UNIVERSITY OF CALIFORNIA SAN DIEGO

Predictive Coding in the Auditory Cortex

A Thesis submitted in partial satisfaction of the requirements for the degree
Master of Science

in

Bioengineering

by

Srihita Rudraraju

Committee in charge:

Professor Timothy Gentner, Chair
Professor Gabriel Silva, Co-Chair
Professor Gert Cauwenberghs

2019

The Thesis of Srihita Rudraraju is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

Co-chair

_____

Chair

University of California San Diego

2019

DEDICATION

This is work is dedicated to my family. They have always supported my endeavors and encouraged me to be the best I can be. I am very grateful for their love and guidance.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Timothy Gentner for his support as the chair of my committee. His support, guidance, and wisdom were instrumental in the completion of this work and have proved to be invaluable.

I would also like to acknowledge Marvin Thielk, Nasim Vahidi, Bradley Thielman, Tim Sainburg, Ezequiel Arneodo, Michael Turvey, Sasen Cain and all members of the Gentner Lab, who have generously provided their time, expertise, and resources towards the success of this work and my growth as a researcher.

I would like to acknowledge Dr. Tatyana Sharpee, and her laboratory at Salk Institute for Biological Studies who developed the MNE model and provided assistance through some of the challenges in this work.

ABSTRACT OF THE THESIS

Predictive Coding in the Auditory Cortex

by

Srihita Rudraraju

Master of Science in Bioengineering

University of California San Diego, 2019

Professor Timothy Gentner, Chair

Professor Gabriel Silva, Co-chair

Characterization of response properties of neurons in higher-level sensory areas is not well defined. Here we show that firing rates of neurons in a secondary sensory forebrain area of songbirds can be modeled by different representations of birdsong. In this work, we modeled neurons in the caudo-medial nidopallium (NCM) of adult European starlings with three different representations of the natural birdsong called signal, prediction, and error. Prediction spectrogram was computed by training the data as a Gaussian distribution on a loss function given by the

negative log likelihood, and then estimating the means and variances of the signal. Using our Maximum Noise Entropy (MNE) model, responses were predicted by the logistic function, the parameters of which are obtained from the MNE model. Predictions of neural responses were computed by using both a full MNE model, and then by only considering the linear parameters of the model. The neural responses to natural stimuli obtained using prediction and error MNEs were close to the actual response in the NCM. The concept of stimulus representations obtained from predictive coding models may be useful for modeling neural responses in higher-order sensory areas whose functions have been poorly understood.

1. Introduction

How does the brain comprehend the world? It is not a passive, receptive, but an active process of constructing reality, it is an interaction between the senses and the cortex that balances new information from the outside world with predictions from the inside of our brain [1].

A popular theory addressing this question is predictive coding (PC). PC states that the brain understands what is 'out there' by constantly predicting what is out there and improving those predictions. The perceiving brain is continuously predicting the incoming sensory input and tries to 'fit' the model. More technically, predictive coding proposes that the brain constructs a generative hierarchical model of the world. This model is capable of generating sensory input activity from the top-down and match it with that external stimuli would elicit from the bottom-up [2]. Human EEG and magnetoencephalographic (MEG) recordings showed responses to mismatched or omitted stimuli from a hierarchical auditory novelty paradigm. They concluded that detection of auditory novelty is organized in several stages: MMN responds to local auditory predictions, and P3b responds to more global and integrative violations of expectations. [3]

Predictive coding (PC) aims to offer a unified theory of cortical function. The framework has drawn a considerable amount of attention, hailed by some as providing a 'grand unified theory of the brain'. Initially, PC was conceptualized in the context of visual processing. However, with many studies capitalizing on the auditory system, it is quickly becoming perhaps the most well-studied neural signature of surprise or error processing [2,4].

1.1 Songbird Auditory System

Humans are experts at vocal learning, an ability shared by few other vertebrates. Songbirds represent an almost unique animal model: song learning has remarkable parallels to human speech

learning, which provides an opportunity for mechanistic investigation of vocal learning and its disorders. Song is processed by circuitry that is specialized for vocal learning and production, but that has strong similarities to mammalian brain pathways (Fig 1.2).

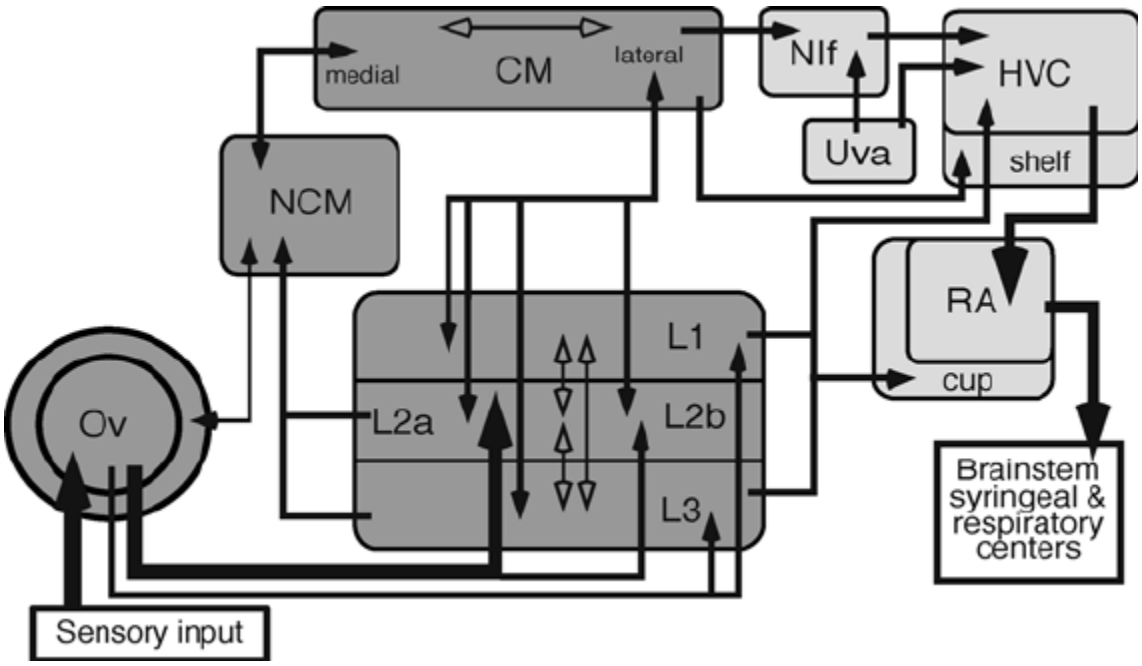

Figure 1.1. Songbird auditory system. Forebrain auditory regions are shown in dark grey and primary motor pathway is shown in light grey. This figure shows major regions in songbird auditory system and their connectivity [6].

The auditory cortex, located in the lateral domains of the temporal lobe, is the location of primary auditory processing in the brain. The acquisition and learning of bird song involves a group of distinct brain areas that are aligned in two connecting pathways [25,26]:

1) Anterior forebrain pathway (vocal learning): composed of area X (homolog to mammalian basal ganglia), lateral part of the magnocellular nucleus of anterior nidopallium (LMAN), and the dorso-lateral division of the medial thalamus (DLM).
2) Posterior descending pathway (vocal production): composed to HVC, robust nucleus of the arcopallium (RA).

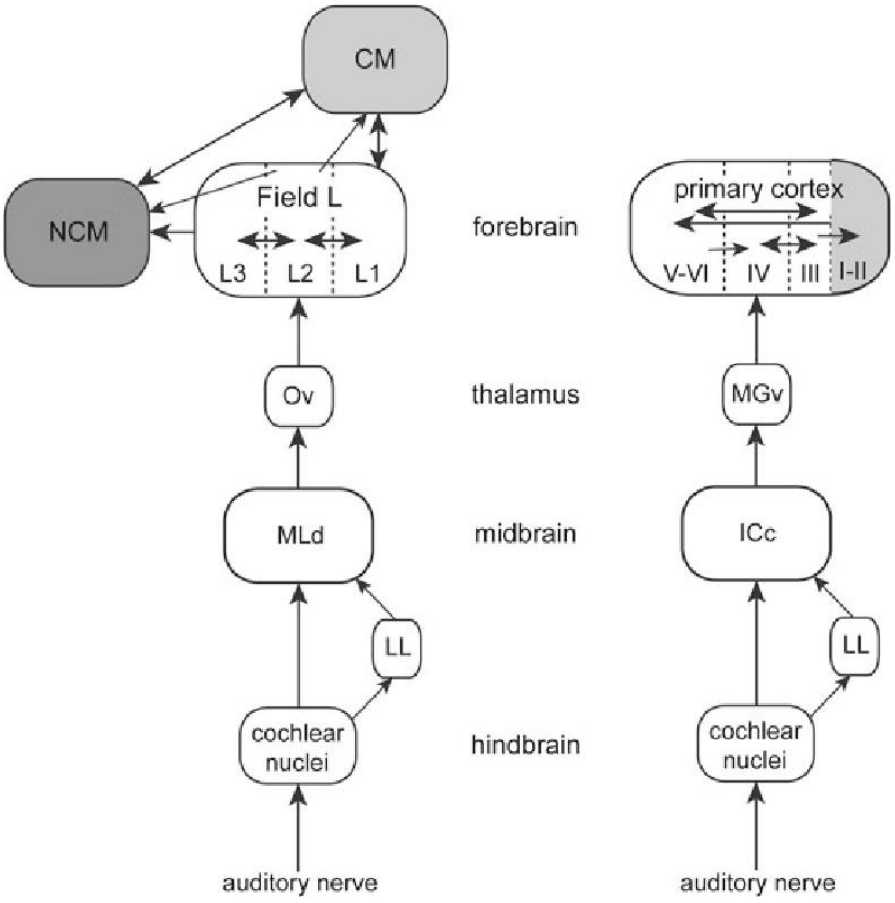

Figure 1.2. Primary pathways of songbird and mammal auditory systems. Songbird auditory system is shown on left and mammalian system is shown on the right. The gray areas are proposed to be homologous based on connectivity and topology. NCM does not have a known homologous region in the mammalian cortex [24].

Responses to songs in higher auditory regions reveal a variety of complex coding properties. Field L subregions L2 and L3 send inputs to NCM, which is posterior to L3. The encoding of sounds and the integrations processes are more complex in songbird. In field L, most neurons are activated only by complex sounds, and responded only to natural sounds. The field L has projections on HVC, which itself projects on the RA.

The response properties of NCM neurons are characterized by linear or non-linear models such as spectrotemporal receptive fields (STRFs). Characterization of the spectrotemporal properties of neurons in these regions have proven to be more difficult than in lower regions [5]. NCM is involved in the processing/ categorization of conspecific songs. Strong conspecific-selective responses have been consistently demonstrated in neurons of NCM, CM and Field L. In European starlings, neurons in NCM habituate to a particular stimulus, and "remember" individual characteristics of songs to which a bird was exposed [6-8]. Enhanced immediate early gene expression and neurophysiological activity in response to songs in CM/ NCM suggest that these sites are critical for vocal learning, with analogies to speech-related regions of the human superior temporal gyrus [27].

1.2 Basics of predictive coding

The sensory cortex has a hierarchical organization. At every level, neurons integrate information received through multiple connections from neurons at the lower level, but also receive inputs from the layer above. The cortex is reciprocally connected. The architecture of the cortex implements a top-down prediction algorithm that constantly predicts incoming sensory stimuli. These predictions are compared with novel incoming inputs. Each cortical area houses an internal model of the environment generated by repeated trials of past inputs.

The difference between the predicted and actual activity at the layer elicits a prediction error. Only this difference is propagated to the layer above. This error is used to generate a new and improved estimate. The prediction error is used to improve the internal model. The process is repeated, at every level in the hierarchy until the most likely estimate is reached, and the stimulus is perceived. This results in an active system that continuously updates its internal models at multiple hierarchical levels. [2]

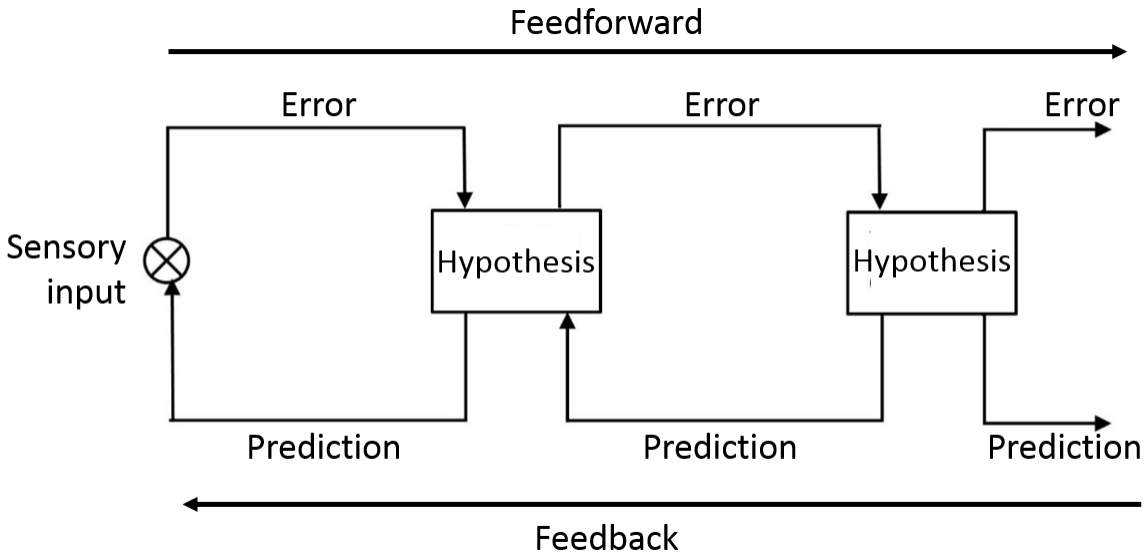

Figure 1.3. Basic structure of predictive coding model. Hierarchical model, higher-level coding units attempt to predict responses of units in the next lower level via feedback connections, while lower-level error detectors signal the difference between the prediction and the actual input.

1.3 Evidence of predictive coding in the animal auditory cortex

According to the predictive coding hypothesis, neurons at higher processing stages generate predictions that bias processing at lower levels. The abstract information at higher levels informs and potentially drives neurons at lower levels by signaling a prior 'best guess' of their activity [9]. Findings from animal, human and computational neuroscience provide converging evidence for the fundamental influence of expectations on neural responses and specifically the notion of prediction error as a model of neural responsiveness.

Most research on auditory prediction focusses on Stimulus Specific Adaptation (SSA). SSA refers to the selective attenuation of responses to repeated stimuli and can be seen as a single cell analog to MMN. Research showed that SSA, defined here as the difference in responses to the same sound presented with different probabilities, depended not just on local context but also on

a longer stimulus history, beyond the order of seconds at which habituation processes are thought to occur [10,11].

The same data was re-analyzed in an attempt to quantify the longer-term dependencies. Results showed representations involving less than 10 preceding stimuli (7.3 s) were almost never in the top 10%. The authors concluded that neurons in A1 signal prediction errors ‘generate predictions’ based on reduced representations that include long-term stimulus history. These results deviate from earlier accounts of SSA, which tend to focus on stimulus-driven explanations [12].

Ulanovsky in 2003 showed that in cats, A1 neurons in the primary auditory cortex responded more strongly to a rarely presented sound than to the same sound when it was common. This was shown for frequency deviants. Their paper showed frequency discrimination is better when processing deviant frequencies, as compared to frequencies that are close. They did not observe any differences in responses in the thalamus. So, they concluded that the origin for this process is above the thalamus. [11]

Gill in 2008 explored surprise as a model for auditory receptive fields. Their paper compared three receptive field models based on natural Zebra finch song: 1) a traditional approach modeling neurons as responding to specific spectrotemporal receptive fields (STRF) showing intensity patterns; 2) a derivative approach, modeling changes in intensities; 3) a model describing neurons as responding to surprise, quantified as the inverse conditional probability of a range of frequencies given the preceding frequencies shown in Fig 1.4. The ‘surprise model’ substantially outperformed traditional models. [13]

The performance of these models also depended on the hierarchical level. In area MLD (homolog of inferior colliculus), models did not differ significantly. In field L (homolog of thalamorecipient neurons in A1), surprise was 20% better than traditional models on average. And in CLM (homolog of higher-order auditory cortex), the surprise model performed a striking 67% better on average. The 'expectations' computed in this research were dependent on very short preceding time windows (3-7 ms). The paper concluded that expectations are increasingly important at higher levels, but the effect is not a direct consequence of high -level 'surprise'. The paper does not explain how the expectations were computed. However, it does successfully show the importance of expectations at the fundamental level of the neural code.

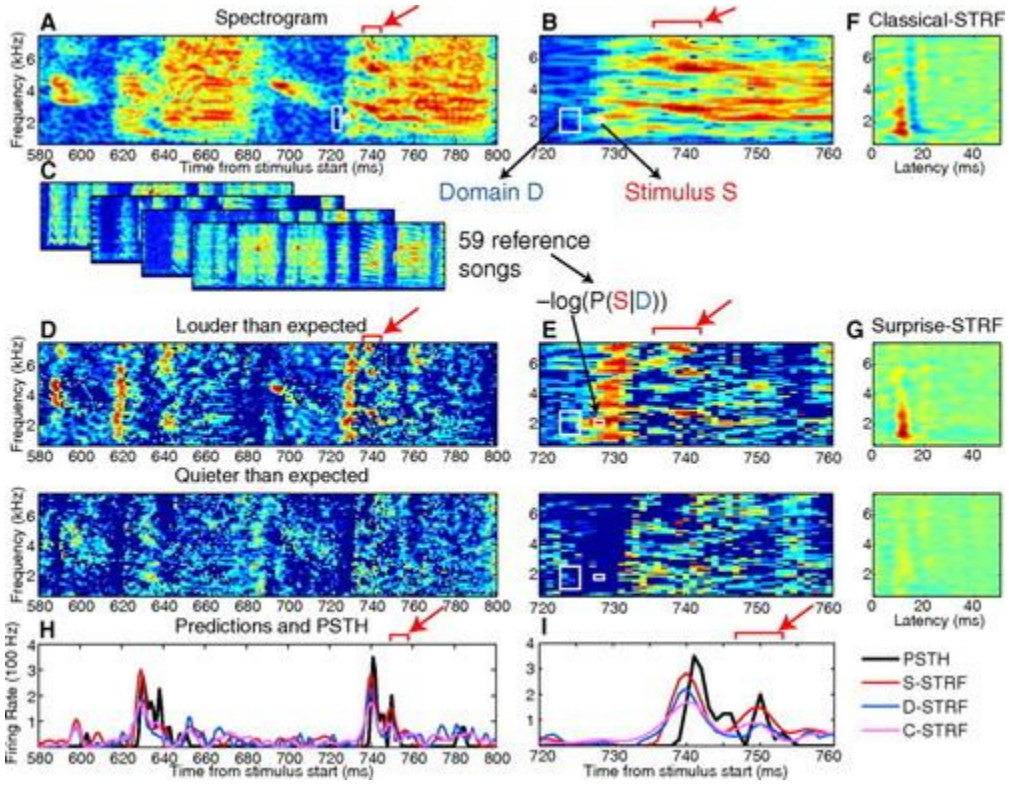

Figure 1.4. Comparisons of predicted PSTHs to the actual PSTHs. Classical-STRF, surprise-STRF and derivative-STRF of segments of zebra finch song calculated by reverse correlation between the PSTH and spectrogram segments. Highlighted red areas show where there is a surprise element within a syllable, captured by the surprise-STRF [13].

Human auditory studies on predictive coding use a variation of the MMN paradigm. MMN is measured using a method in which a sequence of stimuli (typically a repeated tone) establishes a regularity that is violated by a 'deviant' stimulus (oddball paradigm). PC interprets MMN as a mismatch signal between the input and a prospective prediction, it is not a separate evoked response.

Evidence from EEG, MEG, ECoG shows that omissions can evoke responses that are time-locked to omitted stimulus and appear to be generated in the auditory cortex and superior temporal gyrus. Omission responses seem to occur only after unexpected omissions, suggesting a predictive mechanism. However, research showed some remarkable variability. No omission responses were found using MEG, whereas using EEG they could find clearer responses although they were strikingly different from real auditory-evoked potentials (AEPs) [3,28-34].

Omission responses are perhaps the signature finding of PC. By showing that evoked responses fundamentally reflect surprise, these responses are even observable in the absence of sensory input. However, this interpretation critically depends on how prediction error is defined. These studies present highly suggestive, converging evidence of anticipatory mechanisms, operating without conscious expectation, in auditory cortex. However, due to ambiguities in error calculation, it is difficult to directly interpret the implications of omission responses to predictive coding.

Animal model studies relevant to the assumptions of predictive coding are scarce and show mixed results. None of the discussed studies explicitly tested PC, which may contribute to the inconclusiveness of the results. Methodological differences between these studies, and the fact that they did not address the mechanisms of prediction, unfortunately limit their conclusiveness with respect to PC. However, there is a conceptual shift from characterizing neurons as encoding

bottom-up data features, to encoding hypotheses or predictions, and propagating only the divergence from these predictions.

Although human neuroimaging studies – which are confined to investigating the macroscopic level of brain organization – can be informative about cortical function, empirical support for predictive coding understanding at the level of single neurons is lacking. In this work, we explore the concept of predictive coding in single neuron.

In Chapter 2, we applied machine learning models to devise a predictive coding mechanism for a single neuron in the NCM. Although, we did not have strong reasons to prove that the chosen model has better performance over other existing machine learning techniques, Deep Gaussian model presents itself as a good starting point. In Chapter 3, we introduce a technique to analyze the predictive capacity of our model.

2. Deep Gaussian

2.1 Introduction

In this chapter, a deep learning model that uses the negative log-likelihood of Gaussian distribution as a loss was created. The model estimates the mean and variance of the probability distribution of a target as a function of the input, given a Gaussian target error-distribution model.

Deep neural networks in recent years have emerged as flexible parametric models which can fit complex patterns in data. Gaussian processes are a traditional nonparametric tool for modeling. Each layer in the network can be considered as a Gaussian Process (GP). The input to each layer is governed by another GP. The data input to the network is modeled as the output of a multivariate GP. Gaussian Processes govern the mappings between the layers.

2.2 Gaussian Model

2.2.1 Problem setup

It was assumed that the training dataset D consists of N independent and identically distributed data points $D = \{x_n, y_n\}^N{}_{n=1}$, where $x \epsilon R^D$ represents the D-dimensional features. The label was assumed to be real-valued, that is $y \epsilon R$. Given the input features x, a neural network was used to model the probabilistic predictive distribution p(y|x) over the labels. The process for training this neural network is discussed below.

2.2.2 Training criterion

Neural networks usually output a single value, say μ(x), and the parameters are optimized to minimize the mean squared error (MSE) on the training set, given by $\sum_{n=1}^{N}(y_n - \mu(x_n))^2$. However, MSE does not capture predictive uncertainty. The network used in this case outputs two

values in the final layer, corresponding to the predicted mean $\mu(x)$ and variance $\sigma^2(x) > 0$. By treating the observed value as a sample from a (heteroscedastic) Gaussian distribution with the predicted mean and variance, we minimize the negative log-likelihood criterion [14,15]:

$$-\log p(y_n|x_n) = \frac{\log \sigma^2(x)}{2} + \frac{(y - \mu(x))^2}{2\sigma^2(x)} + constant$$

2.2.3 Adversarial training

Adversarial examples are perturbed inputs designed to fool machine learning models. These are 'close' to the original training examples (for example, an image that is visually indistinguishable to the original image) but are misclassified by the neural network. Such examples are injected to increase robustness. To scale this technique to large datasets, perturbations were crafted using fast single-step methods (fast gradient sign method in this case) that maximize a linear approximation of the model's loss.

Given an input x with target y and loss l(x, y) given by $-\log p(y_n|x_n)$, the fast gradient sign method generates an adversarial example as $x' = x + \epsilon(\nabla_x l(x, y))$, where ϵ is a small value. The adversarial perturbation creates a new training example by adding a perturbation along a direction in which the network is likely to increase the loss. This process is called adversarial training [14,15].

2.3 Network Details

2.3.1 Architecture

Both μ and σ^2 were connected to a common large set of hidden units h_j (indexed by j). In addition, a second and a third shared hidden layers were added as per required (Fig 2.1).

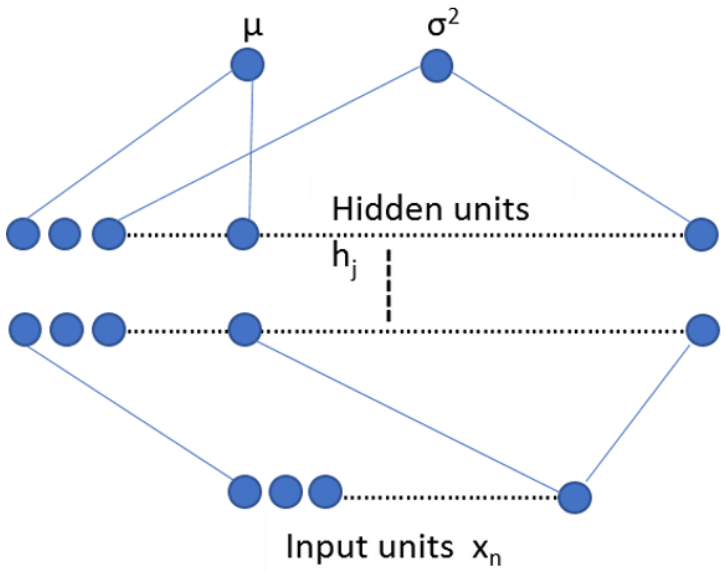

Figure 2.1. Deep Gaussian Neural Network Architecture. Architecture of network with output unit μ and variance unit σ^2. Hidden units connected to input units and have shared connections (not all connections shown). Output units share connections to hidden layer. Network has multiple hidden layers.

2.3.2 Learning Dynamics

Random initialization of the NN parameters, along with random shuffling of the data points, was enough to obtain a good performance. The overall training procedure is summarized in Table 2.1.

Table 2.1. Algorithm. Pseudocode for the training procedure of method.

1. Let neural network parametrize a distribution over the outputs, i.e. $p(y|x)$. Use training criterion $l(x, y)$. Default value for $\epsilon = 10^{-6}$.
2. Initialize Θ (parameters of the NN) randomly.
3. Sample data point n randomly for network (batches for training)
4. Minimize $l(x_n, y_n)$ with respect to Θ.

2.3.3 Experimental setup

The negative log likelihood (NLL) was evaluated, which depends on the predictive uncertainty. NLL is a proper scoring rule and a popular metric for evaluating predictive uncertainty. A batch size of 1024, and Adam optimizer with a fixed learning rate of 0.001 were used in these experiments. Rectified Linear Unit (relu) nonlinearity was added onto layers for more complexity, and default weight initializations were used. The feedback intervals were set to 500. The model has three hidden layers with 64, 64 and 32 units in each layer respectively. The training dataset was split into train and validation sets with a validation ratio of 0.06.

2.3.4 Platform

TensorFlow was used to run this network. Although, Keras is usually the first-choice deep learning framework, it requires backend functions written in TensorFlow or Theano if customized loss functions or layers need to be used. As the negative log-likelihood of gaussian distribution is not one of the available loss functions in Keras, the network was implemented in TensorFlow.

2.3.5 Inputs data

Natural stimuli elicit robust responses of neurons throughout sensory pathways, and therefore their use provides unique opportunities for understanding sensory coding. The test dataset and, spectrograms to be predicted, consisted of spectrograms of five birdsongs, each about one minute long. These stimuli were sampled at a frequency of 44.1 kHz. The neural network was trained on another dataset consisting of 14 birdsong spectrograms, (picked randomly, of varying lengths), also sampled at a rate 44.1 kHz.

Spectrograms of birdsongs used in the model were computed using spectrogram function in Python with parameters: nfft = 128, Hanning window of length 128, and a 50% segment overlap.

Each of the spectrograms were composed of 64 frequency bins and were later downsampled to contain 32 frequency bins.

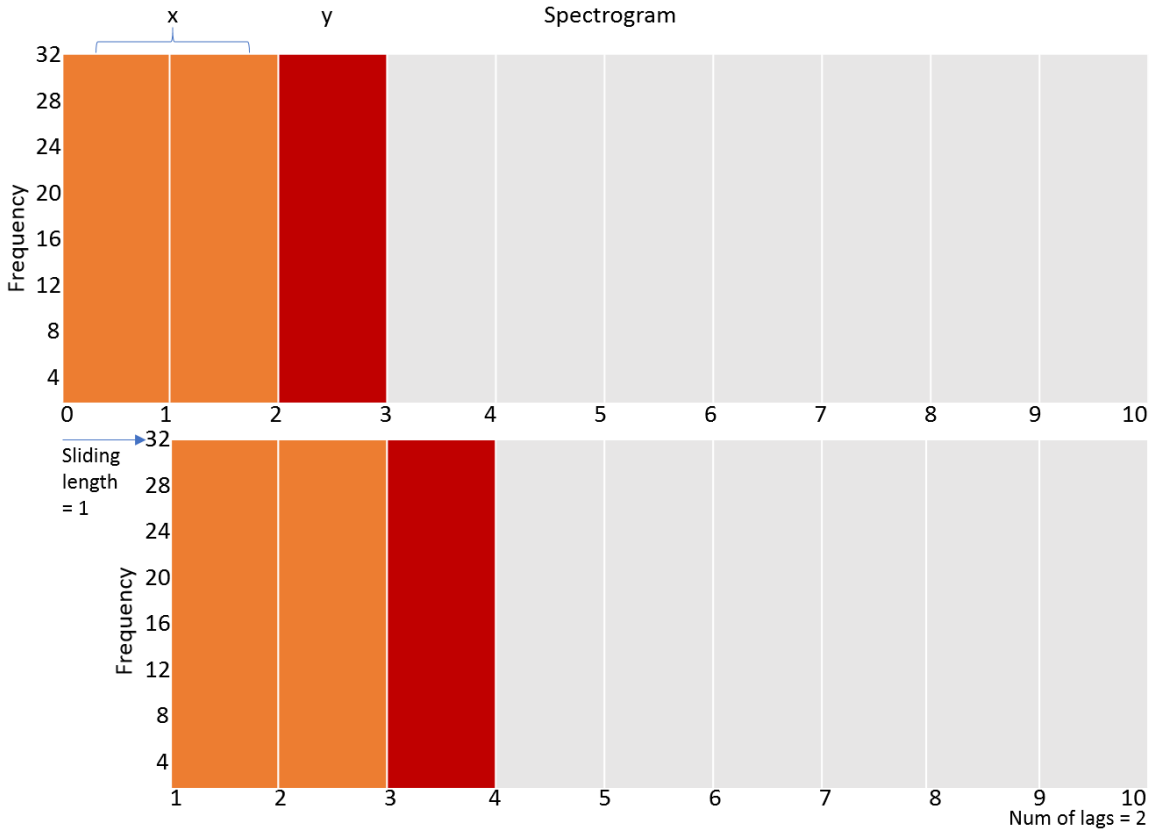

Figure 2.2. Spectrogram windowing for train and test datasets. Train and test datasets are split into input (x) and output sets (y) (no at all time bins shown in the figure). In this case, with number of lags equals 2, the first 2-time bins were considered input, and the following time bin was considered as corresponding output. With a sliding length of 1, the next 2-time bins were considered input with corresponding output. This process was continued for the entire length of spectrogram.

Datasets were segmented into input sets (x) and output sets (y). For example, for a lag number equals 2, a matrix consisting of 2-time bins along all 32 frequency bins was considered to be input x, with the vector consisting of the following time bin along all 32 frequency bins being considered as the corresponding output y. In this case, x has a shape of 2 x 32, while y has a shape of 1 x 32. Similar segmentation was done with number of lags equals 1, 2, 4, 8 and 16 (Fig 2.2).

2.4 Results

2.4.1 Loss function

The model was trained over 50 epochs with early stopping criterion (set at 2). Mean and variance of each frequency time bin was used in training. Fig 2.3 shows the negative log likelihood loss of the model.

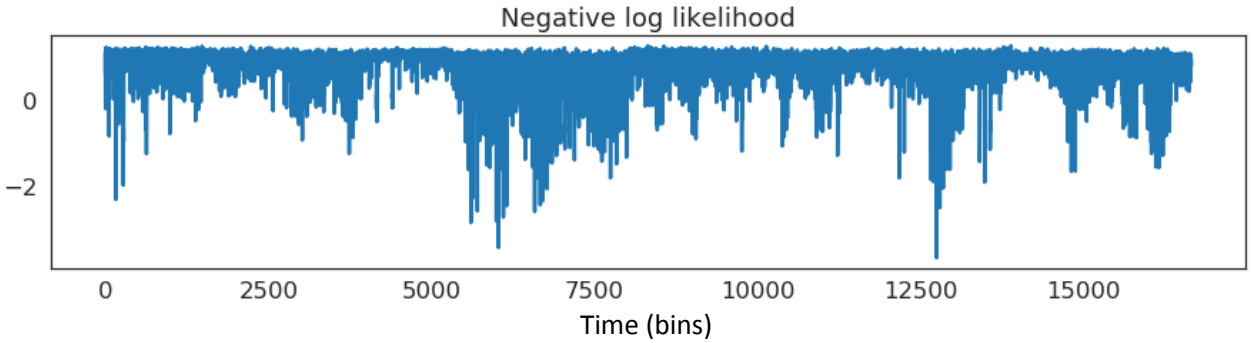

Figure 2.3. Loss function. Negative log likelihood loss of test dataset after training the model. This is used in the estimation of output value (mean).

2.4.2 Prediction with test datasets

A prediction spectrogram was estimated using the Deep Gaussian model, and an error spectrogram was computed by plotting the difference between signal and predicted spectrograms. Spectral power of each time bin was computed by the sum of spectrogram values over all frequency bins, and was plotted for signal and prediction spectrograms.

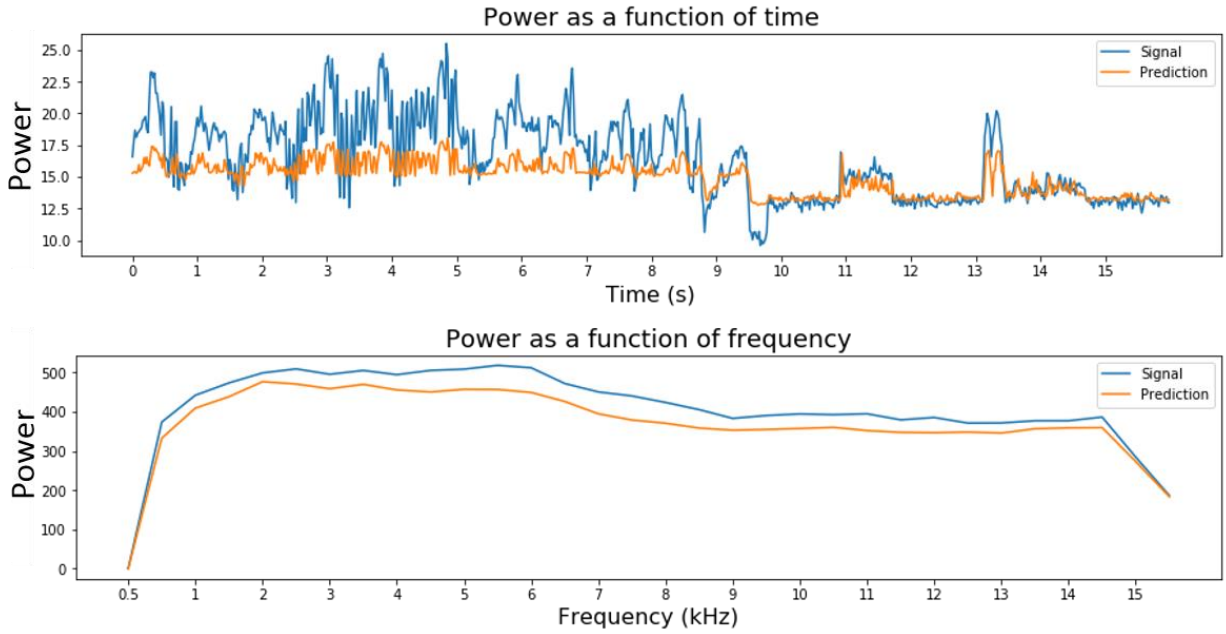

Figure 2.4. Spectral and temporal power of signal and prediction. Spectral power computed by the sum of spectrogram values for each time bin along all frequencies. The prediction spectrogram captures the temporal component of signal. The red portion of spectral power graphs is zoomed in for comparison. The blue trendline corresponds to spectral power or signal, while the orange line corresponds to prediction. The last graph compares temporal powers; blue indicates signal and orange line indicates prediction.

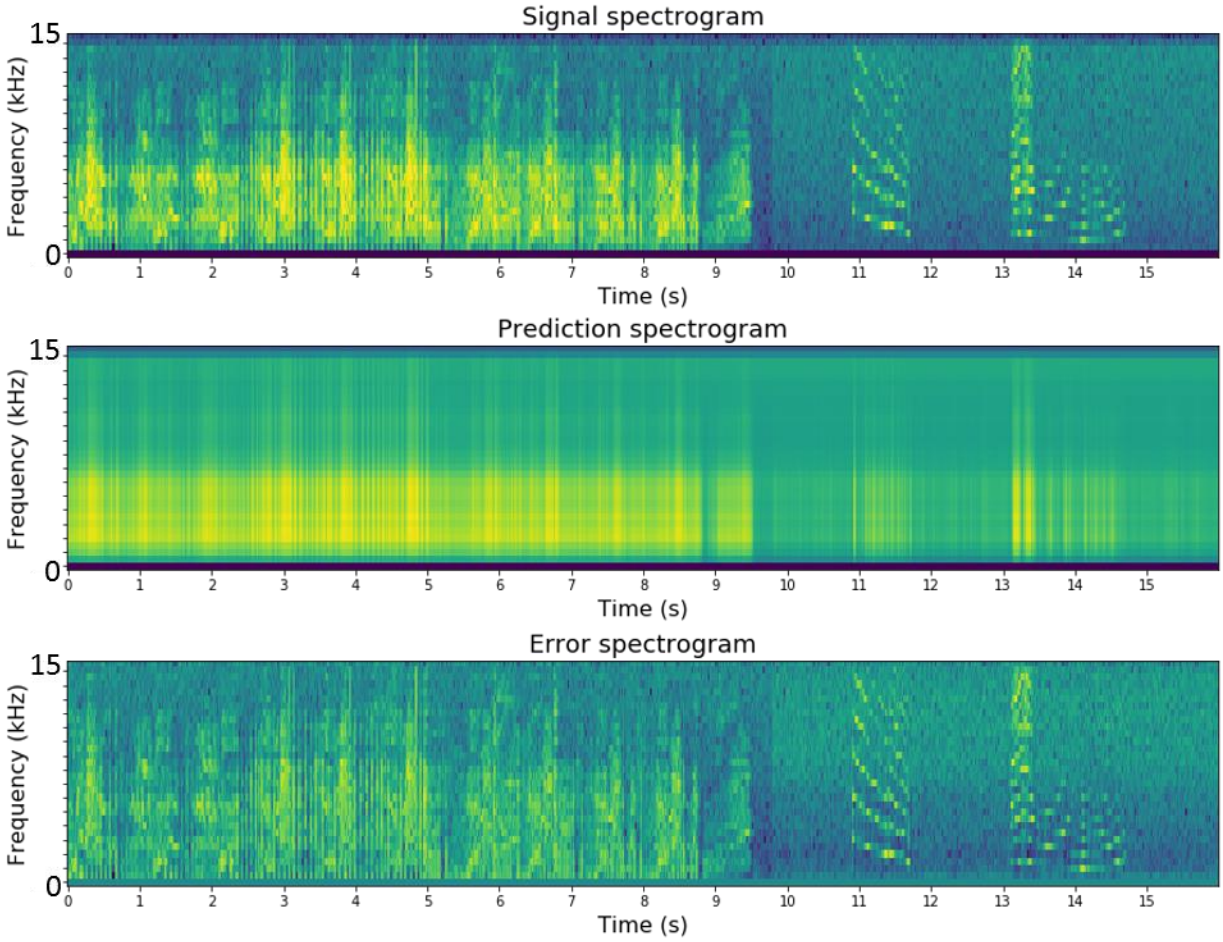

Figure 2.5. Estimated spectrograms. Signal spectrogram and estimated prediction and error spectrograms of five birdsongs (about 1 minute each) computed using the Deep Gaussian model.

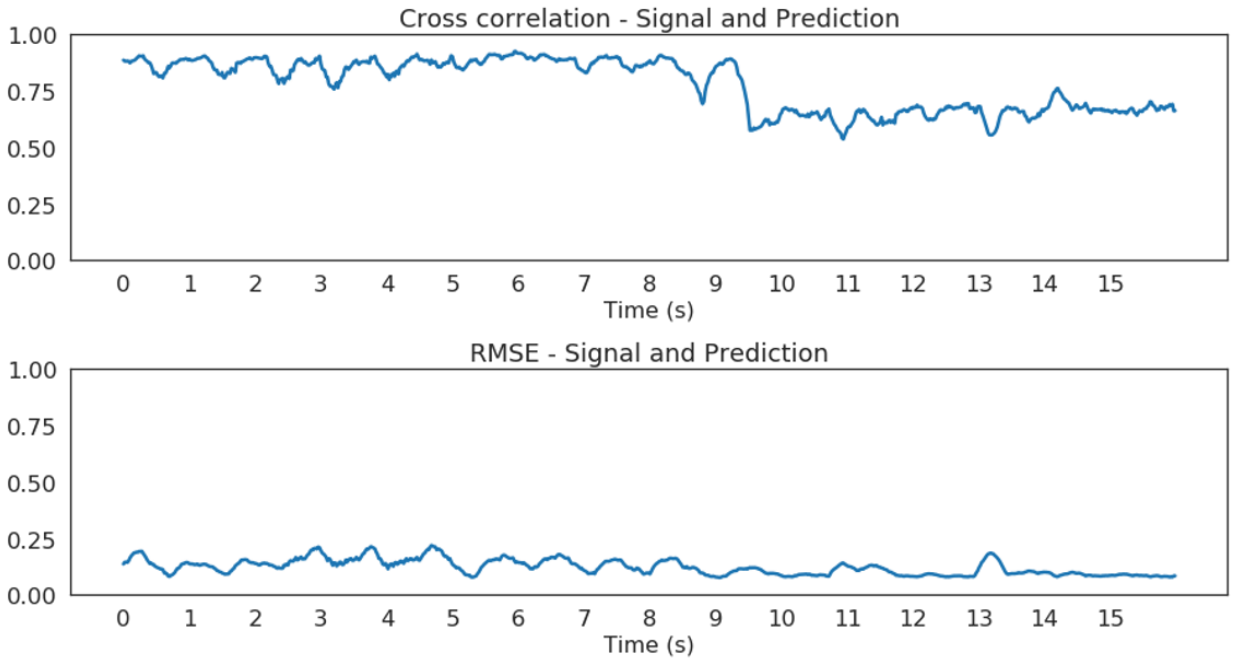

Figure 2.6. Comparison of prediction and error spectrograms to signal. Pearson correlation coefficient values computed for prediction and error spectrogram with respect to signal. The graphs are zoomed in for comparison. The blue indicates correlation of prediction and orange indicates correlation of error with respect to signal.

2.4.2 Comparison of spectrograms

In order to compare the three different representations of stimuli, cross relation values were plotted for prediction and error with respect to signal. The correlation value for each time bin was calculated by computing the correlation between the subsets of spectrograms corresponding to that time bin. *pearsonr* function from scipy was used for this calculation in Python. In addition, RMSE values for the estimated spectrograms were computed with respect to signal. RMSE values for prediction are significantly lower than corresponding values of error spectrogram. A portion of these graphs are shown for comparison. We observed that RMSE of prediction is significantly lower than RMSE of error (Fig 2.6).

Spectral power was plotted by summing all the frequency values along the corresponding time bin. The temporal power graph similarly was calculated by summing spectrogram values across the time-length of spectrogram for corresponding frequency bin. The temporal power of prediction is very close to that of signal. However, the difference in spectral power graph shows that the spectral component of the signal is not captured in our prediction (Fig 2.4).

In this chapter, we successfully observed that our predictive coding model captures the temporal component of the signal. This is critical in the case of temporal processes. The model, however, does not compute good predictions in the frequency dimension. Although there may exist other machine learning models that probably obtain a better performance, this is a good starting point to show predictive coding in a single neuron.

In the next chapter, we introduce a technique used to further compare the performance of our predictive coding model. The MNE technique will allow us to predict the response of this neuron when different stimuli are presented; signal, prediction and error in this case. We then go

on to compare the responses obtained from the model to the actual responses, to better understand what components of the signal are translated into our prediction. The Deep Gaussian predictive model was developed with the help of Marvin, a PhD student in the Gentner Lab.

3. MNE

3.1 Introduction

Selectivity of high-level neurons for processing complex, behaviorally relevant natural stimuli is poorly understood. These neurons are sensitive to conjunctions of features. To understand the principles underlying these process, receptive fields of neurons in the caudo-medial nidopallium (NCM) of the European starling (*Sturnus vulgaris*) were characterized.

Complex receptive fields are a fundamental property of the sensory system, and they map multidimensional stimuli to various behaviors. Individual neurons in the higher-level auditory cortex of starlings have composite receptive fields with several independent features. The representation of features of multidimensional natural stimuli like the starling song is captured by composite receptive fields [16].

Development of dimensionality reduction techniques like spike-triggered covariance (STC) and maximally informative dimensions (MID) have facilitated this discovery. Standard statistical tools only identify one or two features but not complete sets, whereas these techniques involve extracting a hierarchy of features in order to obtain a selective and invariant categorical representation useful for behavior. STC can identify many relevant features for stimuli whose parameters are distributed in Gaussian manner but fail when natural stimuli are used, while MID works well for arbitrary stimuli but requires exponentially larger data sets to find more than a few features [17].

Multiple distinct acoustical features can be explored in individual auditory neurons in songbirds using Maximum Noise Entropy (MNE) technique. The MNE model maximizes the noise entropy of the conditional response distribution. This statistical method is constrained by a given

set of stimuli-response correlations, but is otherwise as unbiased and random as possible. This enables a robust, statistically optimal, representation of complex, real-world signals such as birdsong, speech, or music [16].

In this chapter, the performance of composite receptive fields with respect to signal spectrograms (signal MNE), prediction spectrograms (prediction MNE) and error spectrograms (error MNE) were compared in NCM.

3.2 Receptive Field

Receptive field is a term used to describe the firing properties of the sensory neurons. Receptive fields in the auditory system are modeled as spectrotemporal patterns, which are specific patterns in the auditory domain that modulate the firing rate of a neuron.

The spectro temporal response field (STRF) of a neuron is a useful measure that represents which type of stimuli excite or inhibit a neuron. STRF is the linear characterization of the complex stimulus-response transformations seen in sensory neurons [18,19].

Linear STRFs are created by first calculating a spectrogram of the acoustic stimulus. Firing rate is modeled over time for the neuron, using a histogram combined over multiple repetitions of the acoustic stimulus. Linear regression is used to predict the firing rate of that neuron as a weighted sum of the spectrogram. The weights learned by the linear model are the STRF and represent the specific acoustic pattern that causes modulation in the firing rate of the neuron. STRFs can also be understood as the transfer functions that map acoustic stimulus input to firing rate response output. It can be generalized to capture a rich variety of nonlinear and contextual features observed in sensory neurons. It does not require prior knowledge such as frequency tuning

or threshold and is distinguished from other measures by its broader descriptive power (dynamics and spectral selectivity) [18,20].

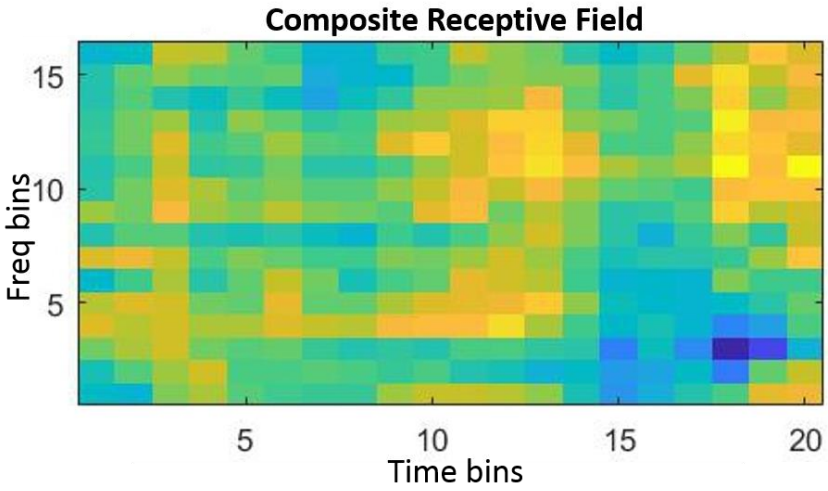

Fig 3.1. Composite receptive field of an auditory neuron. Composite receptive field of an auditory neuron in NCM of auditory cortex computed using MNE method.

3.3 MNE

3.3.1 Introduction

To minimize the bias of the system, the noise entropy was maximized subject to constraints on the stimulus/response moment. It was shown that logistic functions not only maximize noise entropy for binary outputs but provide minimum mutual information solutions when the average firing rate of a neuron is fixed. This idea was used to study single neuron coding to discover what statistics of the inputs are encoded in the outputs [21].

To begin, a system was considered which at each moment in time receives a D-dimensional input $x(t) = (x_1(t), ….., x_D(t))$ from a distribution P(x), such as a neuron receiving a sensory

stimulus or synaptic potentials. The system then performs computations to determine the output y(t) according to its response function P(y|x).

Information about the identity of the input can be obtained by observing the output quantified by the mutual information $I(y;x) = H_{resp} - H_{noise}$. The first term is the response entropy given by $H_{resp} = -\int dx P(x)$, which captures the overall uncertainty in the output. The second term is the so-called noise entropy [21].

$$H_{noise} = -\int dx P(x) \int dy P(y|x) \ln P(y|x)$$

3.3.2 Model

The stimulus features must be correlated in some way with the neural response corresponding to the spiking activity of the neuron. The specific stimulus/response correlations, such as the spike-triggered average (STA), the spike triggered covariance (STC), or the mutual information can be obtained from:

$$I(y;x) = \sum_{y} \sum_{x} P(x)P(y|x) \log \frac{P(y|x)}{P(y)}$$

This equation provides a full measure of the dependence between stimulus and response. These estimates can be used to construct a model of the conditional response probability P(y|x) by constraining to match a given set of observed correlations such as the STA and STC methods. The minimal model of P(y|x) is the one that is consistent with the chosen set of correlations but is otherwise as random as possible, making it minimally biased [17].

3.3.3 Logistic function

This model can be obtained by maximizing the noise entropy $< -\log P(y|x) >$, where <…> denotes an average over P(y,x) = P(x)P(y|x). For a binary spike/ no spike neuron consistent with an observed average firing rate, as well as the correlation of the neural response with linear and quadratic moments of the stimulus, the minimal model is the logistic function [22,17].

$$P_{min}(spike|x) = \frac{1}{1 + \exp(a + h.x + x^T Jx)}$$

3.3.4 Parameters a, h, J

The parameters a, h and J are given by the mean firing rate, experimentally observed spike-triggered average (STA) and spike-triggered covariance (STC) of the model. The relevant stimulus features can be found by diagonalizing the J matrix. The equation can include higher orders of x if correlations between a spike and higher order moments of the stimulus are measured.

The contours of constant probability of the minimal second order models are quadric surfaces, defined by the quadratic polynomial $f(x) = a + h.x + x^T Jx = constant$. The diagonalization of f(x) involves a change of coordinates such that

$$f = a + \sum_{i=1}^{D} x_i z_i + \sum_{i=1}^{D} \beta_i z_i z_i$$

This is accomplished through the diagonalization of the matrix J, yielding D eigenvectors $\{z_i\}$ with corresponding eigenvalues $\{\beta_i\}$. The eigenvectors are the principal axes of the constant probability surfaces, and the magnitude of the eigenvalue along a particular direction is indicative of the curvature, and hence the selectivity of the surface in that dimension [17,22].

3.4 Experimental setup

3.4.1 Stimuli

The stimuli were downsampled to 24 kHz and converted into spectrograms using a spectrogram function in Matlab with parameters: nfft = 128, Hanning window of length 128, and a 50% segment overlap. The DC component was removed, and the adjacent 64 frequencies were averaged pair-wise to obtain 32 frequency bands. These spectrograms were passed through a Deep Gaussian model to obtain the predicted spectrograms, also with 32 frequency bands. For all spectrograms, the adjacent frequencies were again averaged pair-wise to finally obtain 16 frequency bands ranging from 750 Hz to the Nyquist frequency (12 kHz).

The adjacent time bins were averaged three times for a final bin size of 21 ms. 20-time bins were usually used to compute MNE receptive fields. If 32 frequencies were to be used instead of 16 frequencies, a different number of time bins (10, 16, 32) can be used to compute receptive fields with similar results. All spectrograms were converted into the logarithmic scale.

3.4.2 Signal Recording

Under a protocol approved by the Institutional Animal Care and Use Committee of the University of California, San Diego, experiments were performed on adult male European starlings (*Sturnus vulgaris*). For physiological testing, birds were anesthetized (urethane, 7 ml/kg) and head-fixed to a stereotaxic apparatus mounted inside a sound attenuation box. The use of urethane was necessary to obtain the long-stimulus presentation epochs required in this study and is unlikely to alter selectivity significantly.

Songs were played to the subjects at 60-dB mean-level while we recorded action potentials extracellularly using 32-channel electrode arrays (NeuroNexus Technologies) inserted through a small craniotomy into the NCM.

Neural responses to five different 1-minute long songs were recorded, each repeated 20 times. Stimulus presentation, signal recording, and spike sorting were controlled through a PC using Spike2 software (CED). Extracellular voltage waveforms were amplified (model 3600 amplifier, A-M Systems), filtered, and sampled with a 50-μs resolution and saved for offline spike sorting.

3.4.3 Spike Sorting and identification of clusters

After preprocessing, sorting was performed on the spike events using an automated clustering approach with the novel cluster quality metrics MountainSort. Clusters were either accepted or rejected in the automatic annotation phase based on the computed cluster metrics, and then consolidated. The new, efficient, nonparametric, density-based clustering algorithm is termed ISO-SPLIT [23].

The algorithm compromises a series of nonparametric statistical tests for unimodality and makes no assumptions about the shape of clusters. It involves only a few adjustable parameters, and one essentially only needs to specify a statistical threshold for rejecting the null hypothesis for unimodality. The same set of parameters were used for all recordings. The algorithm did not need a priori information about the expected number of clusters nor the expected cluster densities.

Two metrics were used that are specifically suited for spike sorting: isolation and noise overlap. A measure of cluster signal-to-noise ratio (SNR) (large-amplitude extracellular action-

potential waveforms) was also used to exclude clusters contaminated by artifacts. Clusters were categorized into three groups: “single unit”, “noise”, and “non-isolated”.

Of all “single unit” clusters, few were chosen for further analysis based on the quality of their raster plots. These plots were computed using Python scripts. An example of a “good” raster plot is shown in the figure 3.2. This shows that an NCM neuron usually responds to a variety of motifs.

3.4.4 Response Data

Spiking data was divided into two sets for training and testing; the testing set contained one-tenth of the data. Parameters were estimated ten times, each time using a different segment of data for training and testing, and averaged. Early stopping was used for regularization to prevent overfitting. As in STC, diagonalizing the matrix J yields quadratic features with the same time and frequency dimensions as the original stimuli that drove spiking.

3.5 Results

3.5.1 Composite receptive fields

The second-order MNE model’s matrix J for each neuron together with the first-order term defined its receptive field. The composite RF describes the spectrotemporal structure in the stimulus that drives activity at a particular site. This resulting representation was used with relevant stimulus features such as STA and STC from the neural responses.

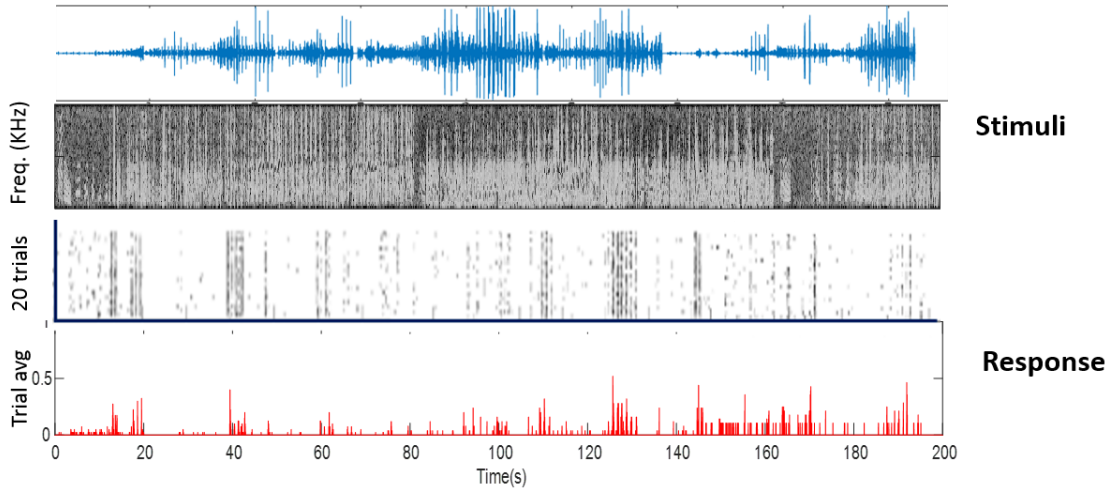

Figure 3.2. Spike raster plots. Spike raster plots showing the response of one neuron in NCM to a birdsong (used here as stimuli). On the top is the waveform, followed by spectrogram of the stimulus. The third plot shows neuron firing in 20 trials, and trial average is calculated in the last plot. This is inputted into the MNE model for training.

The STA is computed by analyzing the change in the mean between the stimulus distribution conditional on a spike and the distribution of all stimuli that were presented in the recording. STC is computed in two steps. First, the difference between the covariance matrix of all stimuli and that of stimuli that elicit a spike. Covariance is encoded in the eigenvalues of this difference matrix.

The stimulus (around 5 minutes) is jackknifed i.e. split 10 times. Training is performed on 9 subsets and responses are predicted on the remaining set not used in parameter estimation. Parameters a, h and J are estimated for each jackknifed and then averaged across all 10 jackknives. An example of the final parameters of a trained model of one jackknife for a cell with respect to signal is shown in Figure 3.3.

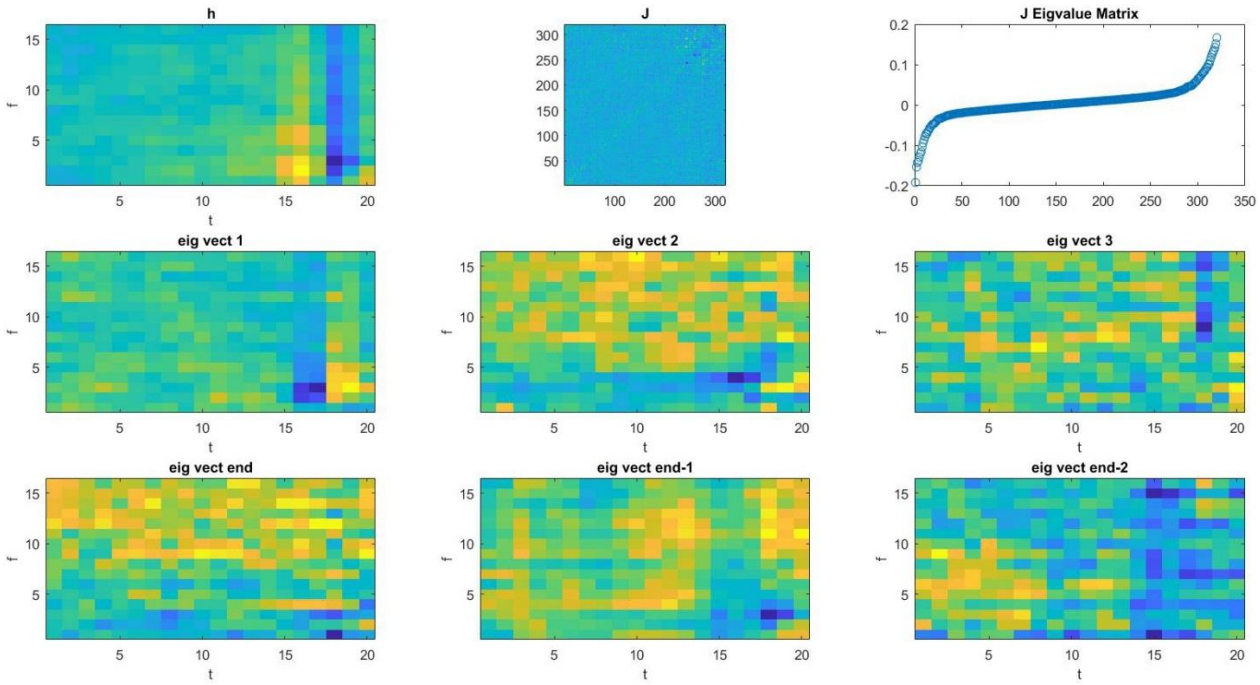

Figure 3.3. Composite receptive field of a single NCM neuron. Top row shows h and J matrices, and eigenspectrum of the matrix J for one NCM neuron. Eigenvectors were normalized for comparison with the data. Second row shows top three excitatory (negative) and last row shows top three inhibitory (positive) features obtained from the same neuron.

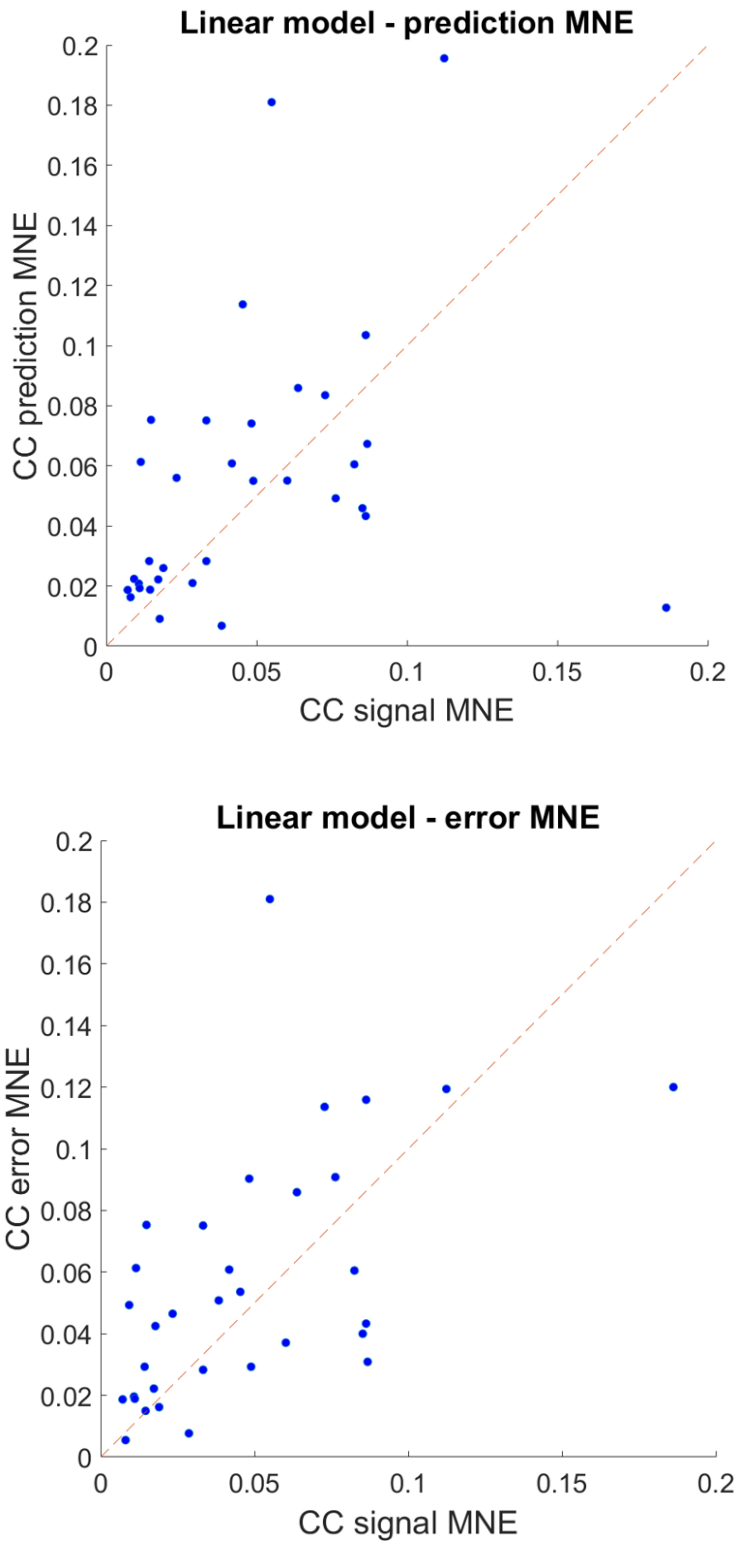

Figure 3.4. Prediction of responses to stimuli using linear MNE model. Full distribution of correlation coefficients obtained with the linear components of MNE model for prediction and error plotted against those obtained with the signal. The diagonal line indicates unity.

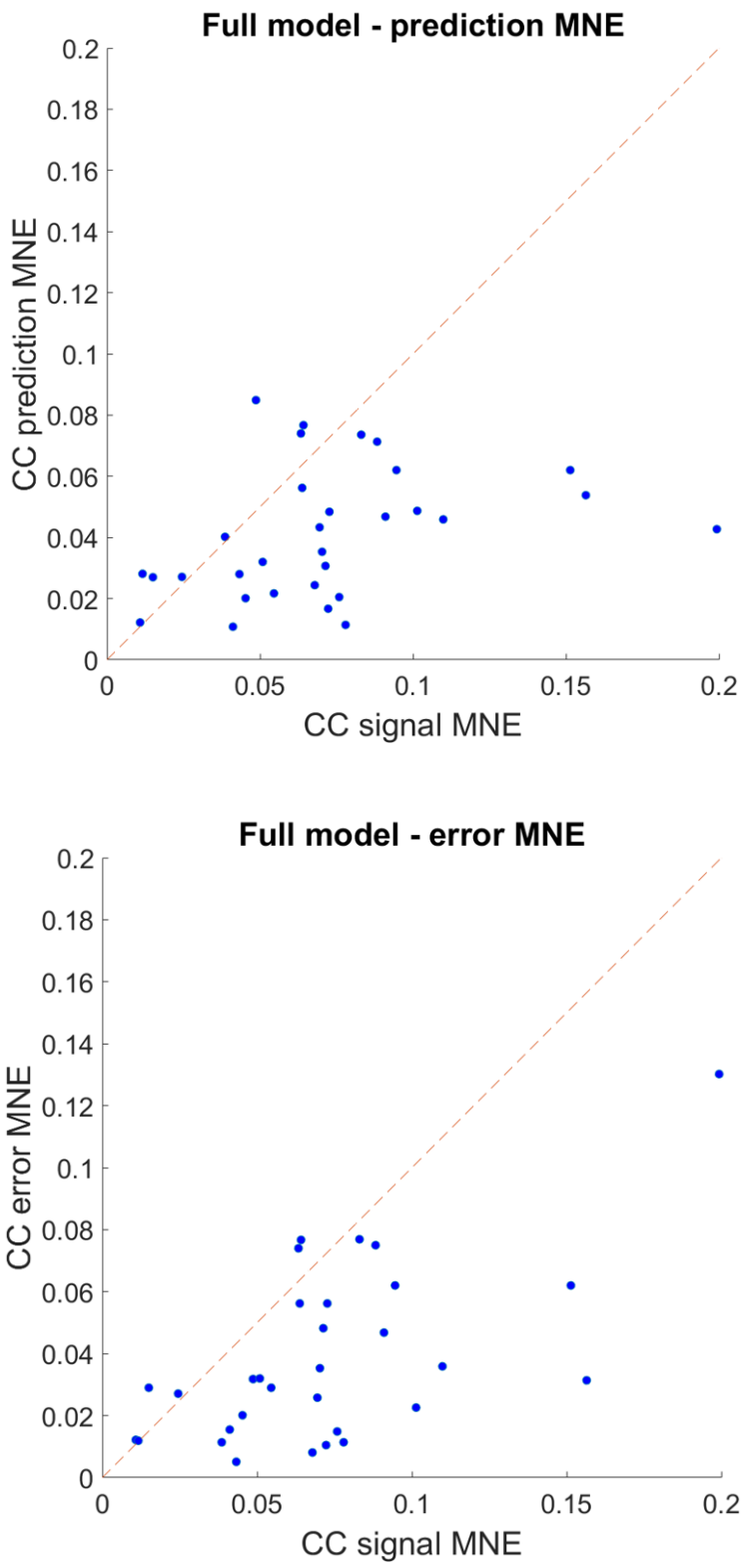

Figure 3.5. Prediction of responses to stimuli using full MNE model. Full distribution of correlation coefficients obtained with the full MNE model for prediction and error plotted against those obtained with the signal. The diagonal line indicates unity.

3.5.2 Performance of signal, prediction and error MNEs

To test the performance of predicted and error MNEs with respect to signal, we obtained the parameters a, h and J for each neuron with respect to all three MNEs. Using these parameters, response predictions were computed using a logistic function. The predictions were then compared to the actual response by estimating respective correlation coefficients (CC) using the Matlab coerrcoef function. CCs of all three MNEs were plotted to compare their predictability (Fig 3.4). In a similar way, responses were predicted considering the linear model, with parameters a and h. CCs of all three MNEs (computed from the linear) with respect to the actual response were plotted to compare their performance, shown in Figure 3.5. As observed from these figures, CCs of prediction and error MNEs are close to that of signal MNEs. Whereas for the full model, the correlation coefficients are lower compared to the signal.

The linear model gives a better predictability than the full model and the difference between these models is that J is excluded in the linear model. As J is given by the spike-triggered covariance of the model, we can conclude that the STC feature is missing in our prediction from the Deep Gaussian model. This information will be helpful in designing predictive coding models in the future. However, these results show the advantages of using lower-dimension MNEs for prediction of neural responses. The prediction and error stimuli could elicit neural activity that was well captured by the MNE.

4. Discussion and Future Directions

In this work, we present three different representations of natural stimuli, birdsong of the European starling. We call them signal, prediction and error. We used a feedforward neural network called Deep Gaussian model to estimate the prediction and error spectrograms. We hypothesize these representations capture stimulus features that are capable of eliciting a neural response comparable to the actual signal.

From the measures used to compare the three representations (spectral power, cross-correlation, and root mean squared error), it can be concluded that the predicted spectrogram captures the temporal component of the signal but fails to capture the spectral component. This is captured in the error spectrogram; however a low degree of temporal correlation is observed. The prediction spectrogram presents a better estimate of the signal confirmed from the corresponding RMSE graphs. The Deep Gaussian model, however, is a good start to explore the predictive coding hypothesis in a single neuron.

MNEs of the three different stimulus representations were computed. The parameters obtained from these models were used to predict responses and their correlation coefficients of MNEs computed estimated their response 'predictability'. The full MNE model fails to obtain good performance for prediction and error MNEs. However, the linear model shows a positive result. The correlation coefficient values of prediction and error MNEs are close to the CC values of signal in the linear model, and are much lower in the full model. This proves that the prediction and error representations of signal can elicit neural activity comparable to the signal. The spike-triggered covariance feature of the signal (captured by J) is important for prediction performance, as J is excluded in the linear model.

Thus far, we obtained different representations of the stimulus using predictive coding models. We then validated the Deep Gaussian model by showing that stimulus representations obtained from this model are successful at eliciting a neural response close to the actual response in the NCM.

There are other existing machine learning models, for example, convolutional neural networks (CNNs), long short-term memory models (LSTMs) especially useful for time-varying signals, combinations of both, generative models like contrastive predictive coding (CPC) [35] that can produce better predictions compared to our model. Our limited knowledge of predictive coding in other areas of the auditory cortex prevents us from drawing strong conclusions about the existence and performance of our model or other PC models in those areas. This calls for more experiments in areas such as CM or areas upstream in the auditory process that have more prominent neuron firing like Field L, since only through comparing different models in each area can we ultimately confirm or falsify the organization of appropriate PC models in the auditory cortex.

The next steps for this project, which are currently underway, involve exploring other predictive coding models that can give a more accurate prediction of the stimulus. Models such as CNNs, which are primarily used with visual stimulus have failed. Models used for image processing cannot be used for audio as: 1) discrete sound events do not separate on a spectrogram. Instead, they sum together into a distinct whole event; 2) axes of a spectrogram are fundamentally different. Offsetting along each axis has a different meaning; 3) the spectral properties of sound are non-local; and 4) while images can be regarded to contain large amounts of parallel information, sound is highly serial.

Models that take the temporal aspect into consideration, such as LSTMs and generative models like contrastive predictive coding, will be developed. Performance of multiple models will be compared to estimate the model that gives better prediction in the NCM. Finally, this work will then be extended to other areas such as CM and Field L, to map prediction models to different regions in the higher-level auditory cortex.

REFERENCES

[1] Adams, R. A., Shipp, S. & Friston, K. J. Predictions not commands : active inference in the motor system. 611–643 (2013). doi:10.1007/s00429-012-0475-5

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

[3] Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschteind, T., Naccache, L., and Dehaene, S. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci.* **108,** 20754–20759 (2011).

[4] Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2,** 79–87 (1999).

[5] Helekar, S. A. *Animal Models of Speech and Language Disorders*.

[6] Gentner, T. Q. & Margoliash, D. Neuronal populations and single cells representing learned auditory objects. 669–674 (2003).

[7] Amin, N. & Doupe, A. Development of Selectivity for Natural Sounds in the Songbird Auditory Forebrain. 3517–3531 (2007). doi:10.1152/jn.01066.2006.

[8] Theunissen, R. I. C. E., Grace, J. A., Amin, N. & Singh, N. C. Selectivity for Conspecific Song in the Zebra Finch Auditory Forebrain. 472–487 (2019).

[9] Mumford, D. cybernetics. **251,** 241–251 (1992).

[10] Ulanovsky, N. Multiple Time Scales of Adaptation in Auditory Cortex Neurons. *J. Neurosci.* **24,** 10440–10453 (2004).

[11] Ulanovsky, N., Las, L. & Nelken, I. Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* **6,** 391–398 (2003).

[12] Rubin, J., Ulanovsky, N., Nelken, I. & Tishby, N. The Representation of Prediction Error in Auditory Cortex. 1–28 (2016). doi:10.5061/dryad.3m5v5

[13] Gill, P., Woolley, S. M. N., Fremouw, T. & Theunissen, E. What ' s That Sound ? Auditory Area CLM Encodes Stimulus Surprise , Not Intensity or Intensity Changes. 2809–2820 (2018). doi:10.1152/jn.01270.2007.

[14] Pritzel, A. & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. (2017).

[15] Nix, D. A. & Weigend, A. S. Estimating the. 55–60 (1901).

[16] Kozlov, A. S. & Gentner, T. Central auditory neurons have composite receptive fields. *Proc. Natl. Acad. Sci.* **113,** 1441–1446 (2016).

[17] Fitzgerald, J. D., Rowekamp, R. J., Sincich, L. C. & Sharpee, T. O. Second Order Dimensionality Reduction Using Minimum and Maximum Mutual Information Models. **7,** 1–9 (2011).

[18] Auditory, C. & Fields, R. Characterizing Auditory Receptive Fields. 829–831 (2008). doi:10.1016/j.neuron.2008.06.004

[19] Theunissen, E., Sen, K. & Doupe, A. J. Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons. **20,** 2315–2331 (2000).

[20] Theunissen, F. E., David, S.V., Singh, N.C., Hsu, A., Vinje, W.E. & Gallant, J.L. Network : Computation in Neural Systems Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli Estimating spatio-temporal receptive fields of auditory. **6536,** (2009).

[21] Fitzgerald, J. D., Sincich, L. C. & Sharpee, T. O. Minimal Models of Multidimensional Computations. **7,** (2011).

[22] Sharpee, T. O. Computational Identification of Receptive Fields. *Annu. Rev. Neurosci.* **36,** 103–120 (2013).

[23] Chung, J. E., Magland, J. F., Barnett, A. H., Tolosa, V. M., Tooker, A. C., Lee, K. Y., Shah, K. G., Felix, S. H., Frank, L. M. and Greengard, L. F. Article A Fully Automated Approach to Spike Sorting Article A Fully Automated Approach to Spike Sorting. *Neuron* **95,** 1381–1394.e6 (2017).

[24] Woolley, S. M. N. The Songbird Auditory System. (2016). doi:10.1007/978-1-4614-8400-4

[25] Nottebohm, F. The Neural Basis of Birdsong. **3,** 759–761 (2005).

[26] Brainard, M. S. & Doupe, A. J. AUDITORY FEEDBACK IN VOCAL BEHAVIOUR. **1,** 1–10 (2000).

[27] Manuscript, A. NIH Public Access. 489–517 (2014). doi:10.1146/annurev-neuro-060909-152826.Translating

[28] Bendixen, A. & Schro, E. I Heard That Coming : Event-Related Potential Evidence for Stimulus-Driven Prediction in the Auditory System. **29,** 8447–8451 (2009).

[29] Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A. and Henson, R. Silent Expectations : Dynamic Causal Modeling of Cortical Prediction and Attention to Sounds That Weren ' t. **36,** 8305–8316 (2016).

[30] Hughes, H. C., Darcey, T. M., Barkan, H. I., Williamson, P. D., Roberts, D. W. and Aslin C. H. Responses of Human Auditory Association Cortex to the Omission of an Expected Acoustic Event. **1089,** 1073–1089 (2001).

[31] Neuroscience, H., Sanmiguel, I., Saupe, K. & Schröger, E. I know what is missing here : electrophysiological prediction error signals elicited by omissions of predicted ” what ” but not ” when ”. **7,** 1–10 (2013).

[32] Sanmiguel, I., Widmann, A., Bendixen, A., Trujillo-barreto, N. & Schro, E. Hearing Silences : Human Auditory Processing Relies on Preactivation of Sound-Specific Brain Activity Patterns. **33,** 8633–8639 (2013).

[33] Todorovic, A., Ede, F. Van, Maris, E. & Lange, F. P. De. Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex : An MEG Study. **31,** 9118–9123 (2011).

[34] Todorovic, A. & Lange, F. P. De. Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. **32,** 13389–13395 (2012).

[35] Oord, A. Van Den. Representation Learning with Contrastive Predictive Coding.