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

Decoding the Rhythms of Avian Auditory LFP

Michael J. Schachter

Acknowledgments

Thank you to my parents. Dad, you told me when I was 4 that if I could read, I could do anything. I took that to heart, although there are still things I'm not able to do. Like I don't know how to fly a plane, for example. But I could probably learn with the right book! Mom, your work in hospitals and fascination with medicine and the human body is one of the things that inspired me to move towards Biology. And you know, being a loving and caring mom was also quite helpful.

Thank you to my brothers. Joe and Tom, you guys are more than I could have asked for. Reliable, supportive, funny, and good drinking buddies. Joe, your devotion to protecting your family and country is exemplary, your wit is top notch, and I have nothing but respect for you. Tom, you're one of the nicest people I know, and an excellent craftsman, whether you're building houses or playgrounds for cats.

Thanks to Northeast High School Magnet program. Philadelphia has run into a rough patch in regards to funding it's public school system, as have many cities, unfortunately, but I was lucky to attend Northeast and participate in the SPARC after school program. It's unfathomable how anyone could rationalize the defunding of public schools. I would not have written this thesis if it wasn't for Northeast Public High School.

Thank you to my former coworkers at Bluestone Software/Hewlett Packard. You really helped show me the ropes of programming, guiding me away from the awful complexities of C++ and into the syntactic Nirvana of Java. You gave me a chance to learn, to travel, and helped me to acquire a proper coffee habit.

Thanks to Dr. Edward Gruberg, who helped introduce me to Neuroscience at Temple University. You noted that I was a bit rough around the edges, but let me perform behavioral experiments on frogs, where I tied crickets to strings, and dangled them at various angles with respect to the frog's field of vision, to test whether lesions in the superior colliculus (or was it the nucleus isthmi?) affected prey/predator responses.

Thanks to Robert G. Smith at University of Pennsylvania, who taught me the ins and

outs of modeling retinal neurons. You took a chance with me, and it (eventually) paid off. I learned so much from you about the biophysical properties of neurons that still sticks with me, even almost ten years later. Thanks to everyone in Peter Sterling's lab.

Thanks to my friends James Bonaiuto and Lex Kravitz, we've worked with each other and grown together. May we some day live forever in robot bodies, or you know, whatever. There are many other friends that I have left out, for brevity. If you're one of them, thank you for reading my thesis, and thank you for being my friend!

Thanks to everyone in the Gallant lab, you provocative thinkers and team players. Whether through your programming skills, your machine learning, or computing cluster, I was always entertained and enlightened by you.

Thank you to the Redwood Institute for Theoretical Neuroscience. It's a special and rare thing to have so many Computational Neuroscientists all in the same building. Thanks for all the equations. Thank you especially to Fritz Sommer, who has advised me from the beginning of graduate school. Your suggestions have always opened my mind, leading me into a forest of knowledge!

Thank you to everyone in the Theunissen lab, for your knowledge, your data, your friendship and laughs, and the cakes! The longer we spend in the lab, the more bird-like we become, together, until the final logical step in our evolution takes place, at which point we will sprout wings and fly to the top of the Campanile, crowing like ravens and building a variety of complex nests.

Thank you Frederic Theunissen. You're literally the nicest PI on the planet, an expert in time frequency analysis, a caring father to your children, an excellent and attentive teacher, hyena tamer and bird whisperer. I couldn't have asked for more.

Thank you Layla. You have emotionally supported me throughout graduate school and dealt with all the ups and downs. I love you and am excited about sharing the remainder of my life with you. You're the best of the best!

Contents

List of Figures

List of Tables

Chapter 1

Decoding the Rhythms of Avian Auditory LFP

1.1 Abstract

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 used encoder models to show how intuitive acoustic features such as amplitude, spectral shape and pitch drive the spike rate of individual neurons and LFP power on electrodes. Using ensemble 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 with comparable performance. In addition we found that adding pairwise spike synchrony to the spike rate decoder boosts performance above that of the population spike rate alone, or LFP power spectra. We also found that decoder performance grows quickly with the addition of more neurons, but there is notable redundancy in the population code. Finally, we demonstrate that LFP power on an electrode can be well predicted by population spike rate and spike synchrony. High frequency LFP power (80-190Hz) integrates neural activity spatially over a distance of up

to 250μm, while low frequency LFP power (0-30Hz) can integrate neural activity originating up to 800μm away from the recording electrode.

1.2 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 modeling approach. Neurons in the auditory system have been probed with simple stimuli such as tones, but it is known that neural responses to complex acoustic sound cannot be understood as the linear combination of responses to the individual tones [1]. One difficulty then is to find the set of acoustic features (e.g. mean frequency, amplitude, spectral shape) that best describes these acoustically complex communication sounds. The chosen acoustic parameters are then used to represent the vocalizations in the analysis so the acoustic dimensions to which auditory neurons are sensitive can be identified. 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, and even to describe tonotopy in human auditory cortex [2].

However, most communication sounds are not fully described by their amplitude and frequency. Human speech, for instance, is a variable and complex sequence of smoothly changing harmonic stacks and noisy bursts. Some bird vocalizations share a similar complexity; Zebra Finch songs are complex but rigid sequence of harmonic stacks, noise bursts, and chirps [3]. The need to utilize natural sound stimuli to more effectively probe neuron re-

sponses 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 [4]. The model that corresponds to the spectrogram representation is a spatio-temporal receptive fields (STRF). A STRF predicts neuronal responses as a weighted sum of the recent spectrogram history. STRFs have been used with much success to describe auditory neurons both in mammals and birds [1]. 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 [5]. Although STRFs can thoroughly explore the respones to all possible spectro-temporal representations of sound, they can be difficult to interpret and dont offer a description of neural tuning in acoustic dimensions that are close to perception. At an intermediate level of representation, communication calls that are short and isolated in time can be represented by a small set of summary statistics that intuitively describe how they vary spectrally, temporally, and spectro-temporally. This approach has been utilized to successfully identify the distinctive acoustic properties of Zebra finch vocalizations [3]. Here, we leverage the same set of acoustic feature to represent Zebra finch vocalizations, and describe 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.

While the perception of auditory stimuli is the result of activity in a large population of neurons, past research mainly focused on the description of the response properties of single auditory neurons. 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. 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 [6]. There is evidence that neurons in the visual and auditory system exhibit robust spatial correlations in their spike patterns. Significant pairwise correlations between spike trains have been observed in retinal ganglion cells ([7], [8]), and V1 ([9]). In the auditory system, [10] showed 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 ([11], [12], [13]). Complementary decoding approaches can be used measure the contribution of correlation activities to information already existing in their independent spike rates. Using a decoding approach, [14] 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 additional 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.

We investigated the population code for zebra finch vocalizations in auditory cortex using both spikes and the local field potential (LFP) as measures of neural activity. The LFP is an aggregate signal comprised of synaptic and transmembrane currents elicited by sodium and calcium spikes [15], and the biophysical origin of LFP power may vary by frequency [16]. 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 [17]. Oscillations in the range of 30-80Hz 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 [18]. 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 [19]. 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 [20]. 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 [21] to show three dimensional propagation of slow wave oscillations (0-5Hz) 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 0-190Hz range in the Zebra finch auditory system, and show that the LFP power spectrum can be used to decode acoustic features from the full repertoire of natural Zebra finch vocalizations. We also directly investigate how much of the LFP for a given frequency band can be predicted by the spike rate and spike synchrony of simultaneously recorded neurons.

1.3 Results

Acoustic Features Covary and Cluster

Our goal is to describe how key features that characterize Zebra finch vocalizations are represented in neuronal spiking, and the LFP, as well as the relationship between neural spiking and the LFP. We segmented Zebra finch vocalizations and quantified their acoustic properties using a rich set of 20 acoustic features that characterized the syllables power

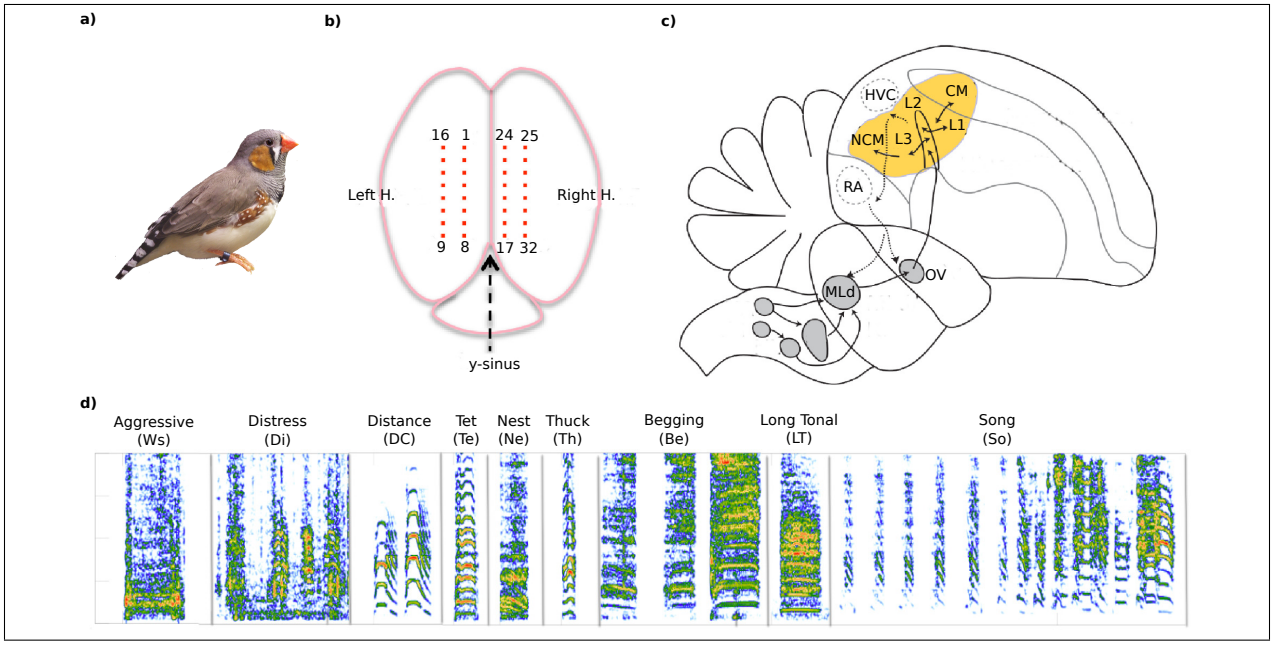

Figure 1.1: Neurophysiology and stimuli: **(a)** Six Zebra Finches were used for the experiment, 4 male, 2 female. **(b)** An 8x2 16 electrode array was placed in each hemisphere over the auditory area of the bird. Local Field Potentials (LFPs) and spikes were recorded simultaneously from the 32 electrodes. Electrodes are portrayed larger in the picture and their placement is not accurate. **(c)** A schematic of anatomical regions in the Zebra Finch auditory system. Recordings were made from thalamorecipient area L2, adjacent processing areas L1 and L3, as well as secondary auditory areas CM and NCM. Also pictured is the auditory thalamus Ovoidalis (OV), and brainstem area MLd, which is homologous to mammalian Inferior Colliculus. **(d)** Example spectrograms of syllables from the full Zebra Finch vocal repertoire.

spectrum, amplitude envelope, and the time-varying fundamental (see Methods - Acoustic Features). With the exception of temporal standard deviation (*Std T*), none of our features were dependent on the duration of syllables, that ranged in duration from 40ms to 400ms. *Std T* was linearly proportional to syllable duration (corrcoef=0.99). This produced a unique 20 dimensional feature vector for each syllable. It will be specifically noted when *Std T* is utilized in later analysis.

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 pitch saliencies (*Saliency*). Saliency is a measure

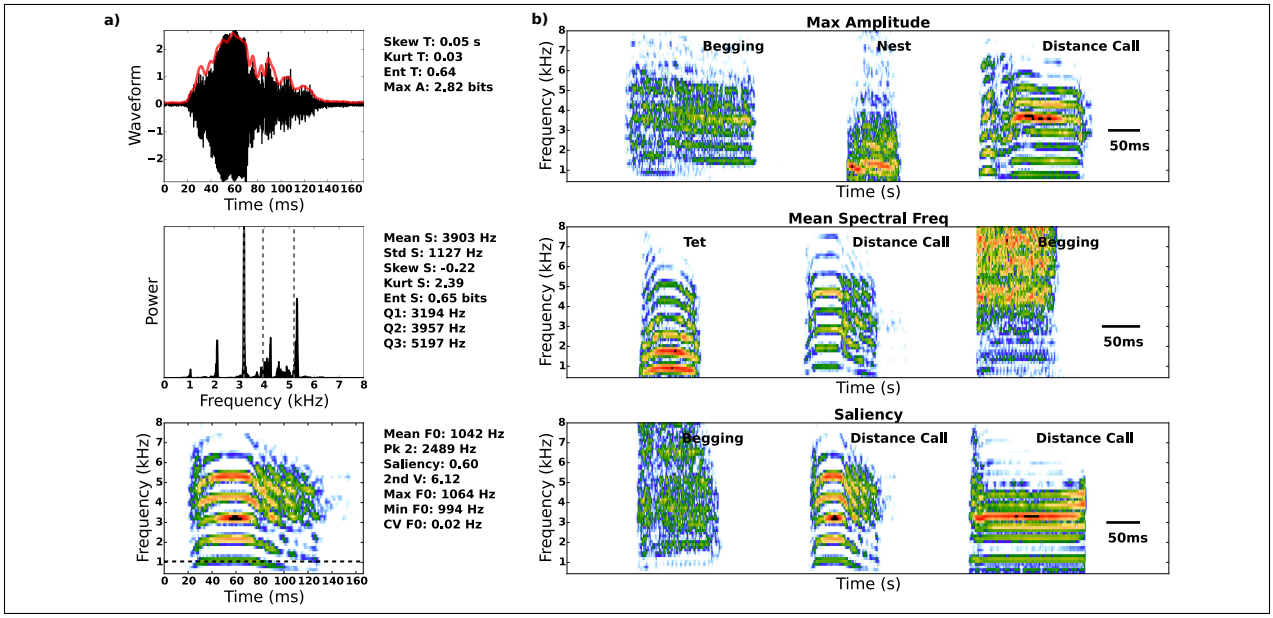

Figure 1.2: Syllables quantified by acoustic features: **(a)** an illustration of the acoustic features quantified for one syllable. The top plot shows the sound pressure waveform of a single syllable, with the amplitude envelope outlined in red. To the right in text are the temporal features quantified from the amplitude envelope and used in the study. The middle plot shows the power spectrum of the same syllable. Gray dotted lines indicate the first, second, and third quartiles of the spectral distribution (*Q1*, *Q2*, and *Q3*, respectively). *Q2* is the spectral median. Text to the right shows the values for the spectral features computed. The bottom plot of the left column shows a spectrogram of the syllable. The black dotted line indicates the mean fundamental frequency (*Mean F0*) computed from the time-varying fundamental. The text to the right shows the variety of features computed for the time-varying fundamental frequency. **(b)** Examples showing syllable variety along each acoustic feature axis. The top plot shows syllables ordered in increasing maximum amplitude, middle plot shows increasing mean spectral frequency, and bottom shows increasing saliency.

of syllable pitchiness, low for noisy syllables and high for harmonic-stack-like syllables. These features have been shown to be vital for determining the behavioral context, and hence semantic meaning, of Zebra finch vocalizations. Moreover, this set of features can be used in supervised classifiers to obtain identical discriminability performance of call types than the one obtained from a complete representation of the sound [3].

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 two parameters that describe the presence of a second voice (*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. We note also that these spectral frequency parameters are only weakly correlated with the mean fundamental; in other words, birds can increase their fundamental frequencies while not changing the spectral envelope of the sound and vice-versa [3]. The spectral kurtosis (*Kurt S*) was correlated to spectral standard deviation (*Std S*). 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. 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

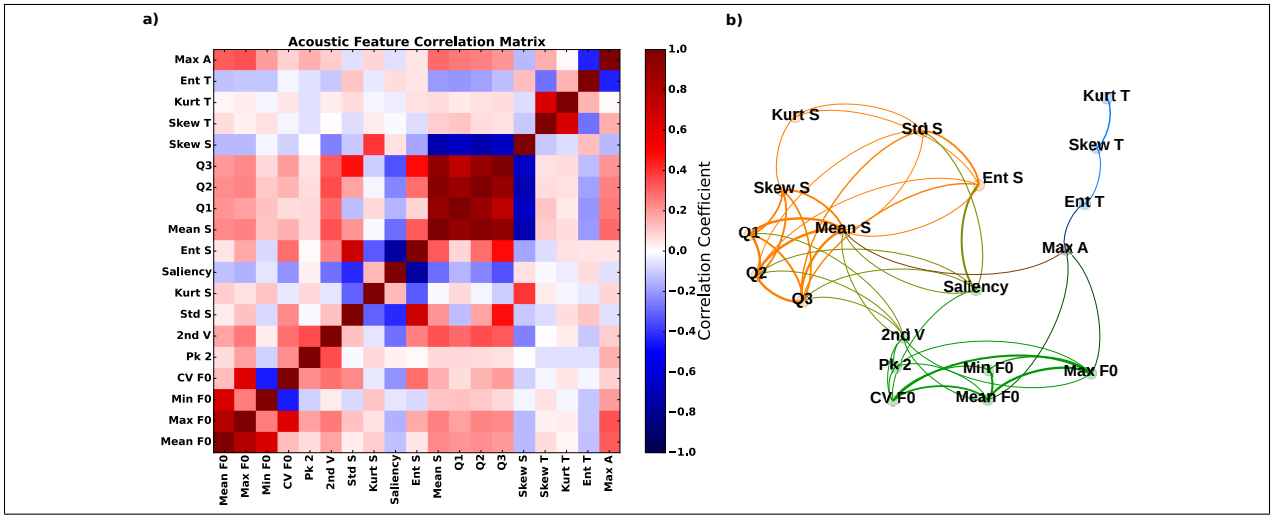

Figure 1.3: Acoustic features cluster into groups: **(a)** We performed a correlation analysis to show how the acoustic features were related to each other. The correlation matrix shows that the features fall into several clustered groups - spectrotemporal features involving fundamental frequency (*Mean F0* to *CV F0*), two clusters of spectral distribution features (*Std S* to *Ent S* and *Mean S* to *Skew S*), and temporal features (*Skew T* to *Max A*). **(b)** A graph of these features is shown, the width of the edge is proportional to the absolute value of the correlation coefficient between two features, and edges with magnitude less than 0.20 were discarded. The features are colored by their grouping (green = spectrotemporal, orange = spectral distribution, blue = temporal). Within-group edges are thicker than between-group edges. Not pictured, but utilized in the analysis, is the temporal standard deviation (*Std T*), which was linearly proportional to syllable duration, and uncorrelated to the other acoustic features.

to have lower maximum amplitudes. Temporal skew (*Skew T*) and kurtosis (*Kurt T*) were correlated with each other but not much with other features.

How Spikes and LFP are Driven by Amplitude, Mean Spectral Frequency, Saliency

We used an encoder analysis 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. 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

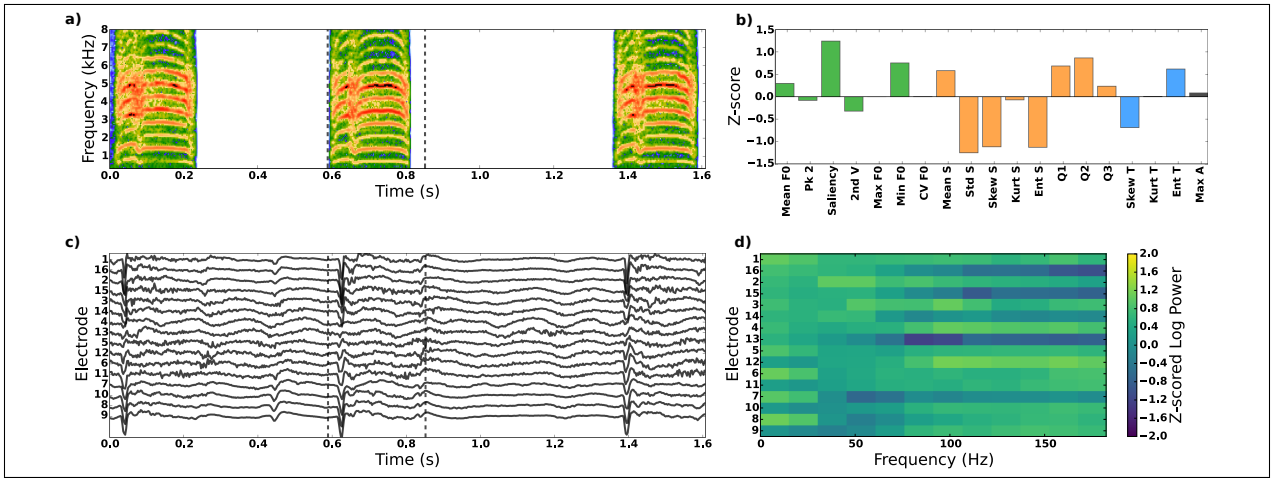

Figure 1.4: Preprocessing of syllables and LFP: An illustration of the methods used to transform the sound stimulus, multi-electrode local field potential, and spike trains. **(a)** A spectrogram of the vocalization, a male distance call (DC) which was emitted three times in a row. Dashed lines indicate the syllable chosen for visualization. The sequence of three call syllables was randomly repeated 10 times over the course of the recording session, along with other types of calls and songs that comprise the full vocal repertoire. **(b)** Acoustic features were computed for the syllable and z-scored across acoustic features for all syllables. The plot shows these properties for the selected syllable. **(c)** One trial of raw local field potential (LFP) recorded from a 8x2 multi-electrode array in the Zebra Finch auditory system, shown here for one trial. Electrodes are ordered rostral-caudal. **(d)** The z-scored multi-electrode LFP log power spectra, computed per trial and then averaged across trials. Each row corresponds to an electrode, and electrodes are ordered rostral-caudal, in the same order as (c). Frequency varies along the x-axis.

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. The frequencies were then summed in three bins - 0-30Hz, 30-80Hz, and 80-190Hz.

We used a stagewise procedure to build the encoder, meaning acoustic features were included in the encoder only if they boosted generalization performance, measured by the cross-validated R2 (see Methods - Encoder to Predict Spikes and LFP from Acoustic Features). However, because some of the acoustic features were highly correlated, we first chose

a set of baseline features that were always included in the regression - maximum amplitude (*Max A*), mean spectral frequency (*Mean S*), and pitch saliency (*Saliency*). We used a non-linear spline basis that allowed us to simultaneously fit nonlinear tuning curves between the neural response (spike rate or log power), and each acoustic feature (see Methods - Spline Basis Representation of Acoustic Features and Methods - Tuning Curves).

Figure 5a shows example tuning curves for several acoustic features, and the top row shows the curves for 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 (not shown). We quantified the tuning curves by computing their linear slope, ignoring slopes from models that had a generalization R2 of less than 0.05. Of the spike rate tuning curves for maximum amplitude (n=590), 62% had negative slope (decreased with increasing amplitude). For 0-30Hz LFP (n=302), 85% had a negative slope, while for 30-80Hz (n=396), 50% of the tuning curves had negative slope, and for 80-190Hz (n=404), 47% of the tuning curves had negative slope. No significant relationship was found between the anatomical corrdinates of a neuron or electrode, and the slope of the maximum amplitude tuning curve. Thus there are neurons that respond to increases in amplitude with increasing spike rate, and neurons that respond to increases in amplitude with a decreasing spike rate, and the same properties are reflected in tuning curves for the LFP.

The second row of Figure 5a illustrates that the relationship between spike rate or LFP power and mean spectral frequency is multimodal and nonlinear. Most spike rate tuning curves for mean spectral frequency had a negative slope (80%, n=652), meaning spike rate decreased as mean spectral frequency increased. Some spike rate tuning curves had a local maximum peak (also called a best frequency) near 3kHz. Examining only the tuning curves that had an identifiable local maximum (not at the endpoints) in the 2-4kHz range, we found that 58% (n=378) had an identifiable center frequency. The median center frequency of these tuning curves was 3.1kHz, with a standard deviation of 0.7kHz. The spike rate tuning curves were most similar to the 80-190Hz LFP band (Table 1). Tuning curves for 0-30Hz

	N	**% Negative Slope**	**# Central Peaked Tuning Curves**	**Median Center Frequency (Hz)**	**Center Frequency SD**
Spike Rate	652	79%	378	3155	746
LFP Power (0-30Hz)	310	96%	68	3236	1082
LFP Power (30-80Hz)	425	96%	138	3351	1147
LFP Power (80-190Hz)	421	78%	193	3351	886

Table 1.1: Tuning Curve Statistics for Mean Spectral Frequency

and 30-80Hz LFP were predominantly negative sloped, and had central peaks less often. No significant spatial relationship, i.e. tonotopy, was found between anatomical location and center frequency. To summarize, increases in mean spectral frequency most often decreased neural activity when measured by spike rate or LFP power. When tuning curves had a central peak, it was most often around 3kHz.

In the third row of Figure 5a we show example tuning curves for saliency. The curves exhibit some bimodality (positive and negative slopes). For spike rate tuning curves (n=620), 44% had a negative slope, while for 0-30Hz LFP (n=301) and 30-80Hz LFP (n=344), only 20% and 30% of the tuning curves had negative slopes, respectively. For the 80-190Hz frequency band, 40% (n=355) of the tuning curves had a negative slope, similar to spike rate tuning curves. The tuning curves for temporal skew (Skew T) overwhelmingly exhibited negative slopes, the percentage of negative-sloped tuning curves was greater than 95% for spike rate, and all LFP frequency bands.

The stagewise approach we utilized to fit our encoder allowed us to determine how important non-baseline acoustic features were in predicting neural activity. Figure 5b shows the fraction of times that a non-baseline acoustic feature was included in the analysis ("Feature Importance"). It is important to note that we allowed temporal standard deviation (*Std T*)

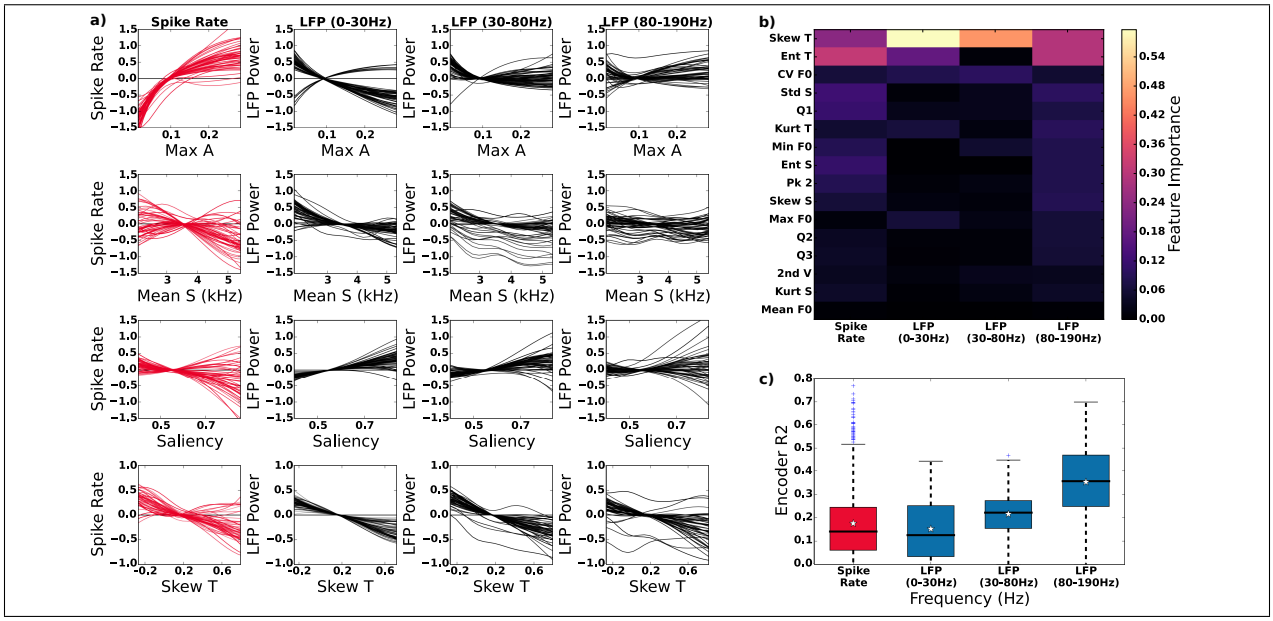

Figure 1.5: How spikes and LFP power are driven by amplitude, spectral mean, saliency: We built encoders to predict spikes or LFP power from nonlinearly mapped acoustic features. The encoder fit tuning curves simultaneously across acoustic features, and a weighted combination of tuning curve outputs were used to predict neural activity. **(a)** Tuning curves across sites for a subset of neurons, relating spike rate to acoustic features (first column). The subsequent columns show the mapping between LFP power in three different frequency bands (0-30Hz, second column, 30-80Hz third column, 80-190Hz fourth column). Tuning curves are shown for maximum amplitude (*Max A*, first row), mean spectral frequency (*Mean S*, second row), saliency (third row), and temporal skew (*Skew T*, fourth row). **(b)** In addition to the baseline features *Max A*, *Mean S*, and *Saliency*, other acoustic features were included using a stagewise regression if they improved encoder performance above baseline. The plot shows the fraction of times an acoustic feature was included in the regression, when predicting spike rate (first column), or one of the LFP power frequency bands (next three columns). **(c)** A boxplot of encoder performance for neurons (first box, red) and LFP power by frequency bands (subsequent columns, blue).

to be a feature. *Std T* is linearly proportional to syllable duration, and was included in a significant fraction ($> 90\%$) of encoders for the 0-30Hz and 30-80Hz frequency bands, though it is not displayed in Figure 5b. The temporal skew (*Skew T*) and temporal entropy (*Ent T*) were the top non-baseline features utilized by the encoders, with the rest of the acoustic features being rarely included.

The overall performance for the encoders, quantified by the cross-validated R2, is shown in Figure 5c for all LFP frequency bands spike rate. There was a significant effect of neural signal representation on encoder performance (ANOVA, $F(3, 1631)=164$, $p<0.01$), with the encoder models performing best for high frequency LFP (80-190Hz). There was also a significant, albeit smaller effect of region on encoder performance (ANOVA, $F(5, 1629)=9$, $p<0.01$); the encoders performed best for neurons/electrodes in region CM, and worst for region NCM. To summarize the results in this section, we successfully predicted spike rates for most neurons, and LFP power for most electrodes, with our encoder models. Neural responses were found to covary with amplitude (Max A) and saliency (*Saliency*) in a bimodal fashion, with some neurons/electrodes decreasing their response to increases in those variables, and others decreasing their response. Neural activity typically decreased with increasing mean spectral frequency (*Mean S*) or temporal skew (*Skew T*).

Regional Specificity in Single Electrode Decoding Performance

We built decoders to predict individual acoustic features from multi-band LFP power on individual electrodes to investigate whether there is regional specificity in decoding performance, before exploring ensemble decoding. The panels in Figure 6a 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 6b shows the regions that correspond to the plotted electrode locations. Figure 6c shows single electrode decoder performance (cross-validated R2) averaged within acoustic feature and

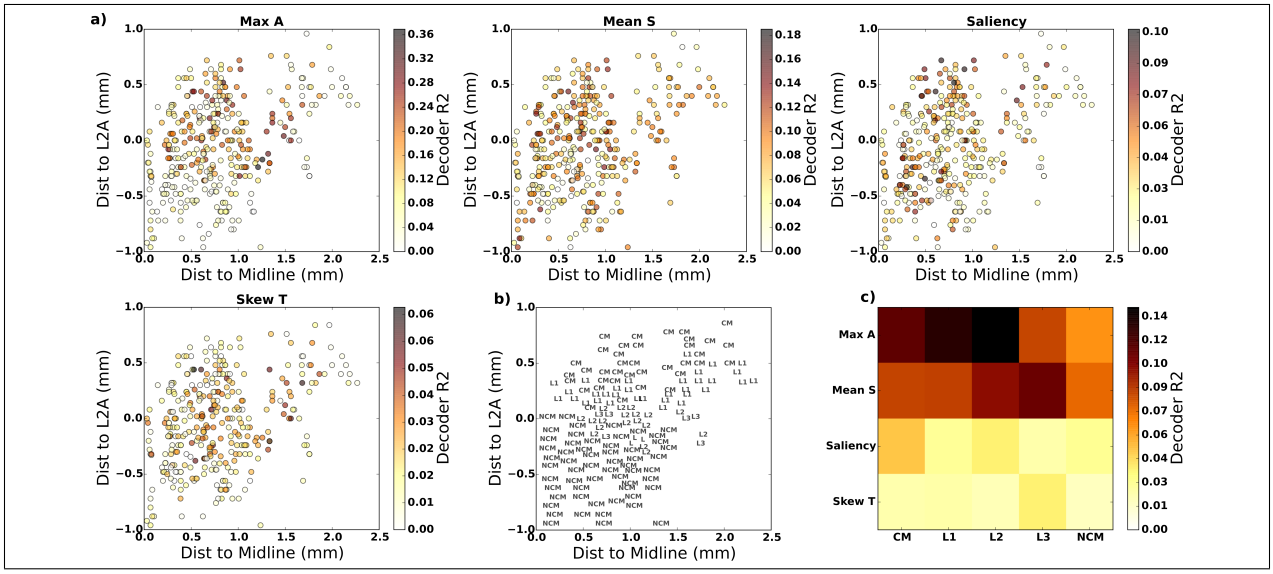

Figure 1.6: Regional specificity of single electrode decoder performance: Decoders were trained to predict acoustic feature values from the full LFP power spectrum of single electrodes. **(a)** Maps of single electrode decoder performance by anatomical location. Electrodes on left and right hemisphere are plotted together, with the left hemisphere points mirrored to correspond to the right hemisphere anatomical coordinates. **(b)** The anatomical region that corresponds to each electrode. **(c)** The R2 across electrodes, averaged within acoustic property and region.

anatomical region. Regional specificity in decoding performance was tested for the representative features shown in Figure 6a. Maximum amplitude (*Max A*) was best decoded from regions L2 and L1 (ANOVA, $F(4, 302)=9.1$, $p<0.01$), while mean spectral frequency (*Mean S*) was best decoded from regions L2 and L3 (ANOVA, $F(4, 309)=9.1$, $p<0.01$). Saliency had less regional specificity, but was best decoded from region CM (ANOVA, $F(4, 311)=7.5$, $p<0.01$). Temporal skew (*Skew T*) had very little regional specificity, but was best decoded from L3 (ANOVA, $F(4, 286)=3.2$, $p=0.01$). These results show that amplitude and frequency can be best decoded from electrodes in regions CM and L1, L2, L3.

Acoustic Features Decoded from Ensemble Activity

Our encoder analysis demonstrated that neuronal spike rates and LFP power are driven by maximum amplitude (*Max A*), mean spectral frequency (*Mean S*), pitch saliency (*Saliency*),

and temporal skew (*Skew T*), and our single electrode decoding analysis showed that these features can be decoded, with some regional specificity. Training decoders to predict individual acoustic features from ensemble activity allows us to explore how much information is contained in the population code and whether pairwise correlations improve decoder performance. So we built ensemble decoders to predict each individual acoustic feature from ensemble activity, represented by the population spike rate vector, LFP power spectra, or pairwise spike synchrony (Methods - Encoder and Decoder Dataset Construction). We utilized a simple measure of zero-lag pairwise synchrony, equal to the normalized dot product between two binary spike trains, and included synchrony terms for all pairs of neurons in an ensemble as input features to the decoder.

Figure 7a shows the mean performance by neural response type and acoustic feature. First, we found that including pairwise spike synchrony improved decoder performance for maximum amplitude (*Max A*) and spectral shape features (*Mean S*, *Q2*, *Q3*, *Skew S*, paired t-test, $p<0.01$ for all comparisons). No significant difference in performance was found between the population spike rate vector and multi-electrode LFP power spectra for those features (paired t-test, $p>0.01$ for all comparisons). Saliency and spectral entropy (*Ent S*) were best encoded by spike rate, which outperformed the multi-electrode LFP power spectra. LFP spectra outperformed spike rate and synchrony for temporal entropy (*Ent T*) and temporal skew (*Skew T*), implying that the power spectra representation contains more information about temporal properties than spike rate. To summarize, syllable amplitude and spectral shape are best decoded from population activity, and pairwise synchrony terms contain information about acoustic features that is not captured by the population spike rate vector.

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, providing up to 32 electrodes which we utilized to decode individual acoustic features. Through spike sorting we obtained 1-3 cells per electrode, and utilized up to 60 neurons. For the population spike rate vector and multi-electrode LFP spectra, we ran a decoder for a

	Max A	***Mean S***	***Saliency***	***Skew S***	***Skew T***
Pop. Spike Rate	30 +/- 2	24 +/- 1	34 +/- 3	28 +/- 2	24 +/- 3
Multi-electrode LFP	18 +/- 1	15 +/- 1	20 +/- 1	21 +/- 1	21 +/- 2

Table 1.2: Average number of neurons electrodes needed to decode acoustic features to 90% of peak ensemble decoding performance. Average was taken over dual-hemisphere recording sites, numbers listed are mean +/- stderr.

variety of combinations of electrodes for a fixed number of electrodes (Methods - Ensemble Decoding Analysis). The results are shown in Figure 7b and 7c. Each curve within a plot shows the average cross-validation R2 for a fixed number of electrodes/neurons, from a dual-hemisphere recording site. We quantified the number of neurons necessary to reach 90% peak decoder performance, across recording sites. We found that, when decoding from population spike rate, 25-30 neurons were needed to reach the 90% peak decoder performance, while decoding from LFP power spectra, 15-20 electrodes were required (Table 2). Given that ensemble decoding performance greatly outperforms single electrode performance, we conclude that the neural representation of these key acoustical features in the avian auditory system is based on a distributed ensemble code. Also, given that only up to half of the number of electrodes/neurons were needed to reach near-peak performance, we conclude that there is redundancy in that code.

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

The relationship between population spike activity and LFP power has not been directly quantified, so we built encoder models to predict LFP power in a given frequency band directly from population spike rate. In addition, we explored if spike synchrony better predicted LFP power (see Methods - Population Spike Rate and Spike Synchrony and Methods

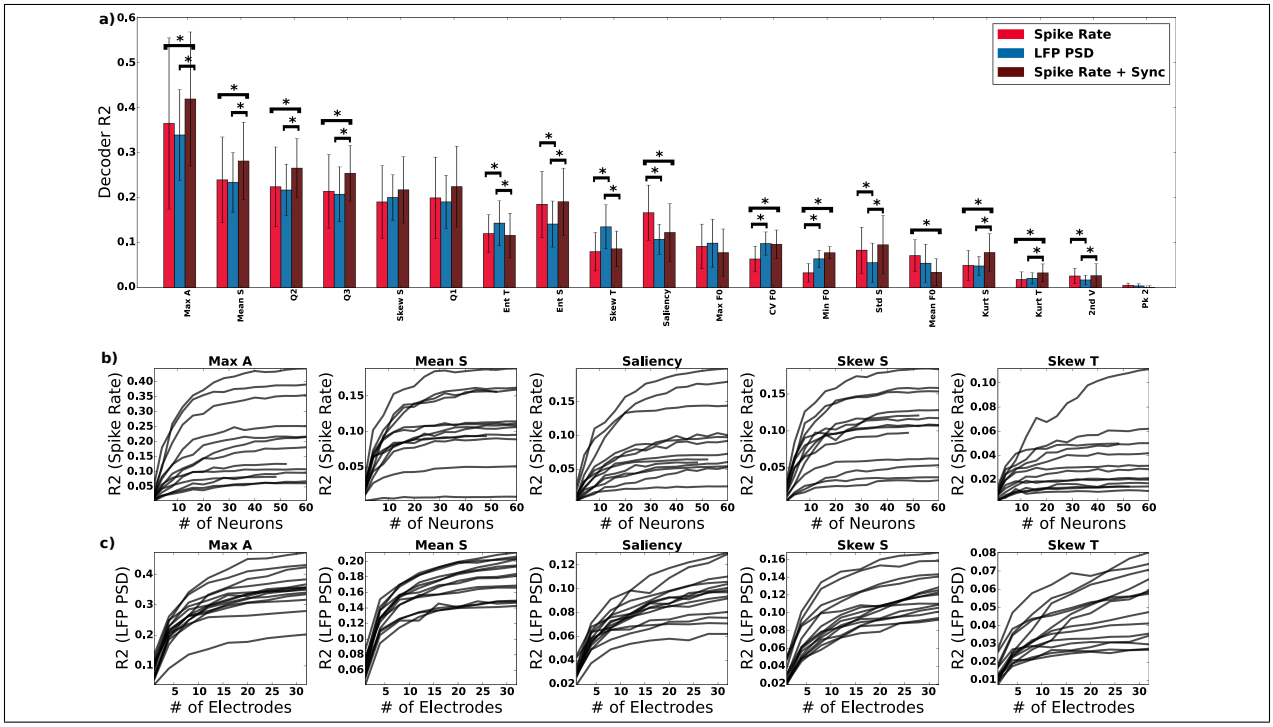

Figure 1.7: Ensemble decoding boosts performance: **(a)** Decoders were trained on individual electrode arrays from each hemisphere (16 electrodes) to predict each acoustic feature, from the population spike rate vector (red), LFP power spectra (blue), and pairwise spike synchrony (brown). Adding pairwise synchrony terms to population spike rate vector typically boosted neuron decoding performance to that of spikes and LFP power. **(b)** Average spike rate decoder performance was estimated as a function of the number of neurons from combined dual-hemisphere recordings. **(c)** Average LFP PSD decoder performance was estimated as a function of number of electrodes for combined dual-hemisphere recordings (32 electrodes total) at each site.

- Spike Rate to LFP Power Encoder). Figure 8a shows that LFP power can be predicted robustly by population spike rate. Adding spike synchrony terms improves predictive performance for the 30-80Hz and 80-190Hz bands (paired t-test, $p<0.01$ for all comparisons). For spike rate alone, performance increased with increasing frequency (ANOVA, $F(2, 875)=62.5$, $p<0.01$). The same was true for spike rate combined with spike synchrony (ANOVA, $F(2, 880)=281.8$, $p<0.01$). To summarize, the results in Figure 8a demonstrate that the LFP can be predicted best at high frequencies, by a combination of population spike rate and spike synchrony.

The LFP is typically described as the summed local electrical activity near a recording electrode [15]. On the one hand, it is therefore a good measure of the population response. On the other hand, by averaging local activity it could also eliminated ensemble codes occurring at that scale. It is important therefore to assess the size of the local area that is recorded in the LFP obtained from a single electrode. To investigate how much of an effect neurons had on LFP power as a function of distance from the electrode, we utilized the weights of the encoder trained to predict LFP power from spike rate. Each neuron, with its associated encoder weight, was a given distance from the electrode whose LFP was being predicted. We fit an exponential curve that mapped distance from predicted electrode to the squared-weight from the encoder for each neuron. For each curve we fit several parameters, and one was a space constant that showed how quickly squared-weight decayed to 36.8% of the maximum. The space constants were 770μm for 0-30Hz (R2=0.02), 237μm for 30-80Hz (R2=0.05), and 212μm for 80-190Hz (R2=0.08). These results show that for higher frequencies, the contribution to LFP power is primarily local ($< 300\mu$m), while low frequency LFP power integrates over a longer distance ($< 800\mu$m). 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 (paired t-test, $p < 0.01$ for all comparisons). To summarize, we have shown that LFP power is predicted from predominantly local spike rates.

1.4 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 skew, 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

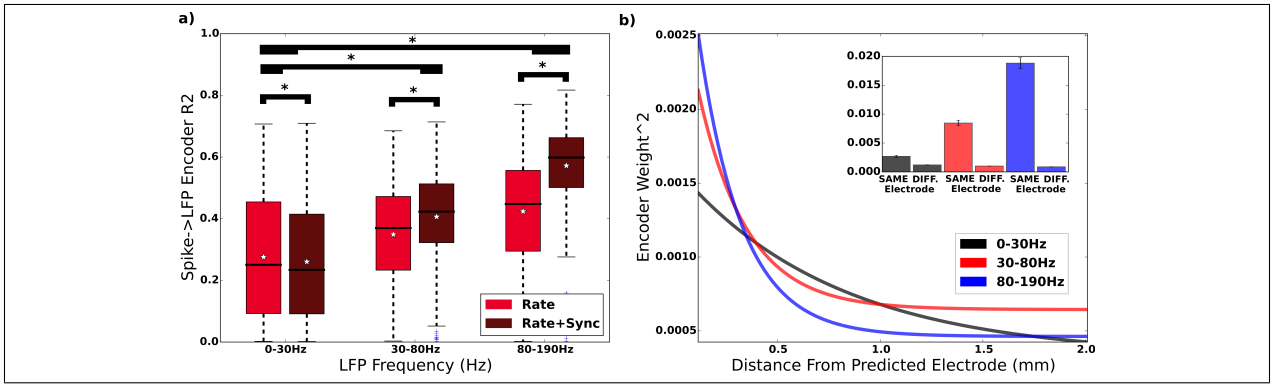

Figure 1.8: LFP power is a mix of local spike rate and synchrony: An encoder was trained to predict LFP power on a given electrode and frequency band from the population spike rate vector ("Rate"), and another encoder was trained that predicted LFP power from population spike rate combined with spike synchrony ("Rate+Sync"). **(a)** A boxplot of encoder performance for each frequency frequency, when predicting LFP power from rate alone ("Rate", red), and rate + spike synchrony ("Rate+Sync", brown). Adding synchrony terms improves predictive performance for the 30-80Hz and 80-190Hz frequency bands. **(b)** To determine the spatial spread of neuronal contribution to LFP power, we fit exponential curves for each frequency band that mapped distance from the electrode whose LFP power is being predicted (x-axis) to the squared-weight of a neuron in the encoder model. The length constants of the curves decrease as function of frequency bands, from 770um (0-30Hz, black), 237um (30-80Hz, red), to 212um (80-190Hz, blue). Inset: The average squared-weight for neurons on the same electrode as the LFP being predicted (Same Electrode), and neurons on a different electrode, for the three frequency bands.

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 [22], while decoders enabled us to determine the nature of ensemble encoding by population activity (Figure 7). The acoustic features that drove neurons the most, such as maximum amplitude and spectral shape, were also acoustic features that were well decoded. We found that decoding from multi-electrode LFP power spectra gave performance on par with that of the population spike rate. This is an important finding for brain-machine interfaces, which could potentially forgo the computationally expensive step of spike identification and sorting for the cheaper alternative of simply computing the

power spectra of the LFP.

We found that decoding from population spike rate plus pairwise spike synchrony outperformed decoding from LFP power spectra or spike rate alone. This lends support to the idea that stimulus information can be carried by the pairwise activity of neurons, and is in potential conflict with findings by [14], who show that pairwise correlations do not contain additional stimulus information. Our approach differs from that research in that they broke their stimuli into categorical tokens prior to using an information theoretic approach to quantify the importance of correlations, which effectively determines whether pairwise correlations contain information about stimulus identity. In contrast, we used a decoding approach to show that acoustic features of stimuli were better decoded when pairwise terms were introduced. [14] are effectively showing that pairwise correlations dont contain information about stimulus identity. We are showing that pairwise correlations contain information about specific acoustic features. Because acoustic features can be similar between stimuli, a pairwise correlation code that carries information about acoustic features, will be similar for two acoustically similar but distinct stimuli, decreasing the ability of a decoder to discriminate stimulus identity but maintaining information about the relevant features.

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 [23]. A detailed analysis of regional specificity by [24] 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. [25] 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. In our work, we found regions L2, L1, and CM were most effective for decoding maximum amplitude and mean spectral frequency, while region NCM was the least effective. We interpret this as showing NCM to be more invariant to amplitude and mean spectral

frequency than other regions. We also found region L3 to less predicted by amplitude, which could contribute to its tolerance to acoustic perturbations in vocalizations found in other work.

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 the 30-80Hz and 80-190Hz frequency bands. The exponential curve fits shown in Figure 8b did not fit the data perfectly; there was much noise in the relationship between distance from predicted electrode and squared-weight. This noise could be due to differences in neuron density and electrical properties over space.

1.5 Methods

Electrophysiological methods and acoustic analyses are fully described in [25] and [3] respectively and are summarized here below. We then describe in detail the computational methods processing and representing local field potentials (LFPs), 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 were performed on 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 [26], acoustic signatures and behavioral contexts were used to classify vocalizations into nine different categories (Figure 1d). A detailed description of call categories can be found [25]. We provide here, a very succinct summary of that characterization for a subset of the calls analyzed here and used in the neurophysiological experiments.

Song is a multi-syllable vocalization emitted only by males. Songs are comprised of repeating motifs of syllables, and in the dataset, have a duration of 1424 +/- 983ms. Song in zebra finches are used in pair bonding and mating behavior. The repertoire contains 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. Zebra finch also produce software calls used principally in the initial stages of pair bonding. 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.

Two calls are emitted by juveniles only. 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 and injected topically with lidocaine in order to remove a patch of skin over the skull and cement a head-holding fixture. On recording day, the subject was fasted for one hour, anaesthetized with urethane, head-fixed in a stereotaxic device, and two small 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 100μm 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 20μm 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 different 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 including Songs and Begging calls. 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.

Acoustic Features

We used a classic bio-acoustical approach to estimate a complete set of acoustic features of each syllable, referred to as Predefined Acoustical Features described extensively in the Methods of [3] and summarized here. The 20 acoustic features fall into three different categories - temporal, spectral, and 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. As was done for the temporal envelope, 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 degree of periodicity or pitch saliency of the syllable. To compute this feature, first the auto-correlation of the raw sound pressure waveform was computed. The peak in the auto-covariance at non-zero lag was found, and the saliency was then computed as the ratio between that peak value and the value of the auto-correlation at lag zero. The saliency feature was labeled as Saliency.

The pitch for all time windows where the saliency was greater than 0.5 was computed by fitting the power spectrum at a time point 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 independently driven vocal folds found in the syrinxes of songbirds [27]. Songs birds are capable of producing two independent voices although this is relatively rare in the zebra finch where the two folds are typically synchronized. 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 $u_k^m(t,s)$. 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 $\Delta\tau$ = 5ms. The window size was W=0.060. The frequency spacing was constant across stimuli due to the fixed window size, equal to f = 9.78 Hz, and ranged from 0 to 190 Hz. The value of the STFT, centered at time and frequency f was computed as:

$$z_k^m(\tau,f,s) = \sum_{t=1}^{T} \; exp(-\tfrac{(t-\tau)^2}{2\sigma_W}) \; exp(i2\pi ft) \; u_k^m(t,s)$$

where i=$\sqrt{-1}$, T is the duration of the stimulus in number of time points at sample rate f_s = 381 Hz, and W was chosen such that 95% of the mass of the Gaussian was contained in the window:

$$\sigma_W = \tfrac{W}{6}$$

From the complex-valued STFT, we averaged power across windowed segments, of which there were $T_W = floor(\frac{T}{W})$, to get the power spectrum for electrode k, trial m, stimulus s:

$$x_k^m(f,s) = \tfrac{1}{T_W} \sum_{\tau=1}^{T_W} \; |z_k^m(\tau,f,s)|^2$$

Once the power spectra were computed for each trial, they were averaged across trials to produce an average power spectrum for stimulus s. Finally, the power spectra were binned into three bins: 0-30Hz, 30-80Hz, 80-190Hz. Power within a band was the sum of values for $x_k^m(f, s)$ within that bands frequencies.

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 $N_i^m(s)$ be the number of spikes that occur during stimulus s, trial m, for cell i. Then the spike rate is given as:

$$r_i^m(s) = \frac{N_i^m(s)}{duration\ of\ s}$$

The spike rate for cell i was averaged across trials to produce an average spike rate $r_i(s)$, 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:

$$r(s) = [r_1(s) \cdot r_Q(s)]$$

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:

$$\gamma_{ij}^m(s) = \frac{\#\ bins\ where\ i\ and\ j\ spiked}{\sqrt{N_i^m(s) N_j^m(s)}}$$

Spike synchrony was then averaged across trials to produce an average synchrony $\gamma_{ij}(s)$.

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 $\mathbf{y}(s)$ to be a collection of

neural states, associated with stimulus s. $\mathbf{y}(s)$ was comprised of one or more of the following neural states: the multi-electrode LFP power spectra, ("LFP PSD"), the population spike rate vector ("Spike Rate"), or the pairwise synchrony for all pairs of neurons ("Spike Sync"). We defined a vector $\mathbf{x}$(s) to be a collection of acoustic features associated with stimulus s. The encoder attempts to predict a single scalar neural feature $y_i(s)$ from the vector of acoustic features $\mathbf{x}(s)$. The decoder attempts to predict a single acoustic feature $x_j(s)$ from the neural feature vector $\mathbf{y}(s)$.

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=600 samples. Each protocol contained vocalizations from eleven different birds - seven adults and four chicks.

Acoustic Feature Decoder Optimization and Cross Validation

The decoder tries to predict a single acoustic feature $x_j(s)$ from a vector of neural responses $\mathbf{y}(s)$. 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 $x_j(s)$ for a different syllable s. For the LFP PSD neural features, M=48 (16 electrodes x 3 frequency bands). 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 $\mathbf{y}$ 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:

$$L(X, y, w) = \|(Xw + b) - y\|^2$$

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:

$$\mathbf{w} = (X^T X - \alpha I)^{-1} X^T \mathbf{y}$$

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 tested 50 values of , chosen from a logarithmically spaced set that ranged from 10^-2 to 10^6. For each candidate value of α, 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 $\mathbf{w}$, 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 α. We trained a final model on the entire dataset using the optimal α, 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, the "cross-validated R2", as a performance measure for our data. The formula for R2 is given as:

$$R2 = \frac{L_{null} - L}{L_{null}}$$

where L_{null} is the sum of squares error for a model that only tries to predict $\mathbf{y}$ 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.

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.

Spline Basis Representation of Acoustic Features

We assumed that the relationship between acoustic features and neural activity was potentially nonlinear by transforming each acoustic feature into a cubic spline basis [28]. A scalar acoustic feature $x_j(s)$ was replaced by a five dimensional projection into the following basis:

$$b_j(s) = [x_j \;\; x_j^2 \;\; [x_j^3 - k_1]_+ \;\; [x_j^3 - k_2]_+ \;\; [x_j^3 - k_3]_+]$$

Where the $[...]_+$ operator is rectification, values less than zero are set to zero. The values for k_i are called the knots and were chosen as the 25th, 50th, and 75th percentile of the distribution of values for acoustic feature j.

Encoder to Predict Spikes and LFP from Acoustic Features

The encoder tried to predict a single scalar neural output, in the form of a spike rate or power at a given frequency band, from a set of acoustic features. Each acoustic feature was z-scored and then projected into a spline basis, as described above. The spline basis values were accumulated into a vector and used as the regressors in an optimization procedure described above for the decoder, but in addition a stagewise procedure was used to select

the smallest predictive subset of acoustic features. To accomplish this, a baseline regression was run with the basis functions of maximum amplitude, mean spectral frequency, and saliency as regressors, and the cross-validation R2 was recorded on this baseline "active set". In the next step, the improvement in R2 was computed for the addition of each remaining acoustic feature. The acoustic feature that produced the largest performance increase was added to the active set, and the process was repeated until no acoustic features were left that improved encoder performance. The algorithm completed by recording the acoustic features in the active set and their incremental improvements to R2.

Tuning Curves

To compute a tuning curve for an acoustic feature, the weights of a trained encoder model that corresponded to that feature were dot multiplied by the spline basis representation of that feature, producing a value representing the contribution to neural response by that acoustic feature. To produce the points that comprised the tuning curve, a spline basis matrix was computed for 20 regularly spaced points across the acoustic feature range, and this 20x6 matrix multiplied the 6 weights corresponding to that acoustic feature, to produce 20 values of neural response contribution (spike rate, LFP power) on the interval.

Predicting LFP Power from Population Spike Rate

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.

Once the encoders were trained, we fit exponential curves to the scatter data that mapped distance from predicted electrode to neuron squared-weight, for the encoder that predicted LFP power from the population spike rate. The form of the exponential function was:

$$f(x) = A\ exp(-\tfrac{x}{\lambda}) + B$$

The function was fit using the curvefit routine of scipy, and we reported the space constant as λ a measure of spatial extent.

Chapter 2

Distinct Oscillatory Subnetworks in the Zebra Finch Auditory System

2.1 Abstract

To understand how an auditory system processes complex sounds, it is essential to understand how the temporal envelope of sounds, i.e. the time-varying amplitude, is encoded by neural activity. We studied the temporal envelope of Zebra finch vocalizations, and show that it exhibits modulations in the 0-30Hz range, similar to human speech. We then build linear filter models to predict 0-30Hz LFP activity from the temporal envelopes of vocalizations, achieving surprisingly high performance for electrodes near thalamorecipient areas of Zebra finch auditory cortex. We then show that there are two spatially-distinct subnetworks that oscillate at different frequency bands, one subnetwork that oscillates around 19Hz, and another subnetwork that oscillates at 14Hz. These two subnetworks are present in every anatomical region. Finally we show that we can improve predictive performance with recurrent neural network models.

2.2 Introduction

Many animals use vocalizations to bond with each other, declare territory, signify the presence of predators, to beg for food, and other essential functions that enable their survival as a species. Human beings arguably have the most sophisticated vocalizations, our acoustically and temporally complex speech, comprised of a highly variable sequence of harmonic stacks and explosive bursts. Songbirds also have a sophisticated vocal repertoire, are capable of learning vocalizations through mimicry, and have auditory systems with a comparable complexity to that of mammals [29]. The study of Avian auditory systems can lead to a better understanding of the auditory systems of their mammalian counterparts.

A fundamental component to both human speech and Avian vocalizations is the temporal envelope, which is low-passed and slowly changing ($<$ 30Hz). The temporal envelope of human speech, ranging in frequency to 2-20Hz, is well represented in human auditory cortex [30]. The temporal modulation spectrum of speech, which is the spectrum of frequencies present in the temporal envelope, was found to have consistent peaks at 2Hz and 5Hz across multiple languages [31]. [32] explored the joint spectro-temporal modulation spectra of Zebra finch song, and found temporal modulation frequencies to be necessarily low ($<$ 50Hz).

Modulations of the temporal envelope have been shown to influence neural activity in Zebra finch auditory cortex [33]. In the work the researchers identified neurons that adapted their temporal integration timescale to the overall temporal envelope magnitude, integrating over shorter timescales for louder stimuli. Linear filter models were used in [34] to decode the temporal envelope of speech from human auditory areas. Taken together, these separate tracts of research show that temporal envelopes of vocalizations are encoded by both Zebra finch and human auditory neurons.

For our study we focused on 5-30Hz activity of the multi-electrode local field potential (LFP), in Zebra finch auditory cortex. The local field potential is a complex mixture of membrane currents from both synapses and voltage-gated ion channels [15]. There is evidence in mammals that the auditory LFP is comprised of a nested hierarchy of timescales [20].

Temporal envelope modulations occur in the 0-30Hz frequency band, and neural activity has been shown to follow the temporal envelope, but it is unknown in Avian auditory systems whether there are other frequencies of oscillation in the 0-30Hz band that coexist with the encoding of the temporal envelope, originated from independent processes. Our work first explores the temporal modulation spectra of vocalizations across the repertoire of the Zebra finch. Then we explore the 5-30Hz LFP, using linear and nonlinear models to predict multi-electrode LFP from the stimulus amplitude envelope.

2.3 Results

The Temporal Modulation Spectrum of Zebra Finch Vocalizations

The temporal modulation spectrum for a set of vocalizations is the set of frequencies for which the amplitude envelope oscillates. Figure 1 shows the approach we used to quantify the modulation spectrum of different types of vocalizations used in the experiment. Figure 1a shows examples of several vocalizations and their amplitude envelopes. Sequences of distance calls and tets comprise a category of "affiliative calls", and were repeated up to three times over two seconds, with random inter-call intervals ranging from 200-500ms. Figure 1b shows the average temporal modulation spectrum for affiliative calls (purple), which exhibit most power in the 2-5Hz range. Song, an example shown in the middle row of Figure 1a, had many fast syllable transitions and the average temporal modulation spectrum exhibited a peak from 6-10Hz. Begging calls, a fast repeating vocalization emitted only by juveniles, also exhibited higher temporal modulation frequencies. Modulation-limited noise (ML-noise) is a sound generated by noise that is constrained to have the same spectral and temporal modulations as Zebra finch song [35]. The bottom row shows an example of an ML-noise stimulus and its amplitude envelope. The temporal modulation spectrum of ML-noise has a large DC component but very little power in the 5-30Hz range.

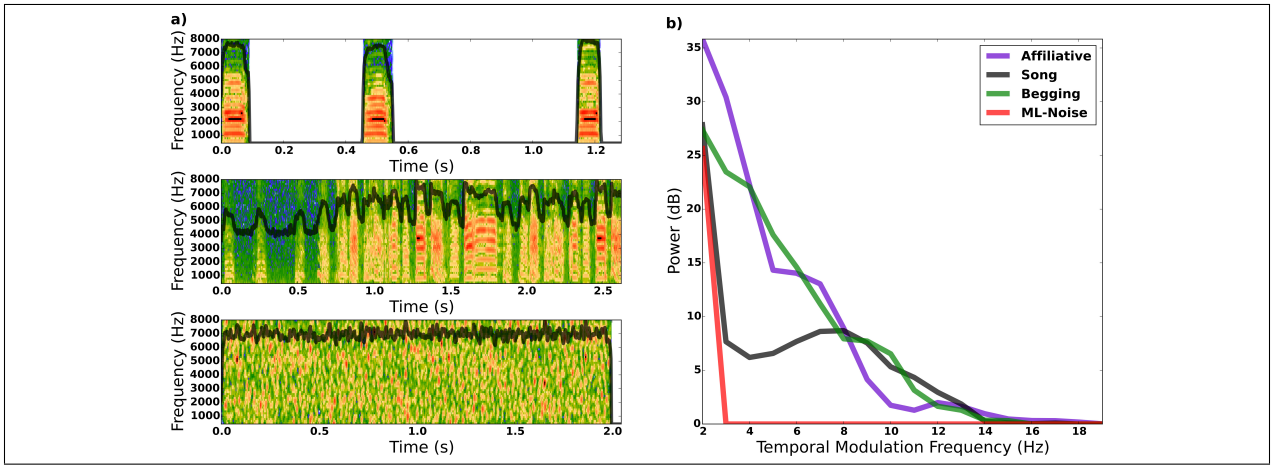

Figure 2.1: Temporal Modulation Frequencies of Vocalizations We computed the temporal envelopes of Zebra finch vocalizations and their power spectra. **(a)** Three examples of Zebra finch vocalizations, shown by their spectrograms, and their temporal envelopes, shown in black. The top row is a Tet, a prolific affiliative communication call. The middle plot is a Zebra finch song, which was comprised of many closely spaced syllables. The bottom plot is modulation-limited (ML) noise, which had a highly variable temporal envelope. **(b)** The average temporal modulation spectra (the power spectra computed from temporal envelopes) for several vocalization categories.

Linear Filter Encoders Predict LFP from Amplitude Envelope

The temporal modulation of sound through the amplitude envelope is vital for speech perception, and has be successfully decoded from multi-electrode activity in human auditory cortex [34]. We showed that temporal modulations varied for different types of Zebra finch vocalizations, and that these modulations occur primarily in frequencies that are less than 30Hz. Our next goal was to investigate whether temporal modulations in this range were represented by neural activity in the Zebra finch auditory system. We trained linear filter encoder models to predict 5-30Hz LFP activity from the stimulus amplitude envelope (see Methods - Linear Encoder Fitting). In this setting, the linear filter encoder is equivalent to a "temporal receptive field", that maps the recent history for the stimulus amplitude envelope to the bandpassed voltage recorded as the LFP, similar to a STRF with only one frequency band [1].

Figure 2 shows encoder predictions for an example song vocalization. Figure 2a shows

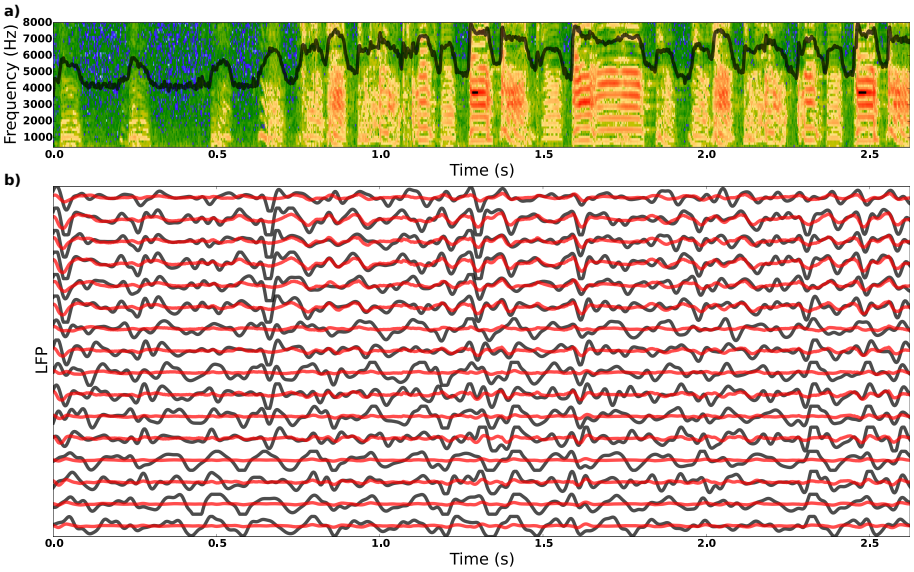

Figure 2.2: Linear Encoder Predictions for 5-30Hz LFP We trained linear filter models to predict 5-30Hz LFP activity on a single electrode from the temporal envelope of Zebra finch vocalizations. **(a)** The spectrogram and temporal envelope (black) of a Zebra finch song. **(b)** The raw 5-30Hz LFP (black) and linear filter encoder prediction (red) for 16 electrodes simultaneously recorded during the song presentation. Electrodes are ordered rostral-caudal from top to bottom.

the song vocalization as a spectrogram, along with the amplitude envelope. Figure 2b shows the multi-electrode LFP in black, and the linear encoder predictions in red. For many electrodes, the LFP responded to high amplitude, temporally sharp, and pitch-salient syllable onsets with sharp downward deflections. The linear model matches some of the stronger downward deflections. Sharp signals such those downward deflections are high amplitude and high-bandwidth, and predicting such a signal requires a filter with a sharp onset response. Following the sharp negative deflections in the LFP, a complex series of lower amplitude, lower bandwidth, more oscillatory upward and downward deflections occur. The linear model can be seen to match some of this activity as well, implying, perhaps surprisingly, that it may be operate on two different time-frequency scales.

We trained our linear encoders using a cross-validation approach (see Methods - Linear Encoder Fitting) and report the average cross-validation correlation coefficient between the model prediction on the holdout set and the actual LFP. Figure 3a shows performance

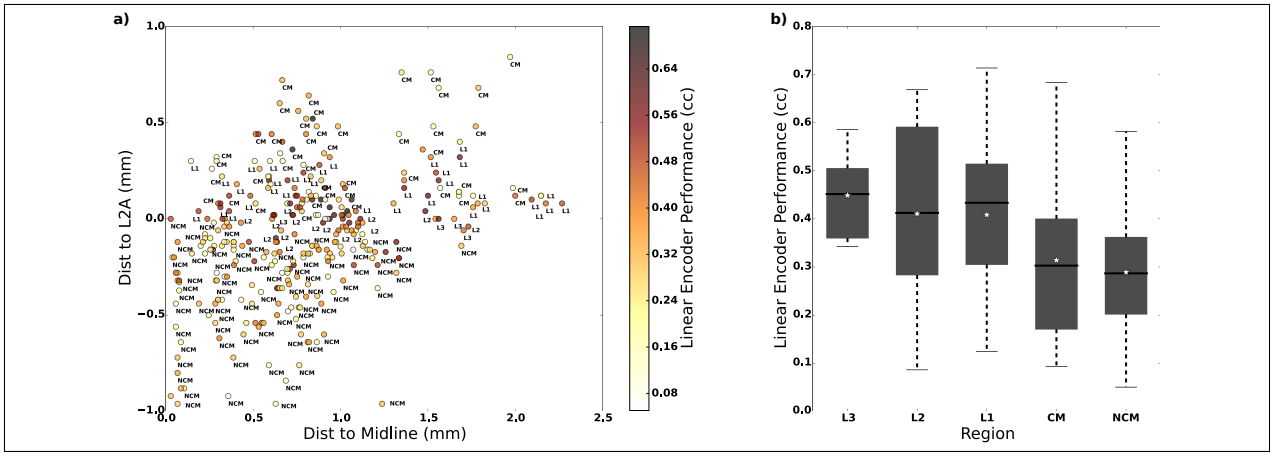

Figure 2.3: Linear Encoder Performance for 5-30Hz LFP (a) A map in anatomical coordinates of linear filter encoder performance, for electrodes across the dataset. Text annotations denote the anatomical region of the electrode. Electrodes on the left hemisphere were mirrored and superimposed with electrodes on the right hemisphere. **(b)** A boxplot of linear filter encoder performance by anatomical region. Performance was best in thalamorecipient region L2 and worst in secondary auditory region NCM.

for each electrode plotted by recording location. Electrodes rostral to the thalamorecipient region L2A had higher performances than electrodes caudal to L2A (ANOVA, $F(1, 307)=16.4$, $p < 0.01$). Lateral distance from the midline was not a significant predictor of encoder performance (ANOVA, $F(1, 307)=0.32$, $p=0.57$). The Zebra finch auditory system is not a homogeneous structure, it is broken into regions with differing neuron densities and functional encoding properties. Figure 3b breaks down decoder performance by anatomical region. Encoder performance was significantly related to anatomical region (ANOVA, $F(4, 317)=13.5$, $p < 0.01$). Performance was highest for regions in Field L, and lowest for region NCM.

Two Oscillatory Subnetworks

We observed in Figure 2 that the 5-30Hz LFP exhibited activity at two different time-frequency scales. The first timescale was a high amplitude, high bandwidth negative deflection that responded to syllable onsets, while the second was a lower bandwidth, lower amplitude, oscillating response. We analyzed the structure of the linear filters, plotted by

region in Figure 4a. The filters had two major components, an initial sharp downward deflection, responsible for the onset response, and a slower oscillatory component that helped predict the resonant properties of the LFP. We quantified this oscillatory component by fitting it with a sine curve (see Methods - Linear Filter Frequency Fitting). In figure 4b we report the center frequencies of the oscillatory components. Surprisingly, the center frequencies fall into a bimodal distribution, one in the 10-15Hz band, and another sharply peaked around 20Hz. To quantify this, we pooled data across regions and fit a bimodal Gaussian Mixture Model to the data. The bimodal GMM outperformed a unimodal model (Likelihood Ratio Test, deviance=6.9, $p < 0.01$), and best fit the data with two Gaussian distributions, one centered at 16Hz with a standard deviation of 3Hz, and another centered at 19Hz with a standard deviation of 4Hz. The analysis was re-run within each anatomical region wtih similar results. To summarize, our linear encoder analysis and investigation of the filters revealed two subpopulations in the auditory network that resonate at different frequencies, found in every anatomical region.

Recurrent Neural Networks Outperform Linear Models

Although linear models performed very well for some regions, the predictions shown in Figure 2 are far from perfect. They miss many of the essential modulations that occur throughout the stimulus. In an attempt to fit the more nonlinear aspects between the temporal envelope and LFP, we trained recurrent neural networks (RNNs) to predict 5-30Hz activity from the stimulus amplitude envelope. A recurrent network is a trainable nonlinear filter, and we utilized the commonly used backpropagation through time algorithm to fit networks to the data (see Methods - Training Recurrent Neural Networks). Our RNN models outperform linear models on virtually every electrode, as shown in Figure 5, providing an average boost in correlation coefficient of 0.08 (paired t-test, t=-28.9, N=447, $p<0.01$).

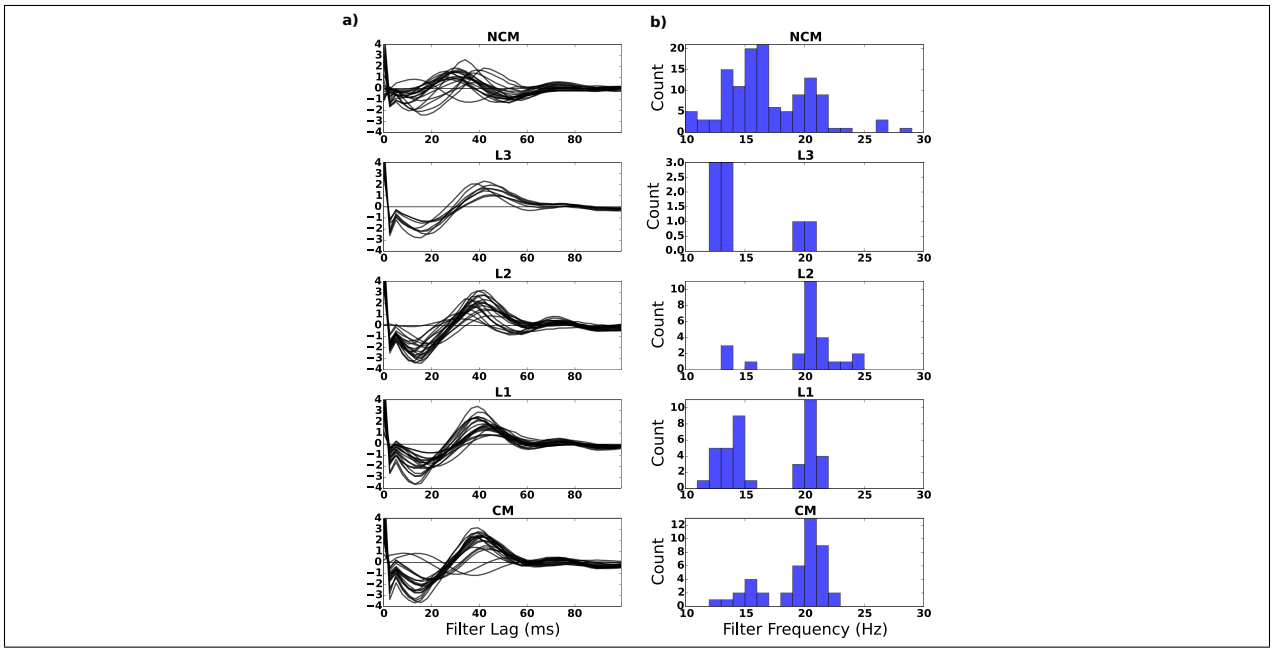

Figure 2.4: Distinct oscillatory subnetworks We analyzed the linear filters of the encoder models to determine how they mapped the temporal envelope to the LFP. **(a)** Filters across anatomical regions had a common theme, a sharp response to amplitude envelope in the first 5ms, followed by a slower oscillatory component. **(b)** We quantified the oscillatory component for each filter by its best fit frequency, and here report the distribution of filter frequencies by anatomical region.

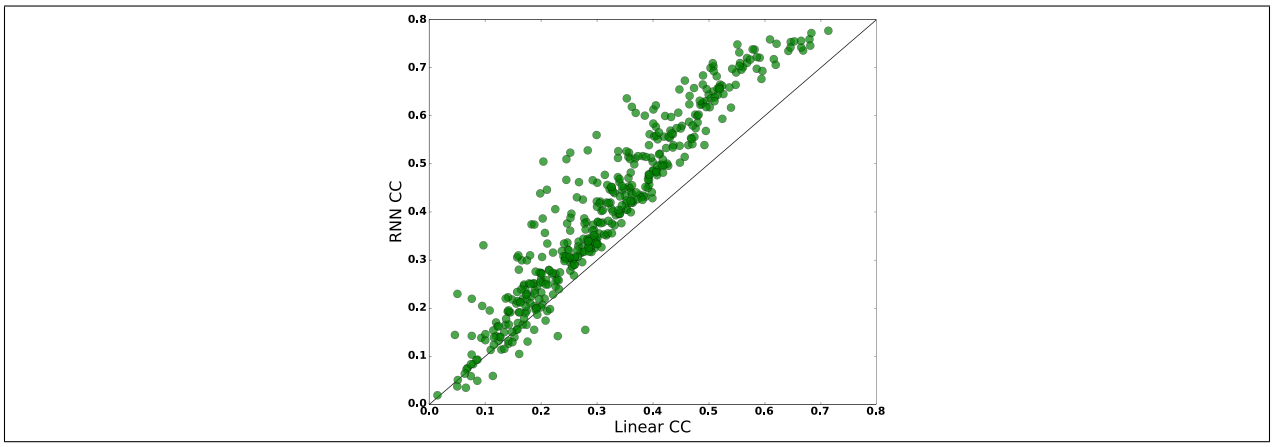

Figure 2.5: Performance enhancements from using RNNs We fit recurrent neural networks to the data to predict the multi-electrode 0-30Hz LFP from the temporal envelope. We compared performance of linear encoders on the x-axis, with the performance of the RNN encoders on the y-axis. Nearly all the points lie above the unity line y=x, indicating that RNNs outperform linear models.

2.4 Discussion

We first showed that the temporal modulations of the stimuli we utilized for the experiment fell in the 0-30Hz range, a range similar to the temporal modulations of human speech ([30], [31]). Notably, song stimuli had significant power in their temporal modulation spectra from 6-10Hz, which may be the "natural" frequencies of Zebra finch song. Begging calls, emitted by juveniles, occupied a similar range of frequencies as song, suggesting perhaps that Zebra finch brains are tuned to resonate at these frequencies, and juveniles take advantage of these natural frequencies to manipulate their parents.

Through linear modeling, we showed that the local field potential can be successfully predicted from the temporal envelope of our stimuli. The reverse direction, predicting the temporal envelope from multi-electrode LFP, was accomplished in human auditory cortex by [34]. We found encoder performance to be region-specific, with encoders performing the best in thalamorecipient regions, and the worst in secondary auditory region NCM. This could be due to differences in NCM activity for awake vs anesthetized subjects, or perhaps the function of neurons in NCM does not require the close following of the temporal envelope. There is evidence that neurons in NCM serve to "de-noise" stimuli and are more invariant to changes in background noise and amplitude ([36], [37]).

We were surprised to find two distinct oscillatory subnetworks in the Avian auditory system, consistent across every anatomical region. This complements the hypothesis of mammalian-like columnar microcircuitry suggested by [23] and [38], who proposed the canonical microcircuit connects auditory thalamus to L2, then L2 to L1, then L1 to region CM. Its possible that in order to communicate with each other, these regions must match oscillation frequencies in order to synchronize information transfer, a method of neural computation termed communication by coherence proposed by [39]. It will be important in future work to study whether these oscillation frequencies change on a per-stimulus basis.

Finally, we showed that we could boost encoder performance by utilizing recurrent neural networks. In Figure 6 we show the predictions of the RNNs vs predictions of linear models.

The RNN does a better job at fitting downward deflections that occur at stimulus onsets, but, disappointingly, does not do much better at capturing the non-onset oscillatory activity. We are actively engaged in work to improve both the performance and interpretability of these models. On the performance front, we had temporal limitations that prevented us from testing a fuller set of hyperparameters, and only utilized up to 100 neurons in the RNN. The LFP exhibits a significant amount spontaneous activity during silent periods, and there are RNN constraints that have been developed to ensure that population activity stays fixed at a baseline level throughout its lifetime [40]. Also, given that gamma oscillations originate from the interplay of excitatory and inhibitory neurons [18], we could constrain our network to be topologically organized and also enforce neurons to be either excitatory or inhibitory, as was done in [41]. Building multi-input, multi-output MIMO encoders with recurrent neural networks constrained to be more biologically plausible will allow us to "peek in the black box" of neural activity to further our understanding of neural dynamics.

2.5 Methods

Electrophysiological methods are fully described in [25] and [3] respectively and summarized here below. We then describe the methods used to preprocess the LFP, fit and analyze linear filter encoders to predict the 5-30Hz LFP, and train recurrent neural networks (RNNs) to predict the multi-variate 5-30Hz LFP waveforms. 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). The electrophysiology experiments were performed on 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 [26], acoustic signatures and behavioral contexts were used to classify vocalizations into nine different categories. A detailed description of call categories can be found [25]. We provide here, a very succinct summary of that characterization for a subset of the calls analyzed here and used in the neurophysiological experiments.

Song is a multi-syllable vocalization emitted only by males. Songs are comprised of repeating motifs of syllables, and in the dataset, have a duration of 1424 +/- 983ms. Song in zebra finches are used in pair bonding and mating behavior. The repertoire contains 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. Zebra finch also produce software calls used principally in the initial stages of pair bonding. 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.

Two calls are emitted by juveniles only. 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 and injected topically with lidocaine in order to remove a patch of skin over the skull and cement a head-holding fixture. On recording day, the subject was fasted for one hour, anaesthetized with urethane, head-fixed in a stereotaxic device, and two small 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. 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 different renditions in series with inter-syllable intervals chosen to match what was observed naturally.

Sound Preprocessing

Vocalizations used in the experiment were transformed into a spectrogram using custom python software. The spectrogram was computed by first applying a short-time Fourier transform (STFT) to the raw sound pressure waveform. The waveform was broken into overlapping segments of length 7ms, spaced apart by a sample interval of 1/381Hz, and segments were multiplied by a Gaussian window of length 6 standard deviations. The Fourier transform was then computed, the absolute value was taken, and squared, to produce the power spectrum corresponding to that segment. The log of the spectrogram was then taken.

To compute the stimulus amplitude envelope, the spectrogram was summed across frequencies at each time point. To compute the temporal modulation frequencies, each stimulus amplitude envelope was isolated, and a spectrogram was computed using a Gaussian window of length of 1s, with increments of 200ms. The windowed segments were averaged to produce the final temporal modulation spectrum.

LFP Preprocessing

The local field potential was recorded with a sample rate of 381 Hz. The LFP on each electrode was z-scored across time for the duration of a stimulus protocol. The LFP was then bandpassed with a pass band of 5-30Hz using a 5th order Butterworth filter implemented in Scipy.

Linear Encoder Fitting

We denote the value of the sound amplitude, for time t during the stimulus protocol played at a site, as $x(t)$. We denote the value of the LFP on electrode k, at time t during the stimulus protocol, as $u_k(t)$. Our goal was to build a linear filter model that predicted the LFP at a given time from the recent history of sound amplitude. To fit a filter using linear regression, we first created a feature vectors comprised of value of the stimulus amplitude envelope at time t, as well as D lags into the past:

$$\mathbf{x}_t = [x(t) \ ... \ x(t-D)]$$

D was chosen so that 500ms of recent stimulus history were taken into account. We then constructed a data matrix X comprised of the feature vectors:

$$X = [\mathbf{x}_1 \ ... \ \mathbf{x}_N]^T$$

where N is the length of the stimulus protocol, with silent periods excluded. 500ms of zeros were inserted between stimuli so that they did not interfere with each other during fitting. N was typically around 10^6. The dependent variable was then the bandpassed LFP time series:

$$\mathbf{y} = [u_k(1) \ ... \ u_k(N)]$$

A linear filter was found that optimally mapped the stimulus history vector to the bandpassed LFP by minimizing the sum of squares error in prediction:

$$L(X, \mathbf{y}, \mathbf{w}) = ||(X\mathbf{w} + b) - \mathbf{y}||^2$$

We utilized Ridge regression with scikits.learn to do this regularization. Ridge regression computes the optimal linear filter $\mathbf{w}$ as:

$$\mathbf{w} = (X^T X - \lambda I)^{-1} X^T \mathbf{y}$$

the value λ is a user-defined hyperparameter, high values of λ force weights towards zero.

The value of the hyperparameter λ was fixed to 1.0, because observations from several electrodes demonstrated very similar results for many different values of λ. 10-fold cross validation was used to compute the cross-validated correlation coefficient. The N samples were split into 10 partitions. For each partition, the rest of the data was trained on, and the correlation coefficient was computed between the prediction on the partition and the actual time series. We report here the correlation coefficients averaged across the ten partitions of data.

Linear Filter Frequency Fitting

To determine the frequency of the oscillatory component of a linear filter, we first isolated the filter between 5ms and 80ms lags, and normalized it by dividing by the absolute maximum. We then fit a sine curve $sin(2\pi tf + \phi)$, where f was the center frequency and ϕ was a phase offset, both free variables. We utilized the Scipy curvefit function to fit the curves.

Training Recurrent Neural Networks

We utilized a recurrent neural network (RNN) architecture to predict multi-electrode LFP activity in the 5-30Hz range from the time-varying stimulus amplitude envelope x(t). Let $\mathbf{u}(t)$ be the time-varying multi-electrode LFP:

$$\mathbf{u}(t) = [u_1(t) \ ... \ u_M(t)]$$

where M is the number of electrodes. Our recurrent neural network architecture uses recurrently connected hidden units to nonlinearly filter the time-varying input $x(t)$. A weighted combination of hidden unit activity is used to make a prediction of the multivariate LFP at time t. Let $\mathbf{z}(t)$ be a vector of hidden unit states at time t for a recurrent network of D neurons:

$$\mathbf{z}(t) = [z_1(t) \ ... \ z_D(t)]$$

The dynamics of our recurrent networks are given by:

$$\mathbf{z}(t+1) = \sigma\left(R\mathbf{z}(t) + Wx(t) + \mathbf{b}\right)$$

where σ is the logistic sigmoid function, R is a DxD matrix of recurrent weights, W is a Dx1 matrix of input weights, and $\mathbf{b}$ is a vector of D bias weights. The prediction of the LFP at time t is given as a weighted combination of hidden states:

$$\hat{\mathbf{u}}(t) = W_{out}\mathbf{z}(t)$$

where W_{out} is a MxD matrix of output weights.

We used a truncated backpropagation through time algorithm (BPTT) [42] to simultaneously fit R, W, b, and W_{out}. The data was broken into segments of length τ_{mem}=50 time points. The cost function that was minimized for each segment was specified as:

$$L(x(t), \mathbf{u}(t), R, W, \mathbf{b}, W_{out}) = \sum_{\tau=\tau_i}^{\tau_f} ||\hat{\mathbf{u}}(\tau) - \mathbf{u}(\tau)||^2 + \lambda_2 \sum_{ij} R_{ij}^2 + \lambda_1 \sum_{ij} |R_{ij}|$$

where λ_1 and λ_2 are hyperparameters for L1 and L2 regularization, and τ_i and τ_f are the start and end times of the segment, respectively. It should be noted that the cost function requires the value of $\mathbf{z}(\tau_i - 1)$ to be specified. We trained segments sequentially, so that the hidden state corresponding to the end of the previously segment was used as the initial state for the next segment. This provided the network the opportunity to extend memory well past the segment size τ_{mem}.

Bibliography

[1] Theunissen F. E., Sen K., and Doupe A. J. "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds". In: *The Journal of Neuroscience, 20(6), 2315-2331* (2000).

[2] Formisano E., Kim D. S., Di Salle F., et al. "Mirror-symmetric tonotopic maps in human primary auditory cortex." In: *Neuron, 40(4), 859-869.* (2003).

[3] 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).

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

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

[6] 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).

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

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

[9] 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).

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

[11] 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).

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

[13] 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).

[14] 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).

[15] 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).

[16] Reimann M. W., Anastassiou C. A., Perin R., 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).

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

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

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

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

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

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

[23] Wang Y., Brzozowska-Prechtl A., and Karten H. J. "Laminar and columnar auditory cortex in avian brain". In: *Proceedings of the National Academy of Sciences, 107(28), 12676-12681* (2010).

[24] Meliza C. D. and Margoliash D. "Emergence of selectivity and tolerance in the avian auditory cortex". In: *The Journal of Neuroscience, 32(43), 15158-15168* (2012).

[25] 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).

[26] R. A. Zann. *The zebra finch: a synthesis of field and laboratory studies*. Oxford University Press, 1996.

[27] Suthers R. A., Goller F., and Hartley R. S. "Motor dynamics of song production by mimic thrushes". In: *Journal of neurobiology, 25(8), 917-936* (1994).

[28] Friedman J., Hastie T., and Tibshirani R. *Elements of Statistical Learning, 2nd Edition*. Springer, Berlin: Springer series in statistics, 2009.

[29] Brainard M. S. and Doupe A. J. "Translating birdsong: songbirds as a model for basic and applied medical research." In: *Annual review of neuroscience, 36, 489.* (2013).

[30] Aiken S. J. and Picton T. W. "Human cortical responses to the speech envelope." In: *Ear and hearing, 29(2), 139-157* (2008).

[31] Ding N., Patel A., Chen L., et al. "Temporal Modulations Reveal Distinct Rhythmic Properties of Speech and Music." In: *bioRxiv, 059683.* (2016).

[32] Singh N. C. and Theunissen F. E. "Modulation spectra of natural sounds and ethological theories of auditory processing." In: *The Journal of the Acoustical Society of America, 114(6), 3394-3411.* (2003).

[33] Nagel K. I. and Doupe A. J. "Temporal processing and adaptation in the songbird auditory forebrain." In: *Neuron, 51(6), 845-859.* (2006).

[34] Pasley B. N., David S. V., Mesgarani N., et al. "Reconstructing speech from human auditory cortex." In: *PLoS Biol, 10(1), e1001251* (2012).

[35] Hsu A., Woolley S. M., Fremouw T. E., et al. "Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons." In: *The Journal of neuroscience, 24(41), 9201-9211* (2004).

[36] Schneider D. M. and Woolley S. M. "Sparse and background-invariant coding of vocalizations in auditory scenes." In: *Neuron, 79(1), 141-152.* (2013).

[37] Moore R. C., Lee T., and Theunissen F. E. "Noise-invariant neurons in the avian auditory cortex: hearing the song in noise." In: *PLoS Comput Biol, 9(3), e1002942.* (2013).

[38] 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).

[39] P. Fries. "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence." In: *Trends in cognitive sciences, 9(10), 474-480.* (2005).

[40] Krueger D. and Memisevic R. "Regularizing RNNs by stabilizing activations." In: *arXiv preprint arXiv:1511.08400.* (2015).

[41] Song H. F., Yang G. R., and X. J. Wang. "Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework." In: *PLoS Comput Biol, 12(2), e1004792.* (2016).

[42] Werbos PJ. "Backpropagation through time: what it does and how to do it." In: *Proceedings of the IEEE, 78(10), 1550-1560* (1990).