Decoding the Rhythms of Avian Auditory LFP

by

Michael J. Schachter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Dr. Frederic Theunissen, Chair
Dr. Fritz Sommer
Dr. Michael DeWeese
Dr. Joan Bruna

Summer 2016

The dissertation of Michael J. Schachter, titled Decoding the Rhythms of Avian Auditory LFP, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Decoding the Rhythms of Avian Auditory LFP

Abstract

Decoding the Rhythms of Avian Auditory LFP

by

Michael J. Schachter

Doctor of Philosophy in Biophysics

University of California, Berkeley

Dr. Frederic Theunissen, Chair

We undertook a detailed analysis of population spike rate and LFP power in the Zebra finch auditory system. Utilizing the full range of Zebra finch vocalizations and dual-hemisphere multielectrode recordings from auditory neurons, we show how intuitive acoustic features such as amplitude, spectral distribution statistics and pitch drive the spike rate of individual neurons and LFP power. Using decoding approaches, we show that these acoustic features can be successfully decoded from the population spike rate vector and the power spectra of the multielectrode LFP, and that multielectrode LFP outperforms spike rate. We find that adding pairwise spike synchrony terms to the spike rate decoder boosts performance to be on par with the LFP power spectra, and that decoder performance grows quickly with the addition of more neurons but quickly saturates. Finally, we demonstrate through predictive modeling that LFP power on an electrode is a mix of local spike rate and spike synchrony.

Acknowledgments

Without X, I would have never done Y, you Z?

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The nature of information encoded by auditory networks in the brain has been described by a variety of experimental approaches that vary in their choice of stimuli, stimulus representation, and predictive modelling approach. Neurons in the auditory system have been probed with simple stimuli such as tones, as well as natural stimuli that tend to elicit more robust neural responses. There is a spectrum of stimulus paradigms and stimulus-response models that can be constructed to better understand the relationship between the properties of sound and the spiking of auditory neurons. These models depend in large part on the richness of the stimulus and the numerical representation used to describe it. At the simple end of the spectrum are artificial pure tones, which can be quantified completely by their amplitude and frequency. Neuron response properties have been described using tuning curves that predict spike rate from the amplitude and frequency of simple tone stimuli. These models have been used with some success to describe neuronal response properties in early auditory areas [**REF**], and even to describe tonotopy in human auditory cortex [**REF**]. However, natural stimuli are not simply described by their amplitude and frequency. Human speech is a variable and complex sequence of smoothly changing harmonic stacks and noisy bursts [**REF**]. Some bird vocalizations share a similar complexity; Zebra Finch songs are complex but rigid sequence of harmonic stacks, noise bursts, and chirps [**REF**]. The need to utilize natural sound stimuli to more effectively probe neuron responses necessitates a stimulus representation more complex than amplitude and frequency alone. Complete information about the time-varying acoustic features of a sound can be quantified using a spectrogram. Spectrograms represent the sound as a set of frequencies that vary over time, and can be inverted to produce the original sound pressure waveform [3]. To accommodate this complex representation, spatio-temporal receptive fields (STRFs) can be fit to describe neuronal responses to natural sounds using a weighted sum of the recent spectrogram history [**REF**]. STRFs have been used with much success to describe auditory neurons both in mammals [**REF**] and birds [**REF**]. Notably, tonotopy has not been observed in higher Avian auditory areas; in its place is a STRFotopy, where temporal memory and spectral bandwidth of neuronal responses vary over anatomical space [12]. At an intermediate level of representation, natural sounds that are short and isolated in time can be represented by a small set of summary statis-

tics that intuitively describe how they vary spectrally, temporally, and spectro-temporally [**REF**]. This approach has been utilized to successfully classify the behavioral context and semantic meaning of Zebra finch vocalizations [6]. We chose the acoustic feature set utilized by [6] to represent Zebra finch vocalizations, and described the relationship between these acoustic features and neuron activity. These stimulus-response models describe neuronal activity as a function not just of amplitude and frequency, but a richer set of features closer to perceptual properties such as pitch and spectral or temporal noisiness. It is unlikely that a clear understanding of the brain can be obtained by the study of single neurons alone. The successful adoption and use of multi-electrode arrays allows Neuroscientists to simultaneously record from many neurons distributed over a large area. By utilizing data from ensembles of auditory neurons presented with natural sounds, we can develop insight into how neurons work together as a population to represent stimulus information. A core observation that sets the context for understanding population coding is that neurons integrate input from many other neurons, and temporally coincident input from multiple input neurons drives stronger spiking activity than non-coincident input [**REF**]. This implies that stimulus information may be encoded and transmitted not only by the idiosyncratic firing of individual neurons, but in addition by the temporal correlations of network firing patterns. Approaches to understand the population code at this level have utilized Information Theory to quantify the amount of stimulus information contained in an ensemble of neurons, as well as Machine Learning approaches to directly decode stimulus features [17]. These approaches have led to a deeper understanding of the population code. There is evidence that neurons exhibit robust spatial correlations in their spike patterns. Analysis of retinal ganglion cell activity by [15] and [9], using random flicker and white noise visual stimuli, respectively, demonstrated the existence of significant pairwise correlations between firing neurons. [14] showed pairwise correlations between V1 neurons using sinusoidal grating stimuli. In the auditory system, [18] showed that pairwise connectivity between neurons in mouse auditory cortex could be modulated by optogenetic activation of inhibitory interneurons. However, the existence of correlated activity does not imply that correlations actually carry stimulus information. Information theoretic frameworks have been constructed to analyze the stimulus information carried by ensembles of neurons independently by their spike rates, and in addition their correlations ([20], [8], [19]). Complementary decoding approaches can be used to disentangle the effects of redundancy and synergy in correlations. Using a decoding approach, [2] show in monkey auditory cortex that the ensemble spike rates of neurons contain non-redundant information about sound stimuli, and decoding performance increases with the number of neurons considered. They found that there is a small group of neurons that contain most of the stimulus information. Following up with an information theoretic approach, they found that correlations in neural activity do not contain stimulus information. In this work we show that including correlated spiking activity in addition to population spike rate improves the performance of decoders trained to predict acoustic features. Electrical activity from the synapses and membrane currents of many neurons contributes to an aggregate signal known as the local field potential (LFP). The LFP is a complex mixture of synaptic and transmembrane currents elicited by sodium and calcium spikes [10], and the

biophysical origin of LFP power may vary by frequency [22]. Many studies show that the mammalian LFP oscillates at several different frequency bands. Very low frequency (¡ 2Hz) slow oscillations, observed during sleep and some types of anesthesia, may originate from the interplay of bursting neurons in the Thalamic Reticular Nucleus and cortex [4]. Oscillations in the range of 30-90Hz are typically labeled as Gamma oscillations. The neural mechanism of Gamma oscillations is thought to involve the spatial and temporal interplay between excitatory and inhibitory networks [11]. Activity in different frequency bands is not mutually exclusive; lower frequency Theta oscillations (7Hz) can modulate higher frequency Gamma oscillations in the Hippocampus in a manner that may help encode ordered sequence of items [7]. A nested hierarchy of frequency bands has been identified in auditory cortex of monkey that controls the excitability of neural activity and may optimize the auditory system for the processing of rhythmic vocalizations [16]. In contrast to the well studied oscillations of mammalian cortex, there have not been many studies of LFP oscillations in the Avian brain. Analysis of multielectrode LFP was used by [13] to show three dimensional propagation of slow wave oscillations in Zebra finch forebrain, but higher frequencies were not studied, and they did not link this activity to sensory stimuli. In this work, we study LFP power in the 15-190Hz range of the Zebra finch auditory system, and show that the power spectrum can be used to decode acoustic features from the full repertoire of natural Zebra finch vocalizations. We then show that LFP power at each frequency can be predicted in large part from the population spike activity and zero-lag correlations of spiking of neurons in the auditory network.

Chapter 2

Predicting Spikes and LFP from Acoustic Features

2.1 Acoustic Features Covary and Cluster

Our goal was to describe the relationship between Zebra finch vocalizations, neuronal spiking, and the LFP. We segmented Zebra finch vocalizations and quantified them using a rich set of acoustic features. Syllables used in the analysis ranged in duration from 40ms to 400ms, which required us to quantify them so that the dimensionality of the feature vector that described a syllable was independent of duration. To achieve this, we quantified the statistical properties of the syllable power spectrum, amplitude envelope, and the time-varying fundamental, detailed in Methods - Acoustic Features. This produced a unique 19 dimensional feature vector for each syllable. In Figure 2, we show an example of the acoustic feature characterization for a single syllable, as well as examples of syllables that span the range of maximum amplitudes (Max A), mean spectral frequencies (Mean S) and saliencies (Saliency). Saliency is a measure of syllable pitchiness, low for noisy syllables and high for harmonic-stack-like syllables. These features, among several others, have been shown to be vital for determining the behavioral context, and hence semantic meaning, of Zebra finch vocalizations [6]. Acoustic features are intuitive quantities for describing syllables, but are not completely independent of each other. By construction, they naturally fall into three groups - those that describe the spectral distribution, the temporal distribution, and the time-varying fundamental frequency. Figure 3a shows a matrix of correlation coefficients between each acoustic feature. The features are ordered according to constructed group, but also naturally fall into several groups given the block-diagonal structure of the correlation matrix. Figure 3b shows a manually organized graphical representation of acoustic feature relationships. Edge thickness depicts the absolute value of the correlation coefficient, and coefficients less than 0.20 are not shown. Taken together, the correlation matrix and graph show that acoustic features cluster into several groups. The time-varying fundamental features form one group (green in Figure 3b), with the inharmonic fundamental parameters Pk

2 and 2nd V forming a distinct subgroup. Features that describe fundamental frequency over time, the mean (Mean F0), max (Max F0), min (Min F0), and coefficient of variation (CV F0) are strongly correlated with each other. Purely spectral features, statistics computed from the power spectrum of the syllable, form another group (orange in Figure 3b). The mean spectral frequency (Mean S), 25th, 50th, and 75th percentiles of the spectral distribution (Q1, Q2, Q3, respectively), and the spectral skew formed a strongly correlated subgroup. The spectral kurtosis (Kurt S) was correlated to spectral standard deviation (Std T). The spectral entropy Ent S, a measure of inharmonicity, was strongly negatively correlated with Saliency - harmonic stack like syllables have low spectral entropies and high saliency, while noisy syllables have high entropies and low saliency. Both saliency and spectral entropy were correlated with spectral standard deviation and kurtosis. Quantities that describe purely temporal features were computed from the amplitude envelope and formed the last group (blue in Figure 3b). The mean and standard deviation of the amplitude envelope (not shown), were linearly proportional to syllable duration, with a correlation coefficient of 0.97. The entropy of the amplitude envelope, temporal entropy (Ent T), was strongly negatively correlated with maximum amplitude (Max A) - syllables with amplitude envelopes with high variation, such as Begging and Nest calls, also tended to have lower maximum amplitudes. Temporal skew (Skew T) and kurtosis (Kurt T) were correlated with each other but not much with other features.

2.2 Spikes and LFP Driven by Amplitude, Temporal Entropy, Mean Spectral Frequency, Saliency

We used an encoder approach to understand how acoustic features drive spike rate and LFP power. Figure 4 shows the isolation and extraction of features for syllables and the LFP. Acoustic feature extraction was described in the previous section. A syllable was isolated (Figure 4a), and the multi-electrode spike trains and LFP were taken for each of the ten trials the syllable was presented for (Figure 4c). The spike rate was computed for each trial, and averaged across trials. The power spectrum was computed from the LFP on each electrode for each trial, and the power spectra were averaged across trials to produce multi-electrode power spectra (Figure 4d). Performance of encoders and decoders for the LFP, described shortly, were contingent on first taking the log of the power spectra, and then z-scoring within electrode and frequency. To quantify the univariate relationship between spike rates and acoustic features, and LFP power and acoustic features, we fit nonlinear tuning curves that estimated spike rate on a neuron or LFP power at a given electrode and frequency from each individual acoustic feature (Methods - Tuning Curve and Generalized Additive Encoder). Figure 5a shows example tuning curves for several acoustic features. The tuning curves illustrate that there are several basic relationships between acoustic features and spiking or LFP power. The top row shows example high performing tuning curves that predict neural response from syllable maximum amplitude (Max A). Strikingly, the

neural response to amplitude is bimodal, some neurons respond to increases in amplitude by increasing their spike rate, while others decrease their spike rate. In contrast, the LFP power spectra on various electrodes did not exhibit the same bimodal relationship at low frequencies (0-16Hz and 33-49Hz bands), but did at high frequencies (165-182Hz band). The second row of Figure 5a illustrates that the relationship between spike rate or LFP power and mean spectral frequency is multimodal and nonlinear. The tuning curves typically had a single peak (also called a best frequency) for spectral mean frequencies greater 2kHz. For neurons whose tuning curves predicted spike rate with a cross-validated R2 of 0.05 or above (n=367), 25% had a best frequency less than 2.5kHz, and declined in spike rate as mean spectral frequency increased. The largest fraction of neurons had a best frequency between 2.5kHz and 3.5kHz (63%), and a small fraction (0.11%) had high best frequencies from 3.5kHz to 5kHz. The distribution of best frequencies had three peaks, one for neurons with low best frequencies (the lowest mean spectral frequency was around 2kHz), another peak at roughly 3kHz, and a third relatively small peak around 4kHz. Examination of the mean spectral frequency tuning curves for LFP power shows that for some electrodes, low frequencies (16-33Hz band) and high frequencies (165-182Hz band) exhibit multiple best frequencies as well. Temporal entropy (Ent T) measures the disorder of the syllable amplitude envelope, and syllables with high temporal entropy elicit more spikes from neurons as shown in the third row of Figure 6a. LFP power had a multi-peaked relationship with temporal entropy. A similar relationship holds for pitch saliency (Saliency, last row). Syllables with high saliency, such as distance calls, are comprised of well-defined harmonic stacks and elicit a stronger spiking response than syllables with low pitch saliency, such as begging calls. LFP power exhibits a similar relationship, as saliency goes up, LFP power increases. The relationship between saliency and spike rate or LFP power was also multi-peaked for some neurons/electrodes. The average performance for spike rate and LFP power at each frequency, for all acoustic features, is shown in Figure 5b. The highest performing tuning curves were for maximum amplitude (Max A), and they were most predictive for frequencies from 0-49Hz. Maximum amplitude was followed by highly correlated spectral features; mean spectral frequency (Mean S) was the second best predictor of activity, followed by spectral quartiles Q2 and Q1. Temporal entropy (Ent T) was also predictive of spike rate and LFP power, as well as the second voice feature (2nd V), a measure of syllable inharmonicity. With the exception of maximum amplitude and second voice, tuning curves that predicted spike rate well did best for predicting the 33-49Hz band. After fitting the tuning curves, we used them as input nonlinearities to a Generalized Additive Model that tried to predict spike rate or LFP power from all the nonlinearly mapped acoustic features. The performance for the encoders across frequencies for LFP power and for spike rate are shown in Figure 6c. To quantify the relationship between frequency, anatomical region, and multivariate encoder performace, a linear model was fit in R to predict the multivariate encoder R2 from frequency and region. Multivariate encoder performance was not highly predictable using this linear model (R2=0.06), but frequency bands had statistically significant differences (Table 1). Frequency bands from 16-49Hz had the highest relative predictability, with less predictability for higher frequencies. The only anatomical region that had a significant difference in predictability from other regions was

NCM, which had a slightly lower predictability than CM and Field L1, L2, L3.

Chapter 3

Decoding Acoustic Features from the Neural Ensemble

3.1 Regional Specificity in Decoding Performance

The Zebra finch auditory system is not a homogenous structure. There is some evidence that it is anatomically layered in a way homologous to mammalian cortex [Wang2010]. A detailed analysis of regional specificity by [Meliza2012] showed that regions L2 and L1 were the least selective and tolerant, responding to most acoustic stimuli in a way that is not invariant to slight changes in acoustic features, while regions NCM and L3 were the most selective and tolerant. [Elie2015a] analyzed the decoding performance of call type using the same dataset analyzed in this work, and found regional differences as well. They found that regions L3 and CM were the best at classifying Distance Calls and Field L were the best at classifying song. They also found that regions L3 and CMM was the most invariant (or tolerant) of variation in Distance calls. The panels in Figure 7a show single electrode decoder performance across space, for all electrodes across recording sites, with electrodes from the two hemispheres plotted together as a function of their distance from the midline along the medial-lateral axis, and their rostral-caudal distance from region L2A. Figure 7b shows single electrode decoder performance averaged within acoustic feature and region. A linear model was fit in R to predict single electrode decoder performance from the interaction between acoustic feature and region. Table 3 shows the effect sizes of the interactions for this model. The panels in Figure 7a show that maximum amplitude has some region specificity, with electrodes in region L2 containing the most amplitude information. Mean spectral frequency also exhibited some regional specificity, being decoded best from regions L2 and L3. The coefficient of variation for the fundamental frequency (CV F0), which would be low for syllables that do not vary much spectrotemporally, like female distance calls, was also decoded best from region L3. Saliency and temporal entropy do not exhibit much regional specificity. Further examination of Table 3 shows that decoders trained on electrodes in region L3 outperform other regions for most spectral features.

3.2 Ensemble Decoding Performs Best for Amplitude and Spectral Features

We built decoders to predict each individual acoustic feature from population activity, encoded by the population spike rate vector, LFP power spectra, or population spike rate vector and in addition, pairwise spike synchrony (Methods - Encoder and Decoder Dataset Construction). Figure 6a shows the mean performance by neural response type and acoustic feature. Visualization makes it clear that maximum amplitude is decoded best from the population data, and also that LFP power spectra outperform population spike rate in decoding performance. Notably, the addition of pairwise synchrony terms to the population spike rate boosts decoder performance to that of the LFP power spectra. We fit a linear model in R to determine the predictability of an acoustic feature as a function of the interaction between acoustic feature and the neural representation used. Table 2 shows the effect size of the interaction terms, demonstrating quantitatively that amplitude and spectral features are best decoded from the population data. We further investigated how decoder performance increased as a function of number of electrodes. To do this, we first merged electrodes from each hemisphere for each recording site, giving us potentially 32 electrodes to predict from. Then we ran a decoder for a variety of combinations of electrodes for a fixed number of electrodes (Methods - Ensemble Decoding Analysis). The results are shown in Figure 7b and 7c for the best decoded acoustic features. Common trends occur for across features. When decoding from population spike rate, decoder performance increased sharply from 1 to 10 neurons. For the LFP power spectra, decoder performance increased as sharply from one to four electrodes. After that point, the slope remained positive, and adding more neurons or electrodes gradually improved decoder performance. For each site, maximum decoder performance was reached using all neurons or electrodes. We conclude that there is much more information in multi-electrode representations than single electrodes.

Chapter 4

The Relationship between Spikes and LFP

4.1 LFP Power is a Mix of Local Population Spike Rate and Synchrony

We attempted to quantify the relationship between population spike rate, spike synchrony, and the LFP as a function of frequency. To do this, we built an encoder that attempted to predict power at a given frequency band for a given electrode from the population spike rate, and in addition spike synchrony (see Methods - Population Spike Rate and Spike Synchrony and Methods - Spike Rate to LFP Power Encoder). Figure 8a shows that on average, the LFP power can be well predicted by population spike rate, and adding spike synchrony terms boost performance for frequency bands from 45-165Hz. The figure also shows variation in performance by frequency. To quantify this, and to explore region specificity, we fit a linear model in R to predict encoder performance from anatomical region, plus the interaction of frequency and population representation type (Spike Rate or Spike Rate + Synchrony). The effect sizes for the interaction terms in the model are shown in Table 4, effect sizes for regions are provided in the figure legend. LFP power was best predicted for Field L regions, and was not as well predicted for CM and NCM. The LFP is typically described as the summed local electrical activity on an electrode. We investigated how much of an effect neurons had on LFP power as a function of distance from the electrode. To do this, we utilized the weights of the encoder trained to predict LFP power from spike rate. Each neuron, with its associated decoder weight, was a given distance from the electrode whose LFP was being predicted. Figure 8b shows smoothed squared encoder weights as a function of distance. The weight-squared are shown for the 33-49Hz band and the 165-182Hz band. Weights-squared for the higher frequency band fall off quickly with distance, while they fall off less sharply for 33Hz, and even begin to increase at long range distances. The inset of Figure 8b shows the average encoder weights-squared for neurons on the same electrode, vs neurons on a different electrode. Neurons on the same electrode contribute much more to

LFP power, an order of magnitude more, than neurons on other electrodes. From this, we conclude that the LFP power is comprised predominantly of local spike rate and synchrony.

Chapter 5

Going Recurrent: Predicting the Time-varying LFP

Chapter 6

Discussion

In this work we have shown that Zebra finch syllables can be quantified by their acoustic features in a duration-independent fashion, that some of these acoustic features, mainly amplitude, spectral distribution statistics, pitch saliency, and temporal entropy, can be used to predict both spike rate and LFP power. We showed that these features can be decoded from the spike rate vector of a population of neurons, and that the decoding performance grows as more neurons are utilized. We showed that the power spectrum of the LFP in the Zebra finch auditory system contains a significant amount of information about the acoustic features of vocalizations, and that regional differences exist in the type of information decoded. Training encoders enabled us to say what causally drove spike rate or LFP power on individual neurons or electrodes [21], while decoders enabled us to determine the types of information that could be successfully extracted from population activity (Figure 6). The acoustic features that drove neurons the most, such as maximum amplitude and spectral distribution statistics, were also acoustic features that were well decoded. However, there was a discrepancy for temporal entropy. Neuron spike rate and LFP power were often tuned to temporal entropy and encoders performed relatively well (Figure 5b), but it was not one of the top decoded acoustic features (Figure 6a). A similar situation existed for the 2nd voice acoustic feature, which drove high frequencies relatively well, but was very poorly decoded. We found that decoding from the LFP power spectra outperformed decoding from the population spike rate vector. This is not necessarily surprising, as the spike rate was computed for syllables of varying duration, and spike timing is likely to play a significant role in encoding time-varying spectro-temporal information. However, we found that including pairwise synchrony terms with the population spike rate vector decoder boosted performance to that of the LFP power spectra. We think that by including the synchrony terms, which are effectively the normalized dot product between two binned spike trains, we are adding back temporal information that was lost when averaging spikes over time to produce rates. Region L3 was found to be best predicted by the univariate tuning curves among other regions, while containing the most information about spectral distribution statistics and the time-varying fundamental frequency, and being more invariant to amplitude than CM, L1, or L2. Other researchers have found that region L3 was found to be more selective and tol-

erant for small changes in acoustic features [**Meliza2012**], better at discriminating distance calls [5], and overall more sparsely firing, noise correlated, and selective [1]. Our results complement the findings of selectivity by these researchers; a neural population must adequately represent the features of a vocalization to be selective. We found L3 to have a higher invariance to amplitude, which could contribute to its tolerance to acoustic perturbations in vocalizations. However, it is somewhat surprising that a region found to be tolerant of perturbations in acoustic features contains so much information about the statistics of the spectral distribution. Perhaps this information is encoded by inhibitory neurons and used to suppress the firing of excitatory neurons, endowing them with their selectivity. Finally, we demonstrated a concrete relationship between the local spike rate and spike synchrony, and LFP power that the population produces. If the LFP is comprised predominantly of synaptic currents, then we are showing that those synaptic currents are directly translated into the average spike rates of neurons, and enabled us to predict the LFP power from spike rate. We found that the addition of spike synchrony boosted our predictive power for frequencies above 50Hz. Synchronous synaptic currents are thought to be a significant contributor to LFP power [**REF**]. Taken together with the results that decoder performance of the LFP power spectrum is higher than the population spike rate vector, and that computing the power spectrum of a raw LFP is computationally cheaper than online spike sorting, we have demonstrated a decoding methodology that could improve the performance of brain-machine interfaces.

Chapter 7

Methods

Electrophysiological methods and acoustic analyses are fully described in [Elie2015a] and [Elie2015b] respectively and are summarized here in the first sections of the methods. Then we describe in detail the computational methods used for the calculation of local field potentials (LFPs), the power spectra and the encoding and decoding analyses. All animal procedures were approved by the Animal Care and Use Committee of the University of California Berkeley, and were in accordance with the NIH guidelines regarding the care and use of animals for experimental procedures.

Animals

The animal subjects studied were adult and juvenile zebra finches (Taeniopygia guttata) from the colonies of the Theunissen and Bentley labs (University of California, Berkeley, USA) (Figure 1a). The electrophysiology experiments described below involved four male and two female adults from the Theunissen lab colony. The acoustic recordings described in the next subsection involved twenty-three birds (eight adult males, seven adult females, four female chicks, four male chicks). Six adults (three males, three females) were borrowed from the Bentley lab. The electrophysiology subjects were housed in unisex cages and allowed to interact freely with their cagemates. All subjects were in the same room and were could interact visually acoustically. The acoustic recordings were performed on pair-bonded adults housed in groups of 2-3 pairs. Chicks were housed with their parents and siblings.

Zebra Finch Vocalization Types

Zebra finches communicate using a repertoire of vocalizations that are dependent on behavioral context. Following [Zann1996], acoustic signatures and behavioral contexts were used to classify vocalizations into nine different categories (Figure 1d). A succinct description of vocalization types recorded in the experiment can be found in Table 1 of [Elie2015a], but here we summarize. Song is a multi-syllable vocalization emitted only by males. Songs are comprised of repeating motifs of syllables, and in the dataset used, have a duration of 1424

+/- 983ms. There are two types of monosyllabic affiliative calls used to maintain contact. Distance calls are loud, used when not in visual contact, and longer in duration (169 +/- 49ms) than Tet calls, emitted when in visual contact during hopping movements, with a duration of 81 +/- 16ms. Nest calls are soft monosyllabic vocalizations emitted by zebra finches looking for a nest or constructing a nest. With a duration of 95 +/- 75ms, they are similar to Tets. Zebra finches emit two types of calls when they are acting out aggressively or being attacked. Wsst calls are noisy (broadband) and often long (503 +/- 499ms) calls emitted by a zebra finch when it is being aggressive. Distress calls are long (452 +/- 377ms), loud, and high-pitched vocalizations emitted by a zebra finch when escaping from an aggressive cage-mate. Both types of vocalizations can be mono or polysyllabic. Juvenile zebra finches emit two types of calls. Long tonal calls are the precursor to the adult distance calls; they are loud, long (durations of 184 +/- 63ms) and monosyllabic, emitted when the chick is separated from its siblings or parents. Begging calls are emitted when a juvenile zebra finch is begging for food from a parent, it is loud, long (duration of 382 +/- 289ms), and monosyllabic.

Electrophysiology and Histology

Twenty-four hours before recording, the subject was deeply anaesthetized with isoflurane, injected topically with lidocaine, and a head-holding fixture was cemented into the skull. On recording day, the subject was fasted for one hour, anaesthetized with isoflurane, head-fixed in a stereotaxic device, and two rectangular openings were made over the auditory area of each hemisphere. An electrode array with two columns of eight tungsten electrodes was lowered into each hemisphere (Figure 1b,c). Electrodes were coated in DiI powder so that their path through the tissue could be analyzed post-experiment. The electrodes ran rostral-caudal lengthwise in eight rows, with two columns that ran medial-lateral. During the experiment, the subject was placed in a soundproof chamber and electrode arrays were independently lowered. Probe stimuli were used to determine visually whether the areas were auditory. Once a reliable site was found, a stimulus protocol was played over speakers within the chamber (described in next subsection). When the stimulus protocol was complete, the electrodes were lowered deeper by at least 100 microns before playing the protocol again at the next site. Once the recordings were finished, typically after 4-5 recording sites, the subject was killed with an overdose of isoflurane, the brain was removed and fixed with paraformaldehyde. Coronal slices of 20 microns were made with a cryostat and Nissl stained. The slices were examined under a microscope and the DiI tracts were used to determine electrode penetration through anatomical regions. Six auditory areas were differentiated: three regions of field L (L1, L2, L3), caudomedial and caudolateral mesopallium (CMM and CML), and caudomedial nidopallium (NCM).

Stimulus Protocol

The vocalizations of ten individuals (three adult females, three adult males, four chicks) were used in the stimulus protocol. The vocalizations of four of the individuals (one male adult, one female adult, one male chick, one female chick) were played at each recording site, and three of each vocalization type were randomly selected from the other birds to be played. Each vocalization was played on average 10 times, randomly interleaved with other vocalizations. The protocol lasted an average of one hour. Monosyllabic vocalizations such as Distance and Tet calls were played with 3-4 renditions in series with inter-syllable intervals chosen to match what was observed naturally.

Syllable Segmentation

For this work we segmented all call types into syllables. The amplitude envelope of the series of call syllables was used for the segmentation. First the spectrogram was computed, and then the standard deviation of power across frequencies was computed at each time point to produce a time-varying amplitude envelope. Syllables began when the amplitude envelope exceeded a threshold value set to the 2nd percentile of the amplitude envelope distribution for all syllables. The syllable was marked as completed when the amplitude envelope subsequently dropped below this threshold. Syllables separated by 20ms or less were considered as one event. Songs and Begging calls were segmented as well.

Acoustic Features

We utilized bioacoustic methods to quantify acoustic features of each syllable, referred to as Predefined Acoustical Features and described extensively in the Methods of [Elie2015b] and summarized here. The 20 acoustic features fall into three different categories - temporal, spectral, and time-varying fundamental features. Temporal features were computed from the temporal envelope of the syllable. The temporal envelope was computed by rectifying the syllables raw sound pressure waveform and low-pass filtering with a cutoff frequency of 20 Hz. The temporal envelope was normalized by its sum, turning it into a probability distribution. The mean (Mean T), standard deviation (Std T), skew (Skew T), kurtosis (Kurt T), and entropy (Ent T) were computed and used as features. The peak amplitude of the syllable was computed as the peak of the non-normalized temporal envelope, and labeled as Max A. Spectral features were computed from the spectral envelope, which is the power spectrum computed from the raw syllable sound pressure waveform. The spectral envelope was normalized by its sum, and the mean (Mean S), standard deviation (Std S), skew (Skew S), kurtosis (Kurt S) and entropy (Ent S) were computed. In addition, the 25th, 50th, and 75th percentile of the distribution were computed, and labeled as Q1, Q2, and Q3, respectively. Time-varying fundamental features were computed from the spectrogram of the syllable and other properties. A feature was computed to quantify the pitchiness of the syllable called the saliency. To compute this feature, first the auto-covariance of the raw

sound pressure waveform was computed. The peak in the auto-covariance at non-zero lag was found, and and the saliency was then computed as the ratio between that peak value and the value of the auto-covariance at lag zero. The saliency feature was labeled as Saliency. For all syllables with a saliency greater than 0.5, a time-varying fundamental frequency was computed by fitting the power spectrum at a time point within the syllable with that of an idealized harmonic stack. Deviations from this idealized harmonic stack were used to quantify inharmonic properties, such as the presence of a second peak in the spectrum not explained by the stack. This double voice phenomenon was the result of the two syrinxes of the singing bird producing two distinct sounds simultaneously. The second fundamental frequency in this situations was computed as the acoustic feature Pk 2, and the acoustic feature 2nd V was defined as the percent of time a second voice was found. Other acoustic features describing the time-varying fundamental are the maximum, minimum, mean, and coefficient of variation in the fundamental frequency over time, labeled Max F0, Min F0, Mean F0, CV F0, respectively.

LFP Power Spectrum Calculation

The local field potential was recorded with a sample rate of 381 Hz, limiting the maximum frequency of analysis to 190 Hz. The LFP on each electrode was z-scored across time for the duration of a stimulus protocol. The LFP was analyzed starting from the onset of a syllable, and the window of analysis was extended to 30ms following the syllable offset. Syllables of duration less than 40ms or more than 400ms were excluded from analysis. We will denote the z-scored LFP conditioned on a stimulus s, for trial m, electrode k as ukm(t, s). We computed the LFP power spectrum from the Gaussian-windowed short-time Fourier Transform (STFT). The time points in the spectrogram were spaced by an increment of = 5ms. The window size was W=0.060, 60ms. The frequency spacing was constant across stimuli due to the fixed window size, equal to f = 9.78 Hz, and ranged from 0 to 185 Hz. The value of the STFT, centered at time and frequency f was computed as: zkm(, f, s)=t=1Texp(-(t-)22W) exp(i2ft) ukm(t, s)

where i=-1, T is the duration of the stimulus in number of time points at sample rate fs = 381 Hz, and W was chosen such that 95

W=W6

From the complex-valued STFT, we averaged power across windowed segments, of which there were TW=floor(TW), to get the power spectrum for electrode k, trial m, stimulus s:

xkm(f, s)=1TW=1TWzkm(, f, s)2

Once the power spectra were computed for each trial, they were averaged across trials to produce an average power spectrum for stimulus s.

LFP Pairwise Coherency Function Calculation

To explore pairwise information in the LFP, we computed a normalized cross correlation function called the coherency [Hsu2004]. To compute the coherency between signals ukm(t,

s) and ujm(t, s) on electrodes k and j, we first computed their autocorrelation functions (ACFs) and cross correlation function (CF), conditioned on stimulus s, for trial m. Then the Fourier transform of the ACFs and CF were taken, and the coherency was computed for stimulus s and time lag as:

ckjm(, s)=IFFTFFTCFkj(, s)— FFTACFk(, s) — — FFTACFj(, s) —

where IFFT is the inverse Fourier Transform. Twenty-one lags were used, ranged from -26ms to 26ms. In practice small values of the ACFs made the quantity in parenthesis very high, and so a threshold was placed on the lowest values of the ACFs, prior to dividing. For these points we set ckj(, s)=0. Once the coherency functions were computed for each trial, they were averaged across trials to produce an average coherency function for stimulus s.

Population Spike Rate and Spike Synchrony

The spike rate for cell i, trial m, stimulus s, was computed as the number of spikes divided by the duration of the stimulus. Let Nim(s) be the number of spikes that occur during stimulus s, trial m, for cell i. Then the spike rate is given as:

rim(s)=Nim(s)duration of s

The spike rate for cell i was averaged across trials to produce an average spike rate ri(s), and the the population spike rate vector for stimulus s was defined as the vector of average spike rates for Q cells recorded at a given site:

r(s)=[r1(s) rQ(s)]

To compute spike rate synchrony for stimulus s, trial m, between cells i and j, we first binned the spike trains for i and j using a bin size of 3ms. Spike synchrony was computed as:

mij(s)=# of bins where both i and j spiked Nim(s) Njm(s)

Spike synchrony was then averaged across trials to produce an average synchrony ij(s).

Encoder and Decoder Dataset Construction

We used an encoding approach to determine what acoustic features drove individual neural responses, and a decoding approach to determine how much information about acoustic features was contained in ensemble activity. We defined the vector y(s) to be a collection of neural states, associated with stimulus s. y(s) was comprised of one or more of the following neural states: the multi-electrode LFP power spectra, (LFP PSD), the pairwise coherency at all lags for all electrode pairs (Pairwise CFs), the population spike rate vector (Spike Rate), or the pairwise synchrony for all pairs of neurons (Spike Sync). We defined a vector x(s) to be a collection of acoustic features associated with stimulus s. The encoder attempts to predict a single scalar neural feature yi(s) from the vector of acoustic features x(s). The acoustic features are chosen by the experimenter and precede the neural response, they have the potential to causally induce changes in the neural state. The decoder attempts to predict a single acoustic feature xj(s) from the neural feature vector y(s). The dataset was constructed from one run of a stimulus protocol on a recording site. Each stimulus

protocol typically contained around 130 vocalizations randomly presented 10 times each. After segmentation and trial averaging, there were roughly D=700 samples. Each protocol contained vocalizations from eleven different birds - seven adults and four chicks.

Optimization and Cross Validation

The decoder tries to predict a single acoustic feature xj(s) from a vector of neural responses y(s). We define the matrix Y to be of size DxM, where D is the number of syllables in the dataset, and M is the number of neural features for a given representation. We define the matrix X to contain D rows and 1 column, each row contains value of the acoustic feature xj(s) for a different syllable s. For the LFP PSD neural features, M=192 (16 electrodes x 12 frequency bands), for the LFP CFs, M=2520 ((16$\hat{2}$ - 16) / 2) electrode pairs, 21 time lags)). For the Spike Rate neural features, there were typically 25-35 cells per site, so M ranged from 25-35, while for the spike synchrony features, M ranged from 300 to 595. The vector y was always z-scored prior to fitting, as was each column of X. Regression finds optimal linear model weights w and scalar intercept b that minimize the sum of squares error between the model prediction and the actual data:

LX, y, w=—— (Xw+b)-y ——2

Given the high dimensionality of some of our feature spaces, it was important to regularize values of w, so that we did not overfit the data. We utilized Ridge regression with scikits.learn to do this regularization. Ridge regression computes the optimal weight vector w as:

w=(XTX- I)-1 XTy

the value is a user-defined hyperparameter, high values of force weights towards zero. The value of the hyperparameter is found using a cross-validated approach. Our goal was to find a value for that maximized generalization performance. We searched 50 values of , chosen from a logarithmically spaced set that ranged from 10-2 to 106. For each candidate value of , we divided the data into a training and validation set 50 different times, and trained the model on the training set to find a set of weights w, evaluating the performance on the test set. The value of that had the best average performance on the 50 test sets was chosen as the optimal . We trained a final model on the entire dataset using the optimal , to produce a final set of weights used for analysis. Vocalizations within the same call category for the same bird can be highly correlated, and may produce very similar neural responses. This could artificially inflate the performance. To control for this, the validation set was comprised of the vocalizations of two randomly chosen adults and two randomly chosen juveniles from the 11 birds in the dataset. The validation set always had at least one example of each call type. We used the R2 averaged across validation sets as a performance measure for our data. The formula for R2 is given as:

R2=Lnull-LLnull

where Lnull is the sum of squares error for a model that only tries to predict y with the intercept term b. It is well known that the R2 increases when the number of features M increases, but this does not apply to the R2 computed on validation sets, which enabled us to compare model performance between models with different numbers of parameters.

Spike Rate to LFP Power Encoder

In addition to trying to predict neural features from acoustic features, we also build an encoder that attempted to predict LFP power for a given frequency and electrode from the population spike rate vector and spike synchrony features. The dataset construction was the same as described for the relationship between neural and acoustic features, but each row of the data matrix X was comprised of the population spike rate vector for a given stimulus, or in addition, the spike synchrony between each pair of cells. Each element of the dependent variable vector y was comprised of the LFP power for a given frequency and electrode. A separate encoder was trained for each frequency/electrode combination.

Tuning Curves and Generalized Additive Encoder

We fit nonlinear tuning curves to determine the relationship between an acoustic and neural feature. To build the tuning curve between an acoustic feature xjand neural feature yk, we first binned the values of xj across stimuli into evenly spaced bins. We then computed the mean and standard error of yk for each bin. The mean and standard error as a function of the acoustic feature value were then fit with a cubic spline, and the spline was evaluated for 50 evenly spaced values that spanned the range of xj to produce the final tuning curve. This process was performed using a cross validation procedure identical to that used for the encoders and decoders. The number of bins was used as a hyperparameter, taking on the value of 5, 7, 10, or 12. The R2 was computed across 25 validation sets and averaged to produce a performance measure, and the number of bins that gave the best validation performance was chosen as the optimal hyperparameter. The tuning curves produce a nonlinear estimate of the spike rate or LFP power given an acoustic feature value. To build a final encoder for spike rate or LFP power, we performed a Ridge regression as described in the previous section on the z-scored output of the tuning curves, to produce a final estimate of the spike rate or LFP power. A model of this type is commonly referred to as a Generalized Additive Model [Hastie1990].

Ensemble Decoding Analysis

We computed the decoder performance for each acoustic feature as a function of the number of electrodes. To do this, we combined data for each site across hemispheres, giving a total of 32 electrodes per recording site. For each site, a number of electrodes was selected ranging from 1, 4, 8, up to 32 in increments of 4. Once the number of electrodes was selected, up to fifty different combinations of that number of electrodes were selected from the site data. A decoder was trained on each combination, using cross validated Ridge regression decoder methods described in previous sections. The validation R2 was computed for each electrode combination, and the mean R2 across combinations was reported as the performance for that site given the number of electrodes specified.

Bibliography

[1] Calabrese A. and Woolley S. M. "Coding principles of the canonical cortical microcircuit in the avian brain." In: *Proceedings of the National Academy of Sciences, 112(11), 3517-3522* (2015).

[2] Ince R. A., Panzeri S., and Kayser C. "Neural codes formed by small and temporally precise populations in auditory cortex." In: *The Journal of Neuroscience, 33(46), 18277-18287* (2013).

[3] L. Cohen. *Time-frequency Analysis*. Prentice-Hall, 1995.

[4] Lewis L. D. et al. "Thalamic reticular nucleus induces fast and local modulation of arousal state". In: *Elife, 4, e08760* (2015).

[5] Elie J. E. and Theunissen F. E. "Meaning in the avian auditory cortex: neural representation of communication calls". In: *European Journal of Neuroscience 41.5: 546-567* (2015).

[6] Elie J. E. and Theunissen F. E. "The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals". In: *Animal Cognition, 1-31* (2015).

[7] Lisman J. E. and Jensen O. "The theta-gamma neural code". In: *Neuron, 77(6), 1002-1016* (2013).

[8] Schneidman E., Bialek W., and Berry M. J. "Synergy, redundancy, and independence in population codes". In: *Journal of Neuroscience, 23(37), 11539-11553* (2003).

[9] Schneidman E. et al. "Weak pairwise correlations imply strongly correlated network states in a neural population". In: *Nature, 440(7087), 1007-1012* (2006).

[10] Buzsaki G., Anastassiou C. A., and Koch C. "The origin of extracellular fields nad currents-EEG, ECoG, LFP and spikes". In: *Nature Reviews Neuroscience, 13(6), 407-420* (2012).

[11] Buzsaki G. and Wang X. J. "Mechanisms of Gamma Oscillations". In: *Annual Review of Neuroscience, 35, 203* (2012).

[12] Kim G. and Doupe A. "Organized representation of spectrotemporal features in songbird auditory forebrain." In: *The Journal of Neuroscience, 31(47), 16977-16990* (2011).

[13] Beckers G. J. et al. "Plumes of neuronal activity propagate in three dimensions through the nuclear avian brain." In: *BMC Biology 12(1)* (2014).

[14] Denman D. J. and Contreras D. "The structure of pairwise correlation in mouse primary visual cortex reveals functional organization in the absence of an orientation map." In: *Cerebral Cortex, 24(10), 2707-2720* (2014).

[15] Shlens J. et al. "The structure of multi-neuron firing patterns in primate retina". In: *The Journal of neuroscience, 26(32), 8254-8266* (2006).

[16] Lakatos P. et al. "An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex". In: *Journal of neurophysiology, 94(3), 1904-1911* (2005).

[17] Quiroga R. Q. and Panzeri S. "Extracting information from neuronal populations: information theory and decoding approaches". In: *Nature Reviews Neuroscience, 10(3), 173-185* (2009).

[18] Hamilton L. S. et al. "Optogenetic activation of an inhibitory network enhances feedforward functional connectivity in auditory cortex". In: *Neuron, 80(4), 1066-1076* (2013).

[19] Nirenberg S. and Latham P. E. "Decoding neuronal spike trains: how important are correlations?" In: *Proceedings of the National Academy of Sciences, 100(12), 7348-7353* (2003).

[20] Panzeri S. and Schultz S. R. "A unified approach to the study of temporal, correlational, and rate coding". In: *Neural Computation, 13(6), 1311-1349* (2001).

[21] Weichwald S. et al. "Causal interpretation rules for encoding and decoding models in neuroimaging". In: *NeuroImage, 110, 48-59* (2015).

[22] Reimann M. W. et al. "A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents". In: *Neuron,79(2), 375-390* (2013).