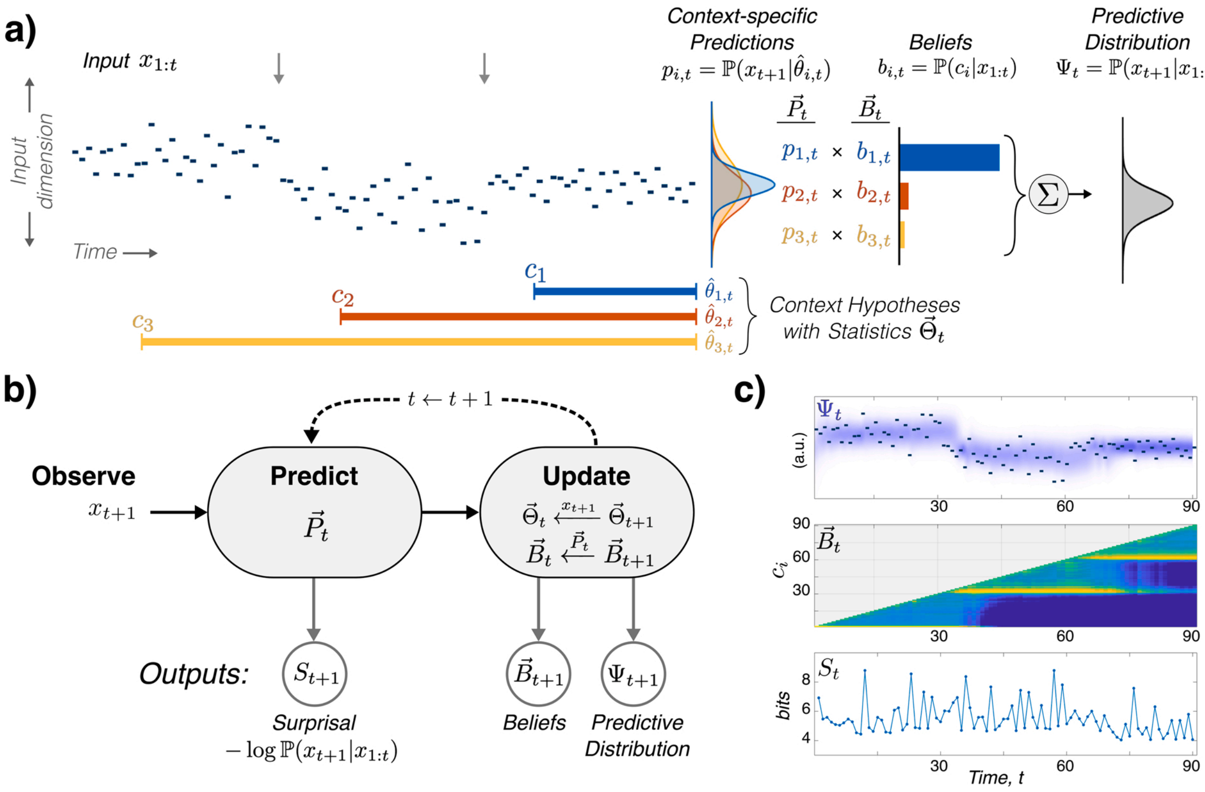

Fig. 1. Model description. a) The model uses multiple context hypotheses to account for unknown changes in the observed sequence. Context-specific predictions $\vec{P}_t$ based on sufficient statistics $\vec{\Theta}_t$ are combined, weighted by corresponding beliefs $\vec{B}_t$, to yield the predictive 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). The top panel shows the predictive distribution at each time (in blue) with the input sequence overlaid, the middle panel shows the context beliefs, with each row corresponding to a particular context hypothesis c_i, and the bottom panel shows the surprisal for each input observation. Note the predictive distribution and context beliefs reflect the underlying change in statistics inferred by the model.

$$\mathbb{P}(x_{t+1}|x_{1:t}) = \mathbb{P}(x_{t+1}|\widehat{\theta}_t) \tag{2}$$

where sufficient statistics $\widehat{\theta}_t$ are estimated from the context $x_{1:t}$.

2.2.2. *Multiple hypotheses for the unknown context*

The choice of context window impacts the quality of the prediction. For example, if θ changed at any point in the observed sequence, 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, each referred to as a *context hypothesis* for parsing 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 beginning of the i^{th} context window 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 predictive probabilities for the next observation given each context, $\vec{P}_t = \{p_{i,t}\}$. For the i^{th} context hypothesis:

$$p_{i,t} = \mathbb{P}(x_{t+1}|c_i, x_{c_i:t}) = \mathbb{P}(x_{t+1}|\widehat{\theta}_{i,t}) \tag{3}$$

Note that this context-specific predictive probability only depends on observations after the context boundary c_i, because observations before c_i are inferred to have been drawn from a different distribution (i.e., with different unknown parameters θ). In this section, time is simplified to be equivalent to the observation index, therefore c_i is unitless (i.e., $c_i = i$). However, in general the c_i's can occur at any point in time—in the examples in the next section, c_i has units in seconds.

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

$$b_{i,t} = \mathbb{P}(c_i|x_{1:t}) \tag{4}$$

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*. Depending on the application, this can be adjusted using the input parameters of the model to represent prior knowledge about when changes occur. In addition, a smaller set of context hypotheses decreases the computational cost of the model. Oftentimes, the beliefs are concentrated on a few context hypotheses (see Fig. 2a-i, middle panels), so reducing the set of context hypotheses by pruning or applying a threshold to the beliefs would not affect performance, and it would result in a sparser and more efficient tracking of the statistical past.

2.2.3. *"Integrating out" the unknown context*

To build the full predictive distribution Ψ_t, the context-specific predictive probabilities $p_{i,t}$ are combined across context hypotheses, weighted by their corresponding beliefs $b_{i,t}$ (see Fig. 1a-right):

$$\begin{aligned}\Psi_t = \mathbb{P}(x_{t+1}|x_{1:t}) &= \sum_{i=1}^{M}\mathbb{P}(x_{t+1},c_i|x_{1:t})\\ &= \sum_{i=1}^{M}\mathbb{P}(x_{t+1}|c_i,x_{c_i:t})\mathbb{P}(c_i|x_{1:t})\\ &= \sum_{i=1}^{M}p_{i,t}b_{i,t}\end{aligned} \tag{5}$$

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

Fig. 1a shows an illustration of how the model builds the predictive distribution for x_{t+1} given an example input sequence $x_{1:t}$ using three context hypotheses (with context windows starting at c_1, c_2, c_3 and statistics $\hat{\theta}_{1,t}$, $\hat{\theta}_{2,t}$, $\hat{\theta}_{3,t}$). For simplicity, time is equivalent to the sequential observation index. 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 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.

2.2.4. Iterative processing

Fig. 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 $\overrightarrow{P}_t$ (see Eq (3)).

Update Sufficient statistics $\overrightarrow{\theta}_t$ are updated sequentially with the newly observed input (e.g., for Gaussian distributions, see Murphy (2007)). Beliefs $\overrightarrow{B}_t$ are are also updated sequentially using the predictive probabilities, where the new beliefs reflect how well each context hypothesis predicted the newly observed input (see Adams and MacKay, 2007 for details).

The updated statistics and beliefs, $\overrightarrow{\Theta}_{t+1}$ and $\overrightarrow{B}_{t+1}$, are used in turn to process the subsequent input x_{t+2}, and so on. For more details on a particular application of this model using Gaussian statistics, see Skerritt-Davis and Elhilali (2018).

2.3. Model outputs

There are three main outputs from the model, as shown in Fig. 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 in bits:

$$S_{t+1} = -log\mathbb{P}(x_{t+1}|x_{1:t}) \tag{6}$$

where the probability is the likelihood of the observed input at time $t+1$ given all previous observations (see Eq (5)). 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 (Samson, 1953).

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 Skerritt-Davis and Elhilali, 2019). 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 (Overath et al., 2007; Barascud et al., 2016).

(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 (5)). 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 (Sedley et al., 2016; Kumar et al., 2013; Arnal and Giraud, 2012). For example, the predictive distribution can be used to examine the evolution of precision-weighted EEG responses in the brain (Barascud et al., 2016).

(iii) $\overrightarrow{B}_{t+1}$, the *context beliefs*, form the discrete posterior probability distribution over context hypotheses (see Eq (4)). The beliefs represent the relative evidence across context hypotheses. Similar to the predictive distribution, measures can be derived from the beliefs to relate it 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 in Fig. 1c can be used to compute the probability at least one change has occurred in the observed sequence:

$$\mathbb{P}(\text{Change}) = \mathbb{P}(c_i > 1|x_{1:t+1}) = \sum_{i:\ c_i > c_1} b_{i,t+1} \tag{7}$$

where the summation of beliefs *after* the initial context hypothesis c_1 represents the probability that the context begins *after* the beginning of the observed sequence (i.e., a change has occurred). Alternatively, the beliefs can be used to define a moment-by-moment measure of shift in the beliefs at each time as they adapt to changing statistics:

$$\delta_t = D_{\text{JS}}(\overrightarrow{B}_t||\overrightarrow{B}_{t+1}) \tag{8}$$

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 Skerritt-Davis and Elhilali, 2018). 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. 1c displays model outputs for an example sequence as it evolves over time (in black, same as in Fig. 1a). For this illustration, the time-axis simply refers to the index of the input observations. This same visual representation of the model outputs will be used in Section 3 below, with the time-axis corresponding to the onset timing of the input observations (in seconds).

The predictive distribution (Fig. 1c-top in blue) adapts to changes in the input observations (with darker blue corresponding to higher probability in the prediction and zero probability in white). Changes in the predictive distribution are a consequence of shifts in the context beliefs (Fig. 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 (here, yellow and blue correspond to larger and smaller beliefs, respectively). 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 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 (Fig. 1c-bottom) shows the momentary mismatch of each input after it has been observed. Note that higher surprisal corresponds with observations that fall farther outside of the predictive distribution in the top panel.

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.

2.4. 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.

2.4.1. Priors: π

The priors π are the initial statistical estimates for a new context hypothesis and take the same form as the sufficient statistics $\widehat{\theta}$ and have the same units. 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. In general, because only a few parameters need to be estimated (e.g., sample mean and variance), not much training data is needed, which is not the case for other statistical models (Pearce and Wiggins, 2012). The priors can also be used to test hypotheses about how prior knowledge affects predictions: for example, the effect of different long-term prior experience on listener responses to the same inputs, or the evolution of trial-to-trial learning over the course of an experiment.

2.4.2. Hazard rate: h_t

The hazard rate h_t is the probability of a change in the underlying statistics generating the sensory inputs (i.e., the parameters θ) 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 the input. The hazard rate can be constant, meaning that 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.

2.4.3. Perceptual parameters: M, N

Previous studies have shown that human listeners do not operate as ideal Bayesian observers (Wilson and Niv, 2012). Two perceptual parameters in the model represent neurally plausible constraints to predictive processing:

Memory M is the maximum number of context hypotheses and represents working memory capacity constraints in the brain (Conway et al., 2001; Just and Carpenter, 1992). 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 $\widehat{\theta}$.

Observation noise N sets a lower bound on prediction uncertainty, representing limitations in perceptual fidelity along the input dimension (Kidd et al., 2007; Wightman and Kistler, 1996). Observation noise is equivalent to adding independent Gaussian noise to the observed input with zero-mean and constant variance N^2, which has the effect of both increasing uncertainty of the prediction *and* decreasing precision of the sufficient statistics $\widehat{\theta}$. N has the same units as the input sequence.

Both of these perceptual parameters affect predictive processing, and they 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 Skerritt-Davis and Elhilali (2018).

3. Examples

In this section, we apply the D-REX model to real-world audio examples to give the reader some intuition behind the causal relationship between sensory inputs and model outputs. Examples were selected to represent a range of everyday sound sources from music, speech, and environmental sounds. These examples 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. While the tracking of the example sequences may seem obvious from an *engineering* perspective, how this is achieved in the brain is not. The model provides a framework to test alternative hypotheses for how the brain tracks statistical structures, infers relevant contexts, and integrates across multiple timescales and dimensions, hence facilitating the comparison between model predictions and experimental data (Skerritt-Davis and Elhilali, 2018, 2019).

Each panel in Fig. 2 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). 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. Table 1 contains a description of each example audio clip in Fig. 2. Audio clips and code used to create Fig. 2 are included in Supplementary Materials.

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

Spatial location. Fig. 2 a 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

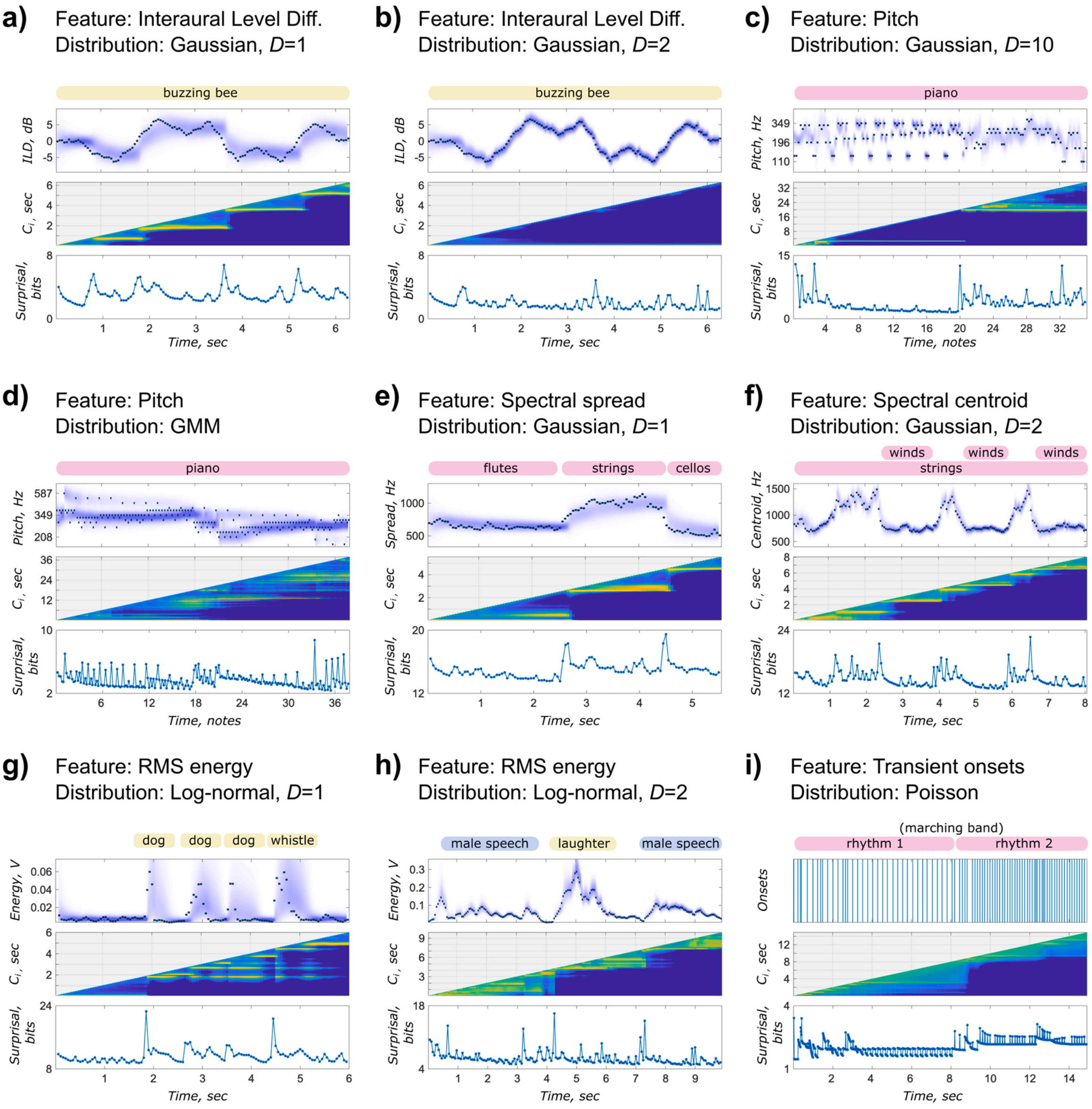

Fig. 2. Model outputs for example inputs from real-world audio clips. Each panel displays the model predictive distribution (top), context beliefs (middle), and surprisal (bottom) over time, with the input sequence overlaid on the predictive distribution (top, in black). The input dimension (feature), distributional choice in the model, and audio event annotation are indicated above. Audio clips can be found in Supplemental Information.

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. 2a and b use a Gaussian distribution in the model, but differ in the temporal dependence *D*. In Fig. 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 Fig. 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. 2a. This offers an alternative predictive interpretation of the same input sequence.

Pitch. Fig. 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. 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. 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. The prediction captures the

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

Table 1
List of audio examples. The first column refers to the panel in Fig. 2, the second column contains filenames of audio clips included in Supplementary Materials, and the third column contains URLs to the source for each audio clip.

Panel	Description	Filename	Source
a,b	Buzzing bee	AudioS1.mp3	ccrma.stanford.edu/azim/220a/hw2.html
c	Bach melody (MIDI)	AudioS2.mp3	en.midimelody.ru/bach-johann-sebastian
d	Bach melody (MIDI)	AudioS3.mp3	en.midimelody.ru/bach-johann-sebastian
e	Orchestral music	AudioS4.mp3	Youtube.com/watch?v=pKOpdt9PYXU
f	Orchestral music	AudioS5.mp3	Youtube.com/watch?v=pKOpdt9PYXU
g	Dog barks, whistle	AudioS6.mp3	Freesound.org/people/conleec/sounds/175917/
h	Conversational speech	AudioS7.mp3	Freesound.org/people/dobroide/sounds/33699/
i	Drum line	AudioS8.mp3	Youtube.com/watch?v=c4S4MMvDrHg

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. 2e and f use Gaussian distributions to predict two features of the spectrum from orchestral music recording: spectral spread and spectral 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. 2g and h 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. 2g, peaks in surprisal correspond 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. 2h, the beliefs capture turn-taking in conversational speech between a male speaker and group laughter.

Onset timing. The final example in Fig. 2i 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 perceptual 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.

4. 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 (Winkler et al., 2009; Friston and Kiebel, 2009; Friston, 2010; Bizley et al., 2013; Winkler and Schröger, 2015). The composition of the D-REX model aligns with previous literature regarding the underlying computations behind predictive processing: the brain builds statistical representations estimated from sounds over time (McDermott et al., 2013; Piazza et al., 2013; Dahmen et al., 2010; Brady et al., 2009), and the brain maintains multiple hypotheses for how much of the past is relevant to the present moment (Luo and Poeppel, 2012; Pieszek et al., 2013, Lau et al., 2017). 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 of sensory inputs with dynamic uncertainty.

By no means a complete picture of predictive coding in auditory perception, the D-REX model is a 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? Existing models in the literature offer a framework to explore these questions, though often constraining the mathematical formulation to a particular set of statistics or type of experiment (Furl et al., 2011; Garrido et al., 2013; McDermott et al., 2013; Barascud et al., 2016). The D-REX model was formulated to address these questions with a broader scope by allowing a comparison of different statistical representations and generalizing to many stimuli and perceptual tasks. For example, in one experiment, the model can be used to determine the statistical representation that best replicates listener responses; this same model can then be used to predict listener responses in a separate experiment. The model can also serve in the design of experiments and stimuli that intentionally tease apart hypothesized statistical representations (as in Skerritt-Davis and Elhilali, 2018).

Many experiments have demonstrated that our perception of the present is modulated by what we have heard in the past, from recent contextual effects (Snyder et al., 2008; Geiser et al., 2012; Luo and Poeppel, 2012; Herrmann et al., 2015; Mcwalter and Mcdermott, 2019) to life-long experience (Strait et al., 2010; Ch et al., 2009), and the D-REX model can be used to investigate these effects 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 of context 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

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

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, variability which has typically not been incorporated into previous models of predictive processing despite its influence on statistical learning (Siegelman and Frost, 2015). In addition to examining individual differences in the processes 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 listener responses.

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) (Sedley et al., 2016; Lecaignard et al., 2015; Hsu et al., 2015). 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 easily extendable, both conceptually and operationally in the code. In Section 3 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 used here are not complete. New probability distributions can be incorporated into the D-REX model, and the model can be applied along any dimension in the acoustic input. One current limitation in the model is that it operates on acoustic inputs that are sampled discretely in time. While this suffices for many experimental and real-world sounds that unfold sequentially in time (i.e., music, speech, alerting sounds), future work could extend the same modeling framework to operate in continuous time. Moreover, the modeling framework can be expanded in other ways to broaden its application. As currently implemented, the model operates at a single level of the sensory input and along a single time-scale, but it could be layered to build hierarchical predictions at different levels of abstraction or multiple time-scales (Heilbron and Chait, 2018). 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 predictive algorithms identified by the model could be implemented in neural circuits (Wilson et al., 2013).

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 practical 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.

Declaration of Competing Interest

The authors report no competing interest.

Acknowledgements

This work was supported by the National Institutes of Health (U01AG058532, R01HL133043, F31DC017629), the National Science Foundation (1734744), and the Office of Naval Research (N00014-19-1-2014, N00014-19-1-2689, N00014-17-1-2736, N00014-16-1-2879).

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jneumeth.2021.109177.

References

Adams, Ryan Prescott, MacKay, David J.C., 2007. Bayesian Online Changepoint Detection. Technical report. University of Cambridge, Cambridge, UK. URL http://arxiv.org/abs/0710.3742.

Agres, Kat, Abdallah, Samer, Pearce, Marcus, 2018. Information-Theoretic Properties of Auditory Sequences Dynamically Influence Expectation and Memory. Cognitive Science. https://doi.org/10.1111/cogs.12477. ISSN 15516709.

Arnal, Luc H., Giraud, Anne-Lise, 2012. Cortical oscillations and sensory predictions. Trends in Cognitive Sciences 16 (7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003. ISSN 13646613. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661312001210.

Barniv, Dana, Nelken, Israel, 2015. Auditory streaming as an online classification process with evidence accumulation. PLoS ONE. https://doi.org/10.1371/journal.pone.0144788. ISSN 19326203.

Barascud, Nicolas, Pearce, Marcus T., Griffiths, Timothy D., Friston, Karl J., Chait, Maria, 2016. Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. Proceedings of the National Academy of Sciences 113 (5), E616–E625. https://doi.org/10.1073/pnas.1508523113. ISSN 0027-8424. 2 URL http://www.pnas.org/lookup/doi/10.1073/pnas.1508523113.

Balaguer-Ballester, Emili, Clark, Nicholas R., Coath, Martin, Krumbholz, Katrin, Denham, Susan L., 2009. Understanding pitch perception as a hierarchical process with top-down modulation. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1000301. ISSN 15537358.

Bizley, Jennifer K., Walker, Kerry M.M., Nodal, Fernando R., King, Andrew J., Schnupp, Jan W.H., 2013. Auditory cortex represents both pitch judgments and the corresponding acoustic cues. Current biology: CB 23 (7), 620–625. https://doi.org/10.1016/j.cub.2013.03.003, 4 URL http://www.ncbi.nlm.nih.gov/pubmed/23523247.

Boubenec, Yves, Lawlor, Jennifer, Górska, Urszula, Shamma, Shihab, Englitz, Bernhard, 2017. Detecting changes in dynamic and complex acoustic environments. eLife 6, 3. https://doi.org/10.7554/eLife.24910. ISSN 2050-084X. URL https://elifesciences.org/articles/24910.

Brady, Timothy F., Konkle, Talia, Alvarez, George A., 2009. Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations. Journal of Experimental Psychology: General. https://doi.org/10.1037/a0016797. ISSN 00963445.

Clark, Andy, 2013. Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. ISSN 14691825.

Chandrasekaran, Bharath, Krishnan, Ananthanarayan, Gandour, Jackson T., 2009. Sensory processing of linguistic pitch as reflected by the mismatch negativity. Ear and hearing 30 (5), 552.

Conway, Andrew R.A., Cowan, Nelson, Bunting, Michael F., 2001. The cocktail party phenomenon revisited: The importance of working memory capacity. Psychonomic Bulletin & Review 8 (2), 331–335. https://doi.org/10.3758/BF03196169. ISSN 1069-9384. URL http://www.springerlink.com/index/10.3758/BF03196169.

Dahmen, Johannes C., Keating, Peter, Nodal, Fernando R., Schulz, Andreas L., King, Andrew J., 2010. Adaptation to stimulus statistics in the perception and neural representation of auditory space. Neuron 66 (6), 937–948. https://doi.org/10.1016/j.neuron.2010.05.018. URL http://www.cell.com/neuron/fulltext/S0896-6273(10)00386-7.

Daunizeau, Jean, den Ouden, Hanneke E.M., Pessiglione, Matthias, Kiebel, Stefan J., Stephan, Klaas E., Friston, Karl J., 2010. Observing the observer (I): Meta-bayesian models of learning and decision-making. PLoS ONE 5 (12), e15554–10. https://doi.org/10.1371/journal.pone.0015554. ISSN 19326203. URL http://dx.plos.org/10.1371/journal.pone.0015554.

Denham, Susan, Böhm, Tamás M., Bendixen, Alexandra, Szalárdy, Orsolya, Kocsis, Zsuzsanna, Mill, Robert, Winkler, István, 2014. Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. Frontiers in Neuroscience. https://doi.org/10.3389/fnins.2014.00025. ISSN 1662453X.

Denham, Susan L., Winkler, István, 2020. Predictive coding in auditory perception: challenges and unresolved questions. ISSN 14609568.

Di Liberto, Giovanni M., Pelofi, Claire, Bianco, Roberta, Patel, Prachi, Mehta, Ashesh D., Herrero, Jose L., de Cheveigné, Alain, Shamma, Shihab, Mesgarani, Nima, 2020. Cortical encoding of melodic expectations in human temporal cortex. eLife 9, 3. https://doi.org/10.7554/eLife.51784. ISSN 2050-084X. URL https://elifesciences.org/articles/51784.

Friston, Karl, Kiebel, Stefan, 2009. Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B: Biological Sciences. https://doi.org/10.1098/rstb.2008.0300. ISSN 14712970.

Friston, Karl J., 2010. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11 (2), 127–138. https://doi.org/10.1038/nrn2787. ISSN 1471-003X. URL http://www.nature.com/articles/nrn2787.

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., Griffiths, T.D., 2011. Neural prediction of higher-order auditory sequence statistics. NeuroImage 54 (3), 2267–2277. URL http://www.ncbi.nlm.nih.gov/pubmed/20970510.

Garrido, Marta I., Sahani, Maneesh, Dolan, Raymond J., 2013. Outlier Responses Reflect Sensitivity to Statistical Structure in the Human Brain. PLoS Computational Biology 9 (3), e1002999. https://doi.org/10.1371/journal.pcbi.1002999. ISSN 1553-7358. 3 URL https://dx.plos.org/10.1371/journal.pcbi.1002999.

Geiser, Eveline, Notter, Michael, Gabrieli, John D.E., 2012. A corticostriatal neural system enhances auditory perception through temporal context processing. Journal of Neuroscience. https://doi.org/10.1523/JNEUROSCI. 5153-11.2012. ISSN 02706474.

Grossberg, Stephen, 1980. How does a brain build a cognitive code? Psychological Review. https://doi.org/10.1037/0033-295X871.1. ISSN 0033295X.

Grossberg, Stephen, Govindarajan, Krishna K., Wyse, Lonce L., Cohen, Michael A., 2004. ARTSTREAM: a neural network model of auditory scene analysis and source segregation. Neural Networks 17 (4), 511–536. https://doi.org/10.1016/j.neunet.2003.10.002. ISSN 08936080. 5 URL https://linkinghub.elsevier.com/retrieve/pii/S0893608003002727.

Hansen, N.C., Pearce, M.T., 2014. Predictive uncertainty in auditory sequence processing. Frontiers in psychology 5 (9), 1052. https://doi.org/10.3389/fpsyg.2014.01052.

Heilbron, Micha, Chait, Maria, 2018. Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? Neuroscience 389, 54–73. https://doi.org/10.1016/j.neuroscience.2017.07.061. ISSN 18737544. URL https://doi.org/10.1016/j.neuroscience.2017.07.061https://doi.org/10.1016/j.neuroscience.2017.07.061.

Herrmann, Björn, Henry, Molly J., Fromboluti, Elisa Kim, Devin McAuley, J., Obleser, Jonas, 2015. Statistical context shapes stimulus-specific adaptation in human auditory cortex. Journal of Neurophysiology 113 (7), 2582–2591. https://doi.org/10.1152/jn.00634.2014. ISSN 0022-3077. URL http://jn.physiology.org/lookup/doi/10.1152/jn.00634.2014.

Hsu, Y.F., Le Bars, S., Hamalainen, J.A., Waszak, F., 2015. Distinctive Representation of Mispredicted and Unpredicted Prediction Errors in Human Electroencephalography. Journal of Neuroscience 35 (43), 14653–14660. https://doi.org/10.1523/JNEUROSCI, 2204-15.2015. 10 URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI. 2204-15.2015.

Just, Marcel A., Carpenter, Patricia A., 1992. A capacity theory of comprehension: Individual differences in working memory. Psychological Review 99 (1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122. ISSN 1939-1471. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X99.1.122.

Karl, J., Friston, 2005. A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences 360 (1456), 815–836. https://doi.org/10.1098/rstb.2005.1622, 4 URL http://rstb.royalsocietypublishing.org/content/360/1456/815.abstract.

Kidd, Gary R., Watson, Charles S., Gygi, Brian, 2007. Individual differences in auditory abilities. The Journal of the Acoustical Society of America 122 (1), 418–435. https://doi.org/10.1121/1.2743154. ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.2743154.

Knill, David C., Pouget, Alexandre, 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in neurosciences 27 (12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007. URL http://www.sciencedirect.com/science/article/pii/S0166223604003352.

Kumar, S., Joseph, S., Pearson, B., Teki, S., Fox, Z.V., Griffiths, T.D., Husain, M., 2013. Resource allocation and prioritization in auditory working memory. Cognitive Neuroscience 4 (1), 12–20. https://doi.org/10.1080/17588928.2012.716416. ISSN 17588928.

Lau, Brian, Monteiro, Tiago, Paton, Joseph J., 2017. The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. ISSN 18736882.

Lecaignard, Françoise, Bertrand, Olivier, Gimenez, Gérard, Mattout, Jérémie, Caclin, Anne, 2015. Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy. Frontiers in Human Neuroscience. https://doi.org/10.3389/fnhum.2015.00505. ISSN 16625161.

Lieder, F., Daunizeau, J., Garrido, Marta I., Friston, Karl J., Stephan, K.E., 2013. Modelling Trial-by-Trial Changes in the Mismatch Negativity. PLoS computational biology 9 (2), e1002911.

Luo, Huan, Poeppel, David, 2012. Cortical oscillations in auditory perception and speech: Evidence for two temporal windows in human auditory cortex. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2012.00170. ISSN 16641078.

McDermott, Josh H., Wrobleski, David, Oxenham, Andrew J., 2011. Recovering sound sources from embedded repetition. In: Proceedings of the National Academy of Sciences of the United States of America. ISSN 00278424. doi:10.1073/pnas.1004765108.

McDermott, Josh H., Schemitsch, Michael, Simoncelli, Eero P., 2013. Summary statistics in auditory perception. Nature Neuroscience 16 (4), 493–498. https://doi.org/10.1038/nn.3347. ISSN 1097-6256. URL http://www.nature.com/articles/nn.3347.

Mcwalter, Richard, Mcdermott, Josh H., 2019. Temporal Integration Windows for Auditory Statistical Estimation. Proceedings of the 23rd International Congress on Acoustics.

Mill, R.W., Bohm, T.M., Bendixen, A., Winkler, István, Denham, S.L., 2013. Modelling the emergence and dynamics of perceptual organisation in auditory streaming. PLoS computational biology 9 (3), e1002925. https://doi.org/10.1371/journal.pcbi.1002925.

Murphy, Kevin P., 2007. Conjugate Bayesian Analysis of the Gaussian Distribution. Technical Report, p. 7.

Nassar, M.R., Wilson, R.C., Heasly, B., Gold, J.I., 2010. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. Journal of Neuroscience 30 (37), 12366–12378. https://doi.org/10.1523/JNEUROSCI. ISSN 0270-6474. 0822-10.2010. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI. 0822-10.2010.

Nix, J., Hohmann, V., 2007. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. IEEE Transactions on Audio, Speech and Language Processing 15 (3), 995–1008.

Overath, Tobias, Cusack, Rhodri, Kumar, Sukhbinder, von Kriegstein, Katharina, Warren, Jason D., Grube, Manon, Carlyon, Robert P., Griffiths, Timothy D., 2007. An Information Theoretic Characterisation of Auditory Encoding. PLoS Biology 5 (11), e288. https://doi.org/10.1371/journal.pbio.0050288.g001.

Pearce, Marcus, 2005. The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition. PhD thesis. URL http://openaccess.city.ac.uk/8459/.

Pearce, Marcus T., Wiggins, Geraint A., 2012. Auditory Expectation: The Information Dynamics of Music Perception and Cognition. Topics in Cognitive Science. https://doi.org/10.1111/j. 1756-8765.2012.01214.x. ISSN 17568757.

Piazza, Elise A., Sweeny, Timothy D., Wessel, David, Silver, Michael A., Whitney, David, 2013. Humans Use Summary Statistics to Perceive Auditory Sequences. Psychological Science. https://doi.org/10.1177/0956797612473759. ISSN 14679280.

Pieszek, Marika, Widmann, Andreas, Gruber, Thomas, Schröger, Erich, 2013. The Human Brain Maintains Contradictory and Redundant Auditory Sensory Predictions. PLoS ONE 8 (1), e53634. https://doi.org/10.1371/journal.pone.0053634. ISSN 19326203. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0053634.

Samson, Edward, 1953. Fundamental Natural Concepts of Information Theory. ETC: A Review of General Semantics 10 (4), 283–297. https://doi.org/10.1016/b978-0-08-010421-8, 50013-8. URL http://www.jstor.org/stable/42581366.

Sedley, William, Gander, Phillip E., Kumar, Sukhbinder, Kovach, Christopher K., Oya, Hiroyuki, Kawasaki, Hiroto, Howard, Matthew A., Griffiths, Timothy D., 2016. Neural signatures of perceptual inference. eLife. https://doi.org/10.7554/eLife.11476. ISSN 2050084X.

Seriès, Peggy, Seitz, Aaron R., 2013. Learning what to expect (in visual perception). ISSN 16625161.

Siegelman, Noam, Frost, Ram, 2015. Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. Journal of Memory and Language. https://doi.org/10.1016/j.jml.2015.02.001. ISSN 0749596X.

Skerritt-Davis, Benjamin, Elhilali, Mounya, 2019. A Model for Statistical Regularity Extraction from Dynamic Sounds. Acta Acustica united with Acustica 105 (1), 1–4. https://doi.org/10.3813/AAA.919279. ISSN 1610-1928. URL https://www.ingentaconnect.com/content/10.3813/AAA.919279.

Skerritt-Davis, Benjamin, Elhilali, Mounya, 2018. Detecting change in stochastic sound sequences. PLOS Computational Biology 14 (5), e1006162. https://doi.org/10.1371/journal.pcbi.1006162. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/29813049.

Strait, Dana L., Kraus, Nina, Parbery-Clark, Alexandra, Ashley, Richard, 2010. Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. Hearing Research. https://doi.org/10.1016/j.heares.2009.12.021. ISSN 03785955.

Snyder, J.S., Carter, O.L., Lee, S.K., Hannon, E.E., Alain, C., 2008. Effects of context on auditory stream segregation. Journal of experimental psychology. Human perception and performance 34 (4), 1007–1016. https://doi.org/10.1037/0096-1523.34.4.1007.

Tabas, Alejandro, Andermann, Martin, Schuberth, Valeria, Riedel, Helmut, Balaguer-Ballester, Emili, Rupp, André, 2019. Modeling and MEG evidence of early consonance processing in auditory cortex. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1006820. ISSN 15537358.

Tenenbaum, Joshua B., Griffiths, Thomas L., Kemp, Charles, 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences 10 (7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009. ISSN 13646613.

Wacongne, Catherine, Changeux, J.P., Dehaene, S., 2012. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. Journal of Neuroscience 32 (11), 3665–3678. https://doi.org/10.1523/JNEUROSCI, 5003-11.2012. 3 URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI. 5003-11.2012.

Winkler, István, Schröger, Erich, 2015. Auditory perceptual objects as generative models: Setting the stage for communication by sound. Brain and Language 148 (C), 1–22. https://doi.org/10.1016/j.bandl.2015.05.003. ISSN 10902155. 9 URL https://doi.org/10.1016/j.bandl.2015.05.003.

Wilson, Robert C., Nassar, Matthew R., Gold, Joshua I., 2013. A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems. PLoS Computational Biology 9 (7). https://doi.org/10.1371/journal.pcbi.1003150. ISSN 1553734X.

Wilson, Robert C., Niv, Yael, 2012. Inferring relevance in a changing world. Frontiers in Human Neuroscience. https://doi.org/10.3389/fnhum.2011.00189. ISSN 16625161.

Wightman, Frederic L., Kistler, Doris J., 1996. Individual differences in human sound localization behavior. The Journal of the Acoustical Society of America. https://doi.org/10.1121/1.415531. ISSN 0001-4966.

Winkler, István, Denham, Susan L., Nelken, Israel, 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends in Cognitive Sciences 13 (12), 532–540. https://doi.org/10.1016/j.tics.2009.09.003. ISSN 13646613. 10 URL https://linkinghub.elsevier.com/retrieve/pii/S1364661309002095.