Journal Section: Behavioral/Systems/Cognitive

Title/Abbreviated Title: Incorporating Naturalistic Correlation Structure Improves Spectrogram Reconstruction From Neuronal Activity in the Songbird Auditory Midbrain

Authors: Alexandro D. Ramirez[12], Yashar Ahmadian[23], Joseph Schumacher[1], David Schneider[1], Sarah M. N. Woolley[14], Liam Paninski[123]

Affiliations: 1. Doctoral Program in Neurobiology and Behavior 2. Center for Theoretical Neuroscience, 3. Department of Statistics, 4. Department of Psychology, Columbia University, New York, NY 10027

Corresponding Author
Alexandro D. Ramirez
Columbia University
Room 930
1255 Amsterdam Ave.
New York, NY 10027
phone: 512-470-6553
email: adr2110@gmail.com

Number of figures: 10

Contents of supplemental material: Appendix, 2 audio files

Number of pages: 36

Number of words: Abstract(175), Introduction(480), Discussion(1319)

Keywords: birdsong, Auditory, extracellular recording, Receptive Field

Acknowledgements
We thank Ana Calabrese and Stephen David for helpful comments and discussions. We thank Columbia University Information Technology and the Office of the Executive Vice President for Research for providing the computing cluster used in this study. ADR is supported by an NSF Graduate Research Fellowship. LP is funded by a McKnight scholar award, and an NSF CAREER award. SMNW is funded by NSF, NIDCD, and the Searle Scholars Fund. LP and SMNW are supported by a Gatsby Initiative in Brain Circuitry pilot grant. YA is supported by Robert Leet and Clara Guthrie Patterson Trust Postdoctoral Fellowship, Bank of America, Trustee.

Abstract

Birdsong is comprised of rich spectral and temporal organization which might be used for vocal perception. To quantify how this structure could be used, we have reconstructed birdsong spectrograms by combining the spike trains of zebra finch auditory midbrain neurons with information about the correlations present in song. We calculated maximum a posteriori (MAP) estimates of song spectrograms using a generalized linear model of neuronal responses and a series of prior distributions, each carrying different amounts of statistical information about zebra finch song. We found that spike trains from a population of MLd neurons combined with an uncorrelated Gaussian prior can estimate the amplitude envelope of song spectrograms. The same set of responses can be combined with Gaussian priors that have correlations matched to those found across multiple zebra finch songs to yield song spectrograms similar to those presented to the animal. The fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations than temporal correlations. However the best reconstructions combine MLd responses with both spectral and temporal correlations.

1 Introduction

Understanding the neural mechanisms that subserve vocal perception and recognition remains a fundamental goal in auditory neuroscience (Eggermont, 2001; Theunissen & Shaevitz, 2006). The songbird has emerged as a particularly useful animal model for pursuing this goal because its complex vocalizations are used for communication (Catchpole & Slater, 1995; Gentner & Margoliash, 2002). Behavioral experiments have shown that songbirds can discriminate between similar, behaviorally relevant sounds (Lohr & Dooling, 1998; Shinn-Cunningham et al., 2007), use song for establishing territorial boundaries (Peek, 1972; Godard, 1991) and in mate preference (O'Loghlen & Beecher, 1997; Hauber et al., 2010). While the ethological importance of songbird vocalizations is well known, the neural basis underlying vocal recognition remains unknown.

The idea that song is processed by neurons which selectively respond to features of the song's time-varying amplitude spectrum (spectrogram) has been quantified by modeling neuronal responses using spectrotemporal receptive fields (STRFs) (Eggermont et al., 1983; Decharms et al., 1998; Theunissen et al., 2000; Sen et al., 2001; Woolley et al., 2006; Calabrese et al., 2010). These models can successfully predict neuronal responses to novel stimuli with a high degree of accuracy. In particular, neurons in the auditory midbrain, the mesencephalicus lateral dorsalis (MLd), region have STRFs that can be categorized into independent functional groups which may function in detecting perceptual features in song such as pitch, rhythm and timbre (Woolley et al., 2009). Midbrain responses from single and multiple neurons have also been used, without the STRF model, to discriminate among conspecific songs (Schneider & Woolley, 2010).

These results provide compelling evidence that zebra finch auditory midbrain neurons are tuned to specific spectrotemporal features that could be important for song recognition. Here, we tested whether responses encode enough information about song so that an 'ideal observer' of MLd spike trains could reconstruct song spectrograms. This method of assessing the information about stimuli preserved in neural responses by reconstructing the stimulus is well studied (Hesselmans & Johannesma, 1989; Bialek et al., 1991; Rieke et al., 1995; Rieke et al., 1997; Mesgarani et al., 2009; Pillow et al., 2010; Koyama et al., 2010) and some of the earliest applications have been in the auditory system. Hesselmans and Johannesma (1989) created coarse reconstructions of a grassfrog mating call, represented using a transformation known as the Wigner coherent spectrotemporal intensity

density, using neural responses from the frog auditory midbrain. Rieke et al (1995) used stimulus reconstruction to show that auditory nerve fibers in the frog encode stimuli with naturalistic amplitude spectra more efficiently than broadband noise and Mesgarani et al (2009) have recently used stimulus reconstruction to study the effects of behavioral state on responses properties of ferret auditory cortex.

Like most natural sounds, zebra finch songs have highly structured correlations across frequency and time, statistical redundancies that the nervous system might use for perceiving sound (Singh & Theunissen, 2003). To test how these statistical redundancies could be used in song recognition we asked whether reconstructions based on MLd responses, and a novel Generalized Linear Model of these responses (Calabrese et al., 2010), improve when responses are combined with prior knowledge of correlations present across zebra finch songs. We tested whether the fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations rather than temporal correlations and we examined how the filtering properties of MLd neurons affect reconstruction. Finally we compare spectrogram reconstructions under a Generalized Linear Model of responses to reconstructions based on the more common method of linear regression.

2 Methods

All procedures were in accordance with the NIH and Columbia University Animal Care and Use Policy. Thirty-six Adult male zebra finches (Taeniopygia guttata) were used in this study.

2.1 Electrophysiology

The surgical and electrophysiological procedures used have been described elsewhere (Schneider & Woolley, 2010). Briefly, zebra finches were anesthetized two days prior to recording with a single injection of 0.04 ml Equithesin. After administration of lidocaine, the bird was placed in a stereotaxic holder with its beak pointed 45 degrees downward. Small openings were made in the outer layer of the skull, directly over the electrode entrance locations. To guide electrode placement during recordings, ink dots were applied to the skull at stereotaxic coordinates (2.7 mm lateral and 2.0 mm anterior from the bifurcation of the sagittal sinus). A small metal post was then affixed to the skull using dental acrylic. After surgery the bird recovered for two days.

Prior to electrophysiological recording, the bird was anesthetized with three injections of 0.03 ml of 20 percent urethane, separated by 20 minutes. All experiments were performed in a sound-attenuating booth (IAC) where the bird was placed in a custom holder 23 cm away from a single speaker. Recordings were made from single auditory neurons in the MLd using either glass pipettes filled with 1M NaCl (Sutter Instruments) or tungsten microelectrodes (FHC, Inc.) with a resistance between three and 10 MΩ (measured at 1kHz). The duration of the recording sessions ranged from 4 to 15 hours. Awake recording sessions were no longer than 6 hours. For a single animal, awake recordings were performed over a period of approximately two weeks. Electrode signals were amplified (1000x) and filtered (300-5000 Hz; A-M Systems). A threshold discriminator was used to detect potential spike times. Spike waveforms were upsampled 4x offline using a cubic spline function, and action potentials were separated from non-spike events by cluster sorting the first three principal components of the action potential waveforms (custom software, Matlab). The number of neurons used in the analysis varied from 1 to 189.

2.2 Auditory Stimuli

Stimuli consisted of a set of 20 different adult male zebra finch songs sampled at 48,828 Hz, and frequency filtered between 250 and 8000 Hz. Each song was presented, in a pseudorandom order, 10 times at an average intensity of 72 dB SPL. Song duration ranged from 1.62 to 2.46 seconds, and a silent period of 1.2 to 1.6 seconds separated the playback of subsequent songs. All songs were unfamiliar to the bird from which recordings were made.

2.3 Bayesian Decoding

In the Bayesian framework, the spectrogram decoding problem is equivalent to determining the posterior probability distribution, $p(s|n, \theta)$, for observing a spectrogram, s, given the measured neural responses, n, and parameters θ. In principle, the posterior contains all available information about s. We use different statistics from this distribution, for example the mode or mean, to reconstruct the particular stimulus presented to the animal.

The encoding model specifies the likelihood, $p(n|s, \theta)$, which assigns probabilities to spike trains given the stimulus and parameters. The posterior distribution is related to the encoding model by Bayes rule,

$$p(s|n,\theta) \quad = \quad \frac{p(n|s,\theta)p(s)}{p(n|\theta)}, \tag{1}$$

where $p(s)$ is the prior distribution over song spectrograms. Here, we reconstruct song spectrograms using single and multiple neurons and different prior distributions (see below) that systematically add information about the birdsong spectrotemporal statistics.

2.4 Encoding Model

For a population of N midbrain neurons, we model the number of spikes fired by neuron i at time t by a random variable n_{it}, where i can range from 1 to N and t from 1 to T. We must assume that neurons are conditionally independent given the stimulus since we recorded cells one by one. Under this assumption, the likelihood in eqn. 1 is given by

$$p(n|s,\theta) \quad = \quad \prod_{t=1}^{T}\prod_{i=1}^{N} p(n_{it}|s,\theta,n_{i1},...,n_{i,t-1}). \tag{2}$$

We discretize time into bins of width dt, and model the conditional distribution for n_{it} given the spectrogram, spike-history up to time t and parameters, θ, as Poisson

$$p(n_{it}|s,\theta,n_{i1},...,n_{i,t-1}) \quad = \quad \exp\left(-r_{it}dt\right)\frac{(r_{it}dt)^{n_{it}}}{n_{it}!}, \tag{3}$$

where r_{it} is the instantaneous firing rate of the ith neuron at time t. r_{it} is given as the output of a generalized linear model (GLM). The GLM, and its application to neural data, has been described in detail elsewhere (Brillinger, 1988; McCullagh & Nelder, 1989; Paninski, 2004; Truccolo et al., 2005; Calabrese et al., 2010) and we only give a brief overview. The GLM for r_{it} applies a non-linearity (we use an exponential) to a linear mapping of input stimuli. As discussed in a recent paper (Calabrese et al., 2010) the model's ability to predict spikes slightly improves with this nonlinearity. In addition, the exponent prevents the model firing rate from taking on negative values and allows us to tractably fit the model to experimental data. The linear mapping is characterized by b_i, a stimulus-independent parameter which models baseline firing, k_i which will be referred to as the spectrotemporal receptive field (STRF) as it performs a linear mapping of

stimulus to response, and a "spike-history" filter, $h_i(\tau)$, which allows us to model neuronal effects such as firing-rate saturation, refractory periods, and/or bursting behavior. Even though the GLM conditional distribution, $p(n_{it}|s, \theta, n_{i1}, ..., n_{i,t-1})$, is Poisson, the joint spike train, $n_{i1}, ..., n_{i,T}$, does not follow a Poisson process because of the feedback from the spike-history filter. This procedure for mapping stimuli onto neural responses is schematized in Figure 1, which shows STRFs derived from data and shows simulated spike responses produced by the GLM.

Denoting the spectrogram by $s(f, t)$ (f indicates the spectral bin number and t denotes the temporal bin number), the firing rate r_{it} is modeled as

$$r_{it} = \exp\left(b_i + \sum_{f'=0}^{F-1}\sum_{\tau'=0}^{M-1} k_i(f', \tau')s(f', t-\tau') + \sum_{j=1}^{J} h_i(j)n_{i,t-j}\right), \quad (4)$$

where F is the total number of frequency bins in the spectrogram, M is the maximal lag-time of the STRF, and J is the maximal lag-time of the spike-history filter. Unless explicitly stated otherwise, spectrograms were temporally binned at 3 ms with 35 linearly spaced frequency bins ($F = 35$) from 400 to 6000 Hz. The power density of all spectrograms is log transformed so that units of power are expressed in decibels. We set $M = 7$ (21 ms) and $J = 10$ (30 ms). Model parameters, $\theta_i = \{b_i, k_i, h_i\}$ for $i = 1...N$ are fit from MLd responses to conspecific song using L1-penalized Maximum Likelihood (Lee et al., 2006). See Calabrese et al., (2010) for full details about the penalized fitting procedures.

2.5 Birdsong Priors

Eqn. 1 shows that song reconstruction depends on $p(s)$, the prior distribution of power spectral densities present in spectrograms. We test how song reconstruction depends on prior distributions that have the same spectrotemporal covariations present in song. We used several Gaussian priors because these distributions only depend on the covariance and mean. Other distributions might lead to better reconstructions by providing information about higher-order statistics in song, but are much more complicated to fit and optimize over. All Gaussians had the same frequency dependent mean but each had its own covariance matrix. All prior parameters were computed using the same songs as those used to collect the data (see Auditory Stimuli above). These songs appear to be sufficient to estimate the prior parameters under the Gaussian models presented below. Estimating

the prior parameters for more complicated models may require more data which can be obtained by using more bird songs than the ones used to collect the neural data. Each song is reconstructed with a prior whose parameters are estimated from all songs in the data set, except the one being reconstructed. The prior mean, $\hat{\mu}_f$, was found by assuming temporal stationarity of song and computing the empirical average power density across all temporal bins in the song data set.

2.5.1 Non-Correlated Gaussian Prior

To measure how well a population of midbrain neurons alone could reconstruct the spectrogram, we used a minimally informative prior. The least informative prior we used is an uncorrelated Gaussian

$$p(s) = \prod_{f=1}^{F} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\hat{\sigma}_f^2}} \exp\left(-\frac{(s(f,t)-\hat{\mu}_f)^2}{2\hat{\sigma}_f^2}\right), \quad (5)$$

where $\hat{\mu}_f$ is the empirical average power density discussed above, and $\hat{\sigma}_f^2$ is the empirical variance of songs in our data set at each frequency bin f. This prior does not provide information about spectral and/or temporal correlations in song. The prior variance is estimated by the empirical variance of songs in our data set. Figure 2 shows a histogram of spectrogram power density values across all spectrogram bins in the song data set (blue dots) and a univariate Gaussian with mean and variance equal to that found in the data.

2.5.2 Spectrally-Correlated Gaussian Prior

Next we measured how well spectrograms can be reconstructed when midbrain neuronal responses are combined with prior knowledge of spectral correlations across multiple conspecific songs. To do this we used a Gaussian prior whose covariance matrix only depended on frequency. Writing the covariance in spectrogram power between one time and frequency bin, $\{t, f\}$, and another, $\{t', f'\}$, as $C(\{t, f\}, \{t', f'\})$, this prior covariance is written as:

$$C(\{t, f\}, \{t', f'\}) = \Phi(f, f')\delta(t - t'), \quad (6)$$

where $\delta()$ is the Dirac delta function. The prior distribution is given by

$$p(s) \quad = \quad \prod_{t=1}^{T} \frac{1}{(2\pi)^{\frac{F}{2}} |\Phi|^{\frac{1}{2}}} \exp\left(-\frac{(s(.,t) - \hat{\mu})^T \Phi^{-1} (s(.,t) - \hat{\mu})}{2} \right), \tag{7}$$

where we use s(.,t) to denote the column vector of power density across frequencies at time t. The Φ matrix is empirically fit from example songs:

$$\Phi(f, f') \quad = \quad \frac{1}{N_t - 1} \sum_{n=1}^{N_t} \Big(s(f,n) - \hat{\mu}_f \Big) \Big(s(f',n) - \hat{\mu}_{f'} \Big), \tag{8}$$

where N_t is the total number of time-bins in the data set. N_t can be different from T, because T refers to the number of time-bins in the spectrogram being reconstructed, whereas N_t is the number of time bins in the entire data set used for training. N_t =13,435 for the data set used here. Figure 3A (upper panel) plots the Φ matrix. The spectral correlations averaged across all songs are larger at higher frequencies.

2.5.3 Temporally-Correlated Gaussian Prior

In order to measure how well songs can be reconstructed when midbrain responses are combined with prior knowledge of temporal correlations across conspecific songs, we reconstructed spectrograms with a prior containing temporal correlations but no spectral correlations

$$C(\{t, f\}, \{t', f'\}) \quad = \quad C_T(t, t') \delta(f - f'). \tag{9}$$

The prior distribution is given by

$$p(s) \quad = \quad \prod_{f=1}^{F} \frac{1}{(2\pi)^{\frac{T}{2}} |C_T|^{\frac{1}{2}}} \exp\left(-\frac{(s(f,.) - \hat{\mu_f})^T C_T^{-1} (s(f,.) - \hat{\mu_f})}{2} \right), \tag{10}$$

where $s(f,.)$ denotes the column vector of power density across time at frequency bin f.

We estimated the covariance matrix C_T by modeling the temporal changes in power density at a given frequency bin f as a stationary, order p, Autoregressive (AR) process:

$$s'(f,t) \quad \equiv \quad s(f,t) - \hat{\mu}_f, \tag{11}$$

$$s'(f,t) \quad = \quad \sum_{i=1}^{p} a_i s'(f,t-i) + \hat{\sigma}'\epsilon_t, \tag{12}$$

the constant terms, $a_i, \hat{\sigma}'$, are model coefficients and ϵ_t is a white noise, Gaussian random variable with unit variance. We used the covariance of this AR process instead of the empirical temporal covariance matrix to construct C_T. This is beneficial because it allowed us to approximate song correlations with far fewer parameters. Without an AR model the number of nonzero values in the matrix C_T^{-1} would grow quadratically with T, the temporal size of the spectrogram. This is troubling because each matrix element must be estimated from data, and therefore the amount of data required for accurately estimating C_T^{-1} grows with T. The inverse covariance matrix, C_T^{-1}, under an AR model is given by the square of a sparse Toeplitz matrix, A (Percival & Walden, 1993)

$$A \quad = \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a_1 & \ddots & 0 & 0 & 0 & 0 \\ \vdots & a_1 & \ddots & 0 & 0 & 0 \\ a_p & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & a_1 & 1 & 0 \\ 0 & 0 & a_p & \cdots & a_1 & 1 \end{pmatrix}, \tag{13}$$

$$C_T^{-1} \quad = \quad \frac{A^T A}{\hat{\sigma}'^2}. \tag{14}$$

As seen in eqn. 14, when we estimate the correlations using an AR model the number of nonzero values in the matrix C_T^{-1} only depends on the parameters $a_i, \hat{\sigma'}$ and, importantly, is independent of T. Thus the amount of data required to accurately estimate C_T^{-1} using an AR model is independent of T. To fit the AR coefficients we used the Burg method to minimize the sum of squares error between the original and AR model power density (Percival & Walden, 1993). We combined the temporal changes across all songs and spectral bins to fit the AR coefficients. Figure 3A (lower panels) compares the correlations of a twenty six order (p=26) AR model with empirical temporal correlations, averaged across songs and spectral bins. There is a trade-off between increasing the order of the AR model for obtaining good fits to the birdsong correlation function and the memory required to store the inverse covariance matrix/computational time to reconstruct spectrograms. We set p=26 because lower order models did not do a sufficient job of capturing the dip and rise

present in the correlation function visible between 0 and 100 ms (see Figure 3A). Note that we do not show a covariance matrix because an AR process assumes that the covariance between time points t and t' only depends on the absolute difference or lag time between these points: $C_T(t,t') = C_T(|t-t'|)$; i.e. all necessary information is contained in the correlation function. It is clear from Figure 3A that the temporal correlation function of the AR model closely matches the empirical correlation function found directly from the data.

2.5.4 Gaussian Prior with Spectrotemporal Correlations

Finally, we measured how well songs can be reconstructed when midbrain responses are combined with the spectral and temporal correlations across conspecific songs. To do this we reconstructed songs using a Gaussian prior with covariance equal to the separable product of the previously described AR covariance matrix and the Φ matrix:

$$C(\{t,f\},\{t',f'\}) \quad = \quad \alpha\Phi(f,f')C_T(|t-t'|). \tag{15}$$

The factor α is set so that the marginal variance of $C(\{t',f'\},\{t',f'\})$ is matched to the average variance of the song spectrograms, $\sum_f \hat{\sigma}_f^2$.

The prior distribution is given by

$$p(s) \quad = \quad \frac{1}{(2\pi)^{\frac{FT}{2}}|C|^{\frac{1}{2}}} \exp\left(-\frac{(\vec{s}-\hat{\mu})^T C^{-1}(\vec{s}-\hat{\mu})}{2}\right). \tag{16}$$

Equation 15 shows that C has a particular block structure in which each element of the Φ matrix is multiplied by the C_T matrix. This structure is known as a Kronecker product and leads to computational advantages when manipulating the C matrix. For example, the inverse matrix C^{-1} also has a Kronecker product form. This is particularly advantageous because we can use this fact to compute the required matrix multiplication $C^{-1}(\vec{s}-\hat{\mu})$ in a time that scales linearly with the dimension T ($O(F^3T)$ instead of the usual $O((FT)^3)$ time). In order to do so, we must construct the spectrogram vector $\vec{s}$ so that same-time frequency bands are contiguous, $\vec{s} = \Big(s(.,1), s(.,2), ..., s(.,T)\Big)^T$.

The matrix C does not exactly match the correlations in birdsong because it assumes that

spectral and temporal correlations can be separated. Using a separable covariance matrix and our AR model is beneficial because we do not need to estimate and store the full (FT) x (FT) covariance matrix, a task which becomes infeasible as we increase the number of time bins in our reconstruction. Importantly, we wanted to find reconstruction algorithms that could be performed in a computationally efficient manner. As discussed below, the separability approximation allows us to reconstruct spectrograms in a manner that is much more efficient than using a non-separable matrix. To examine the validity of the separability assumption we computed an empirical covariance matrix, $\hat{C}(f, f', |t - t'|)$ without assuming separability:

$$\hat{C}(f, f', \tau) \quad = \quad \frac{1}{N_t} \sum_{i=1}^{N_t - |\tau| + 1} \Big(s(f, i) - \hat{\mu}_f \Big) \Big(s(f', i + \tau) - \hat{\mu}_{f'} \Big), \tag{17}$$

where N_t is again the total number of time-bins in the data set. In Figure 3B (middle panel under the title 'True Covariance') we plot the matrix $\hat{C}$ and compare it with the separable matrix used in this study (Figure 3B bottom panel under the title 'Approximate Covariance'). Each lag, τ, can be thought of as an index for an FxF frequency matrix. For example, the upper panels in Figure 3B plot these FxF matrices when the lag equals zero. The matrix $\hat{C}$ and its separable approximation plot these FxF matrices, one for each lag, next to each other. The two matrices fall to zero power at the same rate and are closely matched near zero lags. The separable approximation has less power in the off-diagonal frequency bands at intermediate lags but overall the separable approximation is fairly accurate.

To visualize the information about song provided by this prior, Figure 3C (bottom panel) shows a sample spectrogram drawn from this Gaussian. The differences between this sample and a typical song spectrogram (upper panel) are due to the separable approximation to the song covariance matrix and the Gaussian prior model for the distribution of song spectrograms. Comparing the two-dimensional power spectra (also called the modulation spectra) of song spectrograms and of this prior is another method for assessing the effects of assuming a separable matrix. Figure 3D shows that the prior distribution lacks the peak across spectral modulations at temporal modulations close to zero, but otherwise has a similar spectrum.

2.5.5 Hierarchical Model Prior

One clear failure of the previous prior models is that real songs have silent and vocal periods. We can capture this crudely with a two-state model prior. This prior consists of a mixture of two correlated Gaussian priors, and a time-dependent, latent, binary random variable, q_t, that infers when episodes of silence and vocalization occur. We refer to this model as the hierarchical prior. One of the Gaussian distributions has mean power and spectral covariance determined by only fitting to the silent periods in song, while the other has mean power and spectral covariance fit to the vocalization periods. The two covariance matrices are shown in Figure 4B (upper panel).

Vocalization periods and silent episodes are extracted from the spectrogram data set by using a slightly ad hoc method that works well for our purposes here. A hard threshold is placed on the total power density summed across spectral bins, a variable we call y_t, and on the power density variance, $\hat{\sigma}_t^2$, across spectral bins:

$$y_t = \sum_{f=1}^{F} s(f,t), \tag{18}$$

$$\hat{\sigma}_t^2 = \frac{1}{F-1}\sum_{f=1}^{F}\left(s(f,t) - \frac{y_t}{F}\right)^2, \tag{19}$$

$$q_t = \begin{cases} 1(\text{vocalization period}) & y_t \geq q_1^*,\ \hat{\sigma}_t^2 \geq q_2^* \\ 0(\text{silent period} & \text{otherwise} \end{cases} \tag{20}$$

Figure 4A shows an example spectrogram and associated state transitions found using the above thresholding procedure, with q_1^* set to one standard deviation below the mean power density in the song and q_2^* set to an empirically determined value of 90 dB^2.

We model q_t as a Markov process and fit the transition matrix using maximum likelihood with training data found by taking state transitions from real song. The data set used here consisted of 13,435 state samples. This procedure leads to the transition rates displayed in Figure 4B (lower panel). Temporal correlations come from the AR model covariance matrix (described above) and from the temporal correlations induced by the Markovian model for q. Modeling q_t as a Markov process captures features of state transitions found in song and allows us to decode spectrograms using well-known algorithms (see below). However, by using a Markov model we assume that state

durations are exponentially distributed, which only approximates the distribution of durations found in birdsong.

A sample from this prior is shown in Figure 4C. The differences between vocal and silent periods are more clearly pronounced in this sample than that of the correlated Gaussian prior (Figure 3A). Because of the large differences in spectral correlations between vocal and silent periods, samples from this model also show spectral correlations closer to those found in song.

2.6 Song Reconstructions

Most of our reconstructions will be given by the spectrogram matrix that maximizes the log of the posterior distribution (the *maximum a posteriori* or MAP estimate). Substituting eqns. 2 and 3 into eqn. 1, the objective function that we maximize is then

$$L(s,\theta) = \log p(s|n,\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T} \log p(n_{it}|s,\theta,n_{i1},...,n_{i,t-1}) + \log p(s) + \text{const} \quad (21)$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T} n_{it}\log r_{it} - r_{it}dt + \log p(s) + \text{const}, \quad (22)$$

where N is the total number of neurons used in decoding, θ refers to the encoding model parameters, r_{it} is the firing rate for the ith neuron at time t (computed via eqn. 4), and $p(s)$ denotes the prior distribution. We write the term $\log p(n|\theta)$ as 'const' because it is constant with respect to the stimulus. In general, MAP estimates are found by searching over probabilities for all combinations of power density in a spectrogram and determining the most probable configuration. This task can be extremely computationally difficult as the number of spectral and temporal bins in the estimate grows. However, this problem is computationally tractable using standard Newton-Raphson (NR) optimization methods with the likelihood and prior distributions discussed above (Paninski et al., 2009; Pillow et al., 2010). In general, NR optimization computes the optimal configuration in a time that is on the order of d^3 (written as $O(d^3)$), where d is the dimensionality of the quantity being optimized (in our case $d = FT$). This is because the rate-limiting step in NR optimization is the time required to solve a linear equation involving the matrix of second derivatives of the objective function, L, in eqn.21, which requires $O(d^3)$ time in general. The likelihood and AR model used here yield sparse, banded Hessian matrices which reduces the time for optimization to

$O(F^3T)$ (Paninski et al., 2009; Pillow et al., 2010). This speedup is critical since the dimensionality of the decoded spectrograms is around $d \sim 7000$.

Song reconstructions under the hierarchical prior are created using the posterior mean, $E[s|n]$. The posterior mean is an optimal statistic to use for reconstruction as it is the unique estimate which minimizes the averaged squared error between the reconstruction and presented spectrogram. Using a Gaussian prior we decoded spectrograms with the MAP estimate because it is computationally efficient and because $E[s|n] \approx MAP$ in this case (Ahmadian et al., 2010; Pillow et al., 2010). It is easier to compute $E[s|n]$ using Markov Chain Monte-Carlo (MCMC) sampling when we decode using the hierarchical prior. The idea behind MCMC is that if we can generate samples from the posterior distribution we can use these samples to estimate the mean (Robert & Casella, 2005). It is difficult to sample directly from the posterior distribution using the hierarchical prior described above. However, it is possible to generate samples from the joint distribution, $p(s, q|n, \theta)$ which can then be used to estimate $E[s|n]$ (q again refers to the vocalization state). By definition $E[s|n]$ is given by the following multi-dimensional integral:

$$E[s|n] = \int p(s|n,\theta) s ds \tag{23}$$

$$= \sum_{q_1=0}^{1} ... \sum_{q_T=0}^{1} \int p(s, q_1, ..., q_T|n, \theta) s ds. \tag{24}$$

The relationship in eqn.24 shows how $E[s|n]$ is related to the joint distribution. We do not compute the sum in eqn.24 directly but instead use samples from the joint distribution $p(s, q|n, \theta)$ to evaluate $E[s|n]$. The details are given in the appendix.

2.7 Simulating STRFs

We also examined how our results depended on the spectral filtering properties of the STRF. We compute MAP estimates using simulated STRFs that have no spectral blur, no history dependence and peak frequency locations sampled from a distribution fit to real STRFs. This distribution was empirically constructed by creating a histogram of spectral bins at which STRFs obtained their maximal value. Denoting the ith neuron's filter at frequency bin f and temporal lag τ by $k'_{if\tau}$, our

simulated STRFs take the form:

$$k'_{if\tau} = a'_i \delta_{f,f'_i} \delta_{\tau 0}, \tag{25}$$

where δ_{ij} is the Kronecker-delta function. We choose the values of a'_i and the new encoding model bias parameters, b'_i, to obtain model responses whose first and second moment are roughly matched to those of the true responses. For each neuron, we use k and b to compute the linear mapping:

$$x_{it} = b_i + \sum_{f'=0}^{F-1} \sum_{\tau'=0}^{M-1} k_i(f', \tau') s(f', t - \tau'). \tag{26}$$

Then we compute the median, $\tilde{x}_i$, and absolute median deviation, $|x_i \tilde{-} \tilde{x}_i|$ across time of x_i. Given $\tilde{x}_i$ and $|x_i \tilde{-} \tilde{x}_i|$, we algebraically determine values of a'_i and b'_i which yield equivalent linear medians and absolute median deviations when convolved with the input spectrogram s. In other words, we solve the following linear equations for a'_i and b'_i:

$$\tilde{x}_i = b'_i + a'_i \tilde{s}_{f'_i} \tag{27}$$

$$|x_i \tilde{-} \tilde{x}_i| = |a'_i| |s_{f'_i} \tilde{-} s_{\tilde{f}'_i}|. \tag{28}$$

Spike trains generated using simulated STRFs with parameters fit as described have first and second moments roughly matched to spikes generated from real STRFs (compare the raster plot in the middle panel of Figure 10B with the raster in the bottom panel of Figure 10B).

2.8 Optimal Linear Estimator

We compare our estimates with the optimal linear estimator (OLE) (Bialek et al., 1991; Warland et al., 1997; Mesgarani et al., 2009). In brief, this model estimates the spectrogram by a linear filter g_{ift} which linearly maps a population of spike responses n_{it} to a spectrogram estimate $\hat{s}(f, t)$:

$$n'_{it} = n_{it} - \frac{1}{T} \sum_{j}^{T} n_{ij} \tag{29}$$

$$\hat{s}(f,t) \quad = \quad \sum_{i=1}^{N}\sum_{j=0}^{\tau-1} g_{ifj} n'_{it-j} + \frac{1}{T}\sum_{j}^{T} \hat{s}(f,j), \tag{30}$$

where N is again the total number of neurons used in decoding and τ is the maximal lag used in the decoding filter. n' is the mean subtracted spike responses and is used to ensure that the OLE and spectrogram have the same means. The function g is found by minimizing the average, mean-squared error between $\hat{s}$ and the spectrogram s at each frequency bin. The solution to this problem (Warland et al., 1997; Mesgarani et al., 2009) is given by

$$g_f \quad = \quad C_{nn}^{-1} C_{ns(f)}, \tag{31}$$

where C_{nn} denotes the auto-covariance of neural responses and $C_{ns(f)}$ denotes the cross-covariance of the response with the temporal changes in bin f of the spectrogram. The amount of data required to accurately estimate the matrices C_{nn} and $C_{ns(f)}$ increases as the filter length and the number of neurons used in the estimation increases. We did not implement any regularization on the matrices C_{nn} and $C_{ns(f)}$ to deal with this problem (see Pillow et al (2010) for further discussion). As is customarily done (Theunissen et al., 2001), we assume stimuli and responses are stationary so that temporal correlations between two points in time, say t and t', only depend on the distance or lag between these points, $t - t'$. We compute the covariances up to a maximal lag of 18 ms using spectrograms with time binned into 3 ms intervals with 35 linearly spaced frequency bins from 250 to 8000 Hz. These values were chosen in an attempt to maximize OLE performance.

2.9 Measuring Reconstruction Accuracy

The quality of reconstructions is measured using the Signal-to-Noise ratio (SNR) which is defined as the variance in the original spectrogram divided by the mean-squared error between the original and estimated spectrograms. Each song is reconstructed 4 times using the responses to different presentations of the same song. Since there are 20 songs in the data set, we obtain 80 different samples of mean-squared error between the estimated spectrograms and original. The mean-squared error is estimated by averaging these estimates together. The estimator's stability is measured using the standard error, which is the sample standard deviation of these estimates divided by the square-root of our sample size (80). Songs were reconstructed using different numbers of neurons. The

neurons used for reconstruction were chosen by randomly sampling without replacement from the complete data set of neural responses.

We also examined reconstruction quality in the Fourier domain. For each prior used, we computed the coherence between the estimated spectrogram and the original. The coherence between the original spectrogram, S, and the reconstructed spectrogram $\hat{S}$ is defined as

$$C(\nu_1, \nu_2) = \frac{|R_{S,\hat{S}}(\nu_1, \nu_2)|}{[R_{S,S}(\nu_1, \nu_2) R_{\hat{S},\hat{S}}(\nu_1, \nu_2)]^{\frac{1}{2}}}, \tag{32}$$

$$R_{XY}(\nu_1, \nu_2) = \sum_{u=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \exp(-i2\pi u\nu_1) \exp(-i2\pi m\nu_2) \mathrm{cov}\{X(u,m)Y(0,0)\} \tag{33}$$

The cross-spectral density function, R, for each reconstruction-spectrogram pair was estimated using Welch's modified periodogram method with overlapping segments (Percival & Walden, 1993). For each pair, the spectrograms were divided into segments that overlap by 25% and whose length, L, was $\frac{1}{8}$ the number of time bins in the spectrogram. Within each segment we computed a modified periodogram of the form:

$$\hat{X}(\nu_1, \nu_2) = \sum_{u=1}^{U} \sum_{m=1}^{U} \exp(-i2\pi u\nu_1) \exp(-i2\pi m\nu_2) h(u,m) X(u,m) \tag{34}$$

$$\hat{R}_{XY}(\nu_1, \nu_2) = \frac{\hat{X}(\nu_1, \nu_2)\hat{Y}^*(\nu_1, \nu_2)}{\sum_{u=1}^{N} \sum_{m=1}^{N} h^2(u,m)}, \tag{35}$$

where Y^* denotes the complex conjugate of Y, h is known as a data taper, and U denotes the window size. Data tapers and the use of overlapping windows were used because they can reduce the bias and variance associated with estimating R by a naive periodogram using data of finite length (Percival & Walden, 1993). Data was zero-padded so that U equaled 256 even though the window length was variable. h was chosen to be the product of two Hanning windows:

$$h(u,m) = 0.25(1 - \cos(2\pi u/L))(1 - \cos(2\pi m/L)) \text{ for } u = 1, ..., L\,; m = 1, ..., L. \tag{36}$$

R was estimated by averaging the estimates across segments. C was computed as in eqn. 32 substituting in the estimated R for the cross-spectral density function. We plot coherence values in decibels, given by the base-10 logarithm of the coherence multiplied by a factor of 10.

3 Results

3.1 Two-alternative forced choice discrimination

We begin by asking if an ideal observer of MLd neural responses could discriminate between conspecific songs using the GLM encoding model. Zebra finches can accurately discriminate between songs based on a single presentation, so it is interesting to determine if MLd responses can also discriminate between songs. We performed multiple two-alternative forced choice tests using an optimal decision rule and counted the fraction of correct trials from the different runs. Each test consisted of two segments of different conspecific songs and the spike trains, from multiple neurons, produced in response to one of the segments (Figure 5A). Using the log-likelihood eqns(2 - 4), we evaluated the probability of observing the spike trains given both segments and chose the segment with the higher likelihood. This procedure is optimal for simple hypothesis tests (Lehmann & Romano, 2005). The likelihood depends on the STRFs of the neurons used for discrimination and thus this test directly measures if the information about song provided by these STRFs can be used for song discrimination.

Figure 5B shows the fraction of correct trials for four different response/segment durations, when single spike trains produced by 1, 104, and 189 neurons are used for decision-making. As expected, the probability of responding correctly increases with longer response durations and as more neurons are used for discrimination. Given the very short response duration of 3 ms single neurons discriminated at chance level. As the response duration increased to 102 ms the average fraction correct increased to around 70%. Combining the responses of 189 neurons led to a discriminability accuracy as great as that seen in behavioral experiments, 90-100% (Shinn-Cunningham et al., 2007), after response durations around 27 ms and perfect discrimination after durations of 100 ms. These results show that MLd responses can be used for single presentation conspecific song discrimination.

3.2 Decoding Song Spectrograms Using Single Neurons

The results discussed above are in agreement with previous studies showing that responses from auditory neurons in the forebrain area field L (Wang et al., 2007) and MLd (Schneider & Woolley, 2010) can be used for song discrimination. As in previous studies, MLd responses and our encoding

model could be used to determine the most likely song identity, given a pre-defined set of possible songs. Instead of directly comparing our results with previous methods we focused on a different problem; we asked if these responses contain enough information to reconstruct song spectrograms. Spectrogram reconstruction is a more computationally demanding task than song discrimination and is a better test of the information about song encoded by neuronal responses. As explained in the methods we use the *maximum a posteriori* (MAP) value to estimate the spectrogram. We first compute spectrogram reconstructions using single-trial responses from single MLd neurons to understand how MAP estimates depend on the STRF and prior information.

The upper panel of Figure 6A shows 250 miliseconds of a spectrogram which elicited the two spikes shown below the spectrogram. The spikes are plotted at the frequency at which that neuron's STRF reaches a maximum (the BF or best frequency). Below the evoked spike-train is the MAP estimate (see methods) computed without prior information of spectrotemporal correlations in song power (Figure 6B, upper panel) and with prior information (Figure 6B, lower panel).

When the stimulus is only weakly correlated with the neuronal response, i.e. when the convolved spectrogram, computed by $\sum_{f'=0}^{F-1}\sum_{\tau'=0}^{M-1} k(f',\tau')s(f',t-\tau')$, is much less than one, it is possible to calculate the MAP solution analytically (Pillow et al., 2010). As discussed in Pillow et al.(2010), the analytic MAP solution when a spike occurs is approximately equal to the neuron's STRF multiplied by the prior covariance matrix. Under our uncorrelated prior (see methods section 2.5.1) this is equivalent to estimating the spectrogram by the STRF scaled by the song variance. In the absence of spiking the analytic MAP solution is equal to the prior mean. We plot the frequency averaged prior mean, $\sum_f \hat{\mu}_f$, in green and adjacent to the MAP estimate we plot the STRF for this particular neuron. Comparing the STRF with the MAP estimate we see that the analytic solution for the weakly correlated case is valid for this example neuron. This solution is intuitive because it reflects the fact that the only information a MLd neuron has about the spectrogram is from its song filtering properties. It also illustrates a fundamental difficulty in the problem of song estimation; an MLd neuron only responds to a small spectral and temporal area of song power. Because of this, spikes from a single neuron, without prior correlation information, can only contribute information on the small spectrotemporal scales encoded by that neuron.

Information independent of MLd spike responses can aid in song reconstruction by using spectral and temporal correlations to interpolate bands filtered out by the STRF. The MAP solution using

our correlated Gaussian prior displays this intuitive behavior. Next to the MAP solution using a correlated prior we plot the neuron's STRF temporally convolved and spectrally multiplied with the prior covariance matrix

$$k'(f,\tau) \quad = \quad \sum_{f',\tau'} C_T(|\tau-\tau'|)\Phi(f,f')k(f',\tau'). \tag{37}$$

Comparing the MAP solution with k' we see that when a spike occurs, to a good approximation, the MAP estimates the spectrogram with a covariance multiplied STRF. The MAP estimate shows values of power after a spike occurs because it uses prior knowledge of the temporal correlations in song to infer spectrogram power at these times.

3.3 Population Decoding of Song Spectrograms

We expect reconstruction quality to improve as more neurons are used in decoding. From the intuition developed performing single neuron reconstructions, we guessed that each neuron, without prior information of song correlations, would estimate a STRF each time it spikes. With the diverse array of STRF patterns known to exist in MLd (Woolley et al., 2009), a population of MLd neurons might faithfully estimate a spectrogram, without prior information, by having each neuron estimate a small spectrotemporal area determined by its STRF.

In Figure 7A (upper panel) we plot 1.25 seconds of an example song that we attempt to reconstruct given the spike trains from 189 neurons. In Figure 7A (lower panel) we have plotted these responses, with each neuron's spikes plotted at the best frequency (BF) at which that neuron's receptive field reaches a maximal value. Neurons with the same BF are plotted on the same row. Figure 7B shows the MAP estimate using the uncorrelated prior. As in the single neuron case, during segments of song that result in few spikes from the population the song is roughly estimated by the prior mean (green segments). This MAP estimate does a good job of distinguishing areas of high power from silent periods. Examination of the population spike responses show that this property is due to the population's ability to respond in a manner temporally-locked to vocalization episodes.

3.4 Effect of Prior Information on Song Estimation

Reconstructions without prior information show discontinuous gaps in power during vocal periods. They also show sparse spectrotemporal correlations at frequencies above 4 kHz. As in the single-neuron case, these features probably reflect the fact that each neuron only filters small spectrotemporal segments of song. In addition, most STRFs have peak frequencies below four kHz (this is evident in the plot of spike responses ordered by BF). Intuitively we expect MAP estimates constructed with prior information of song correlations to enhance reconstructions by filling in these 'gaps' in power density.

Figure 7 shows how the MAP estimate given the responses from 189 neurons changes as prior information is added. We plot MAP estimates using a prior covariance matrix that contains temporal information but no spectral information (Figure 7C), spectral information but no temporal information (Figure 7D), and both spectral and temporal correlations (Figure 7E; see section 2.5 in the methods for details and Figure 10 for a plot of preferred frequency tuning across the neural population). Comparing these estimates with the estimate using an uncorrelated prior (Figure 7B) shows that information about the second-order statistics in zebra finch song enhances reconstructions by interpolating correlations in spectrogram areas not covered by MLd STRFs.

A clear improvement in estimates using the spectrally correlated priors occurs at times where spiking activity is sparse. At these times the MAP estimate using the uncorrelated prior equals the prior mean. When given knowledge of song correlations, the MAP uses the sparse activity in MLd to correctly infer the deflections from the mean that occur in the original song spectrogram. The correlations also help the MAP infer that, during episodes of high spiking activity, spectral bands above four kHz should have similar levels of power as below four kHz. In the supplementary material we provide the reconstruction in the time-domain created by combining the spectrogram in Figure 7E with random phase. For comparison purposes we also provide the original song constructed with randomized phase.

In Figure 8 we quantify how reconstruction quality improves as a function of the number of neurons used in decoding. We use the signal-to-noise ratio (SNR) (see methods for a definition of the SNR) as a quantitative method for evaluating reconstruction accuracy. As described in the methods, the neurons chosen for reconstruction were randomly chosen from the population. Figure 8A plots

example MAP estimates using the Gaussian prior with spectrotemporal correlations as a function of the number of neurons for a single example song. The associated value of SNR is given above the MAP estimate. The solid lines in Figure 8B show the SNR averaged across all songs and dashed lines show the standard error about these lines. A SNR value of one corresponds to estimating the spectrogram by a single number, the mean. Improvements in SNR reflect improved estimates in the correlations of song power about the mean. The colors denote which prior was used in computing the MAP estimate. As expected, the SNR from MAP estimates using prior spectrotemporal information (black line) grows the fastest, followed by the SNR from MAP estimates which only use spectral prior information (green line). The faster growth in SNR only using spectral prior information versus temporal information is probably due to the facts that MLd population responses already capture a good deal of temporal information, spectral information helps infer deflections from the mean at times of sparse activity, and most MLd neurons have STRFs with peak frequencies below 4 kHz.

Figure 8C plots the coherence between the reconstructions and original spectrograms. The coherence is a normalized measure of the cross-correlation between the original two-dimensional signal and estimate in the frequency domain. In all the plots the vertical axis is spectral modulations (in units of cycles/kHz). These frequencies are often referred to as the ripple density. The horizontal axis is temporal modulations (in units of Hz). We note that the coherence plot is not the same as the modulation power spectrum shown in Figure 3D. In all plots, the range of the coherence is limited from -10 decibels (dark blue), a coherence of 0.1, to 0 decibels, i.e. perfect coherence (shown in red). With the exception of the non-correlated prior, we see a high coherence for temporal modulations between -50 and 50 Hz and ripple densities between 0 and 0.6 cycles/kHz. When we analyzed the coherence within these frequencies we found that the average coherence is highest for the spectrotemporal prior, second highest for the spectral prior and smallest for the prior without covariance information. From this plot we conclude that prior knowledge of the stimulus correlations primarily aids in reconstructing lower temporal modulations and ripple densities.

It is interesting to compare the decoding performance just described with the optimal linear estimator (OLE), a simpler and more commonly used decoder (Mesgarani et al., 2009). As discussed in the methods, the OLE finds the estimate that minimizes the average-squared euclidean distance between the spectrogram being estimated and a linear combination of the responses. Figure 8

(magenta line) shows the growth in the SNR of the OLE using the same real responses as those used for the non-linear, Bayesian model. The OLE depends on spectrotemporal correlations in the stimulus so we compare its performance with the prior that contains both spectral and temporal correlations (black line). Comparing these two shows that when the number of neurons is low the two estimates perform similarly. As more neurons are added to the estimate, the MAP estimator outperforms the OLE. Recent work (Pillow et al., 2010) has shown that this behavior is expected if the encoding model is a good model for spike responses to stimuli and if the prior model does a good job of capturing stimulus correlations. Pillow et al.(2010) showed that when the number of neurons used for estimation is low, the MAP estimate and OLE are equivalent. As the number of neurons grows, the MAP estimate can outperform the OLE because the MAP estimator is not restricted to be a linear function of spike responses.

3.4.1 Hierarchical Prior Model

We observed visible differences in power density covariance and mean during silent and vocal periods (see the covariance matrices plotted in Figure 4 and the differences in color between silent and vocal periods in the plotted spectrograms). These differences are averaged together when constructing the covariance matrix and mean used in the single Gaussian prior. Averaging together the correlation information from these two different periods smooths the spectral correlation information (compare the covariance matrix in Figure 3A with that of Figure 4B (left panel)). We reconstructed songs using a hierarchical prior (see methods sec 2.5.5) to test whether this smoothing hinders the reconstruction performance. This prior includes a state variable which determines the mean and spectral covariance. We first study the case where all possible state trajectories are used for decoding, with trajectory probabilities determined by neural responses and the transition probabilities in our model (see methods and Figure 4). Each trajectory yields a different reconstructed spectrogram and the final estimate is determined by averaging across these reconstructions. This is equivalent to estimating the song using the posterior mean. This estimate should be better than an estimate using a single Gaussian if the neural responses provide sufficient information to properly infer state transitions.

In Figure 9A (left column) we plot an example song spectrogram with evoked single-neuron, single-trial responses immediately below. We have again plotted the responses at the neuron's

best frequency, which in this case is 1.8 kHz. Below this we have plotted the MAP estimate using these responses and a single correlated Gaussian prior (Figure 9B upper panel) and the posterior mean using the hierarchical prior (Figure 9B lower panel). The estimates show surprisingly similar behavior. Under the hierarchical prior we see power densities slightly closer to those in song, around the neuron's BF, compared to the estimate using a single Gaussian prior. Otherwise no large differences between the two estimators are seen.

It is possible that estimates based on the hierarchical model are not much better than using a single Gaussian because single neuron responses do not provide enough information to infer the state transitions. Figure 9C shows the average state transition given the neural response. We see that this is indeed the case and on average the inferred state transitions do not match those in the song being estimated. Given the above result we asked if the hierarchical model would outperform the single Gaussian prior when more neurons are used for decoding. In the right column of Figure 9A we plot the responses of 49 additional neurons (for a total of 50 neurons) with BFs slightly greater than the single neuron used in the left column. These responses are again plotted below the spectrogram being estimated. Examining the average state changes given responses in Figure 9C we see a closer resemblance between the inferred state transitions to those present in the estimated song. In the right column of Figure 9B we plot the posterior mean under the hierarchical prior and the MAP estimate using the same subset of neural responses combined with a single Gaussian (non-hierarchical) prior. The two estimators do not show any prominent differences. Adding more neurons to the estimation should only cause the two estimators to look more similar since the reconstructions will have less dependence on prior information when more data is included. Therefore we did not compute estimates with more than 50 neurons using the Hierarchical prior. Finally, we eliminated the portion of reconstruction error due to problems associated with estimating the state transitions by computing the MAP estimate of the Hierarchical prior given the true underlying state in the song being estimated. We compared this estimate, which has perfect knowledge of the underlying song transitions, to the MAP estimate using a single Gaussian prior. Even in this case we do not see any large differences between the estimators (data not shown). These results demonstrate that spectrogram estimates do not necessarily improve as more complicated prior information of song is included in the posterior distribution. While samples from the hierarchical prior contain more statistical information of song and arguably show more

resemblance to song than samples from a single, correlated Gaussian prior (compare Figure 4C with Figure 3C), this advantage does not translate into better spectrogram reconstructions.

3.5 Reconstruction Dependence On STRF Frequency Tuning

The information for reconstruction provided by an individual MLd neuron depends on its STRF. Neurons that have STRFs which overlap in their spectrotemporal filtering properties will provide redundant information. While this redundancy is useful for reducing the noise associated with the spike generation process (Schneider & Woolley, 2010), good spectrogram reconstructions also require enough neurons that provide independent information. We asked if our results would improve if we used neurons that had either no overlap in their filtering properties or complete overlap. We computed MAP estimates using simulated STRFs, which we will refer to as point STRFs, that have no spectral blur and no history dependence (see methods for how these receptive fields were constructed). Figure 10A plots a STRF (upper, left panel) calculated from real responses using the method of maximum likelihood ('full' STRF) and a point STRF (upper, right panel) with an equivalent peak frequency. In Figure 10A (lower, left panel), we show the extent of the blurring behavior in our neuronal population. For each neuron, we plot the spectral axis of its STRF at the latency which that STRF reaches its maximum value. The right panel shows the same information for the point STRFs (see methods for our determination of the number of neurons with a particular peak frequency).

Figure 10B (upper panel) shows an example spectrogram we attempt to reconstruct. For both STRFs we reconstructed songs using simulated responses. We did not use real responses because we wanted to reduce the differences in reconstruction performance caused by the poorer predictive performance of point STRFs on real data. Using simulated responses allowed us to better control for this effect and focus on differences in reconstruction performance due to spectral blurring. For comparison purposes, we plot the real responses of 189 neurons, aligned according to their BF, immediately below this spectrogram. The middle panel shows simulated responses to this example song created using the generalized linear model with point STRFs. The lower panel shows simulated responses using full STRFs. Using a correlated Gaussian prior, we reconstructed the spectrogram using the point STRFs and the simulated responses generated from them (middle panel) and using the full STRFs and their associated simulated responses (lower panel).

Stimulus reconstructions using point STRFs show slightly finer spectral detail compared to reconstructions using full STRFs. However, overall we do not find that spectral blurring of the full STRFs leads to much degradation in stimulus reconstructions. The growth in SNR for point STRFs and full STRFs as a function of the number of neurons is shown in Figure 10C. On average point STRFs have slightly higher signal-to-noise ratios as the number of neurons increases; however the difference between the two curves is not too great. It is important to point out that these results depend on the fact that reconstructions were performed using a correlated prior trained on natural stimuli. The spectrotemporal width of the covariance is broad compared to that of the full STRFs. When we reconstructed songs using a prior with no correlations, we found that full STRFs decode slightly better than the point STRFs (data not shown). Also, for the reasons stated above, we used simulated responses which also influences the results. Reconstructions using point STRFs are slightly worse than reconstructions with full STRFs when real data is used. We attribute this difference to the better predictive performance of the full STRFs on real data.

4 Discussion

We asked if the responses of zebra finch auditory midbrain neurons to song encode enough information about the stimulus so that an 'ideal observer' of MLd spike trains could recognize and reconstruct the song spectrogram. We found that 189 sequentially recorded MLd responses can be combined using a generalized linear model (GLM) to discriminate between pairs of songs that are 30 ms in duration with an accuracy equivalent to that found in behavioral experiments. These results are in agreement with prior studies showing that responses from auditory neurons in the forebrain area field L (Wang et al., 2007) and MLd (Schneider & Woolley, 2010) can be used for song discrimination. Importantly, this previous work did not use the GLM to evaluate the discriminability and thus provides an independent benchmark to compare with our GLM-dependent results.

We tested the hypothesis that the statistics of zebra finch song can be used to perform vocal recognition by decoding MLd responses to conspecific song using *a priori* knowledge of the joint spectrotemporal correlations present across zebra finch songs. We explicitly used prior information lacking higher-order information of song to test if MLd responses only require knowledge

of correlations to be used for spectrogram reconstruction. When we evaluated the reconstructed spectrograms in the Fourier domain, we found that these responses do a fair job of reproducing temporal and spectral frequencies, i.e. temporal modulations and ripple densities, between -50 and 50 Hz and below 0.6 cycles per kHz. When combined with the joint spectrotemporal correlations of zebra finch song we found an improvement in the coherence in these regions. These results did not change greatly when we used STRFs with non-overlapping best frequencies, suggesting that the spectral blur or 'bandwidth' limitations of the STRF did not strongly affect reconstruction performance using these responses combined with spectrotemporal correlations in zebra finch song.

None of the reconstructions using MLd neurons and the correlations present in song reproduced all the details of a particular spectrogram. These results are qualitatively similar to previous findings showing that the auditory system of zebra finch, as well as other songbirds, can recognize songs even when some of the fine details of the song signal have been degraded by various types of background noise (Bee & Klump, 2004; Appeltants et al., 2005; Narayan et al., 2006; Knudsen & Gentner, 2010). This may be similar to the finding that humans can recognize speech even after the spectral and temporal content has been degraded (Drullman et al., 1994; Shannon et al., 1995).

It is interesting to speculate if the song features that were reproduced in this study are relevant to the bird for song recognition. For example, we found that reconstructions were most accurate at low ripple densities and temporal modulations. Song recognition based on these features would be consistent with existing evidence that zebra finch are better able to discriminate auditory gratings with lower ripple density/temporal modulations (Osmanski et al., 2009). Because of the complexity of song it is difficult to quantify behaviorally relevant song features birds use for recognition and communication (Osmanski et al., 2009; Knudsen & Gentner, 2010). The spectrogram reconstructions reported here may serve as a useful probe for future discrimination studies. For example, one could compare discrimination thresholds between songs whose amplitude spectrums have been degraded according to the regions where reconstructions have low coherence with songs whose amplitude spectrums are randomly degraded. If the MAP reconstructions are relevant to the bird, we would expect performance to be worse on songs with randomly degraded amplitude spectrums. This idea is similar to the previously mentioned Osmanski et al. (2009) study testing discrimination of auditory gratings in birds; however the ripple density/temporal modulations used for probes would be more complex than simple gratings. Working with ferret auditory cortical neurons, Mesgarani et

al.(2009) have also recently examined the effects of stimulus correlation on spectrogram decoding. Similar to our findings, this group finds improvements in reconstruction quality when they use prior information of sound correlations. This suggests that the use of natural sound correlations for vocal recognition might be a general strategy employed across species. However, there are important distinctions between the Bayesian approach used here for reconstruction and the optimal linear decoder used in this previous work. The optimal linear decoder incorporates stimulus correlations via the stimulus-response cross-covariance matrix and the response auto-covariance matrix. The Bayesian decoder incorporates stimulus statistics using a prior distribution that is independent of the likelihood distribution used to characterize neural responses (eqn. 1). Therefore this decoder allows one to estimate song correlations independent of the amount of neural data available. This is beneficial for obtaining good estimates of song correlations when it is easier to obtain song samples than physiological data. Another important distinction between the linear decoder and the Bayesian method is that the Bayesian decoder does not have to be a linear function of the spike responses. This seems to be the reason for the Bayesian method's slight improvement over the linear decoder. When we decode songs using a linear, Gaussian, Bayesian decoder with the same correlated Gaussian prior as the one in this study we find worse reconstruction performance than the GLM. This suggests that the nonlinearity is an important factor in the GLM's improved performance.

Another advantage of separating prior information from neural responses is that we could systematically change the prior to study which statistical properties of song are most important for stimulus reconstruction without refitting the filters applied to the observed spike trains. We found that reconstructions based on MLd responses with *a priori* information of spectral correlations yielded better estimates of song than did reconstructions using temporal correlations present in song. While we cannot conclude from this study whether or not the bird actually uses a prior, we speculate that these results suggest what information, in addition to MLd responses, maybe used when the bird recognizes song. These results suggest that there is a greater benefit to the bird, in terms of vocal recognition capabilities, if MLd responses are processed by neuronal circuits that have access to the joint spectrotemporal or spectral song correlations rather than temporal correlations. This interpretation would be consistent with recent work showing that zebra finch appear to be more sensitive to frequency cues than temporal cues when categorizing songs belonging to

one of two males (Nagel et al., 2010). However, even though much work has been done relating information encoded within a prior distribution to neuronal spiking properties (Zemel et al., 1998; Beck & Pouget, 2007; Litvak & Ullman, 2009), it is unclear how to predict response properties of cells based on the statistical information about a stimulus they may be encoding. To better understand this relationship future experiments could perform a similar decoding analysis using the responses from other brain areas to look for spiking activity in which it is more beneficial to store temporal correlations rather then spectral correlations. If such activity exists, these responses could be combined with MLd spike trains to perform reconstructions which presumably would only show marginal improvement when combined with prior knowledge of either temporal or spectral correlations.

There has been much recent interest in determining good priors for describing natural sounds and stimuli(Singh & Theunissen, 2003; Karklin & Lewicki, 2005; Cavaco & Lewicki, 2007; McDermott et al., 2009; Berkes et al., 2009). With the two-state model we briefly explored the effects on reconstruction quality of prior distributions which contain more information than just the mean and covariance of birdsong, however none of the priors used in this study explicitly contain information about the subunits such as song notes, syllables or motifs typically used to characterize song (Catchpole & Slater, 1995; Marler & Slabbekoorn, 2004). Future work could examine if reconstruction quality changes using more realistic, non-gaussian prior distributions of birdsong which contain higher-order information. For example, neurons in the songbird forebrain nucleus HVc are known to be sensitive to syllable sequence (Margoliash & Fortune, 1992; Lewicki & Arthur, 1996; Nishikawa et al., 2008) suggesting that there are neural circuits which could provide prior information of sound categories such as syllables and motifs. One could therefore reconstruct songs using this prior information, for example by using a Hidden Markov Model (HMM) with the hidden states trained on sound categories (Kogan & Margoliash, 1998). While we didn't find much of an improvement in reconstruction quality using the two-state prior compared to a Gaussian prior, more realistic priors may yield better reconstructions. If so, one could determine additional statistical information about song stimuli, other than stimulus correlations, also useful for song recognition.

References

Ahmadian, Y., Pillow, J., & Paninski, L. (2010). Efficient Markov Chain Monte Carlo methods for decoding neural spike trains. *In press, Neural Computation*.

Appeltants, D., Gentner, T. Q., Hulse, S. H., Balthazart, J., & Ball, G. F. (2005). The effect of auditory distractors on song discrimination in male canaries (serinus canaria). *Behav Processes*, *69*, 331–41.

Beck, J. M., & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation*, *19*, 1344–1361.

Bee, M. A., & Klump, G. M. (2004). Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *J Neurophysiol*, *92*, 1088–104.

Berkes, P., Turner, R. E., & Sahani, M. (2009). A structured model of video reproduces primary visual cortical organisation. *PLoS Comput Biol*, *5*, e1000495.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., & Warland, D. (1991). Reading a neural code. *Science*, *252*, 1854–1857.

Brillinger, D. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cyberkinetics*, *59*, 189–200.

Calabrese, A., Schumacher, J., Schneider, D., Woolley, S. M. N., & Paninski, L. (2010). A generalized linear model for estimating receptive fields from midbrain responses to natural sounds. *Frontiers in Neuroscience Conference Abstract, Computational and Systems Neuroscience*.

Catchpole, C., & Slater, P. (1995). *Bird song: Biological themes and variations*. Cambridge.

Cavaco, S., & Lewicki, M. S. (2007). Statistical modeling of intrinsic structures in impacts sounds. *J Acoust Soc Am*, *121*, 3558–3568.

de Gunst, M., Kunsch, M., & Schouten, J. (2001). Statistical analysis of ion channel data using Hidden Markov Models with correlated state-dependent noise and filtering. *JASA*, *96*, 805–815.

Decharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, *280*, 1439–1443.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*, 197–208.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, *95*, 1053–64.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*, 216 – 222.

Eggermont, J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*, 1–42.

Eggermont, J. J., Aertsen, A. M. H. J., & Johannesma, P. I. M. (1983). Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. *Hearing Research*, *10*, 167–190.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gentner, T. Q., & Margoliash, D. (2002). *Acoustic communication*, chapter 7, The Neuroethology of Vocal Communication: Perception and Cognition. Springer-Verlag.

Godard, R. (1991). Long-term memory for individual neighbors in a migratory songbird. *Nature*, *350*, 228–229.

Hauber, M., Campbell, D., & Woolley, S. M. N. (2010). Functional role and female perception of male song in zebra finches. *Emu-Austral Ornithology*.

Hesselmans, G. H., & Johannesma, P. I. (1989). Spectro-temporal interpretation of activity patterns of auditory neurons. *Mathematical Biosciences*, *93*, 31 – 51.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comput*, *17*, 397–423.

Knudsen, D. P., & Gentner, T. Q. (2010). Mechanisms of song perception in oscine birds. *Brain and Language*, *In Press, Corrected Proof*, –.

Kogan, J. A., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *Journal of the Acoustical Society of America*, *103*, 2185–2196.

Koyama, S., Eden, U. T., Brown, E. N., & Kass, R. E. (2010). Bayesian decoding of neural spike trains. *Annals of the Institute of Statistical Mathematics*, *62*, 37–59.

Lee, S. I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient L1 Regularized Logistic Regression. In *Proceedings of the twenty-first national conference on artificial intelligence (aaai-06)*.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Lewicki, M. S., & Arthur, B. J. (1996). Hierarchical organization of auditory temporal context sensitivity. *J Neurosci*, *16*, 6987–6998.

Litvak, S., & Ullman, S. (2009). Cortical Circuitry Implementing Graphical Models. *Neural Computation*, *21*, 3010–3056.

Lohr, B., & Dooling, R. (1998). Detection of changes in timbre and harmonicity in complex sounds by zebra finches (Taeniopygia guttata) and budgerigars (Melopsittacus undulatus). *Journal of Comparative Psychology*, *112*, 36–47.

Margoliash, D., & Fortune, E. S. (1992). Temporal and harmonic combination-sensitive neurons in the zebra finch's hvc. *J Neurosci*, *12*, 4309–4326.

Marler, P. R., & Slabbekoorn, H. (2004). *Nature's music: The science of birdsong*. Academic Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall/CRC. Second edition.

McDermott, J. H., Oxenham, A. J., & Simoncelli, E. P. (2009). Sound texture synthesis via filter statistics. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. *Journal of Neurophysiology*, *102*, 3329–3339.

Nagel, K. I., McLendon, H. M., & Doupe, A. J. (2010). Differential influence of frequency, timing, and intensity cues in a complex acoustic categorization task. *J Neurophysiol*, *104*, 1426–37.

Narayan, R., Grana, G., & Sen, K. (2006). Distinct time scales in cortical discrimination of natural sounds in songbirds. *Journal of Neurophysiology*, *96*, 252–258.

Nishikawa, J., Okada, M., & Okanoya, K. (2008). Population coding of song element sequence in the bengalese finch hvc. *Eur J Neurosci*, *27*, 3273–3283.

O'Loghlen, A. L., & Beecher, M. D. (1997). Sexual preferences for mate song types in female song sparrows. *Animal Behaviour*, *53*, 835–841.

Osmanski, M. S., Marvit, P., Depireux, D. A., & Dooling, R. J. (2009). Discrimination of auditory gratings in birds. *Hearing Research*, *256*, 11 – 20.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*, 243–262.

Paninski, L., Ahmadian, Y., Ferreira, D., Koyama, S., Rahnama, K., Vidne, M., Vogelstein, J., & Wu, W. (2009). A new look at state-space models for neural data. *Journal of Computational Neuroscience*, *29*, 107–126.

Peek, F. W. (1972). An experimental study of the territorial function of vocal and visual display in the male red-winged blackbird (agelaius phoeniceus). *Animal Behaviour*, *20*, 112–118.

Percival, D. B., & Walden, A. T. (1993). *Spectral analysis for physical applications: Multitaper and conventional univariate techniques*. Cambridge University Press.

Pillow, J., Ahmadian, Y., & Paninski, L. (2010). Model-based decoding, information estimation, and change-point detection in multi-neuron spike trains. *(in press) Neural Computation*.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286.

Rieke, F., Bodnar, D. A., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B: Biological Sciences*, *262*, 259–265.

Rieke, F., Warland, D., de Ruyter, & Bialek, W. (1997). *Spikes: Exploring the neural code*. The MIT Press.

Robert, C., & Casella, G. (2005). *Monte Carlo statistical methods*. Springer.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, *16*, 351–367.

Schneider, D., & Woolley, S. M. (2010). Discrimination of communication vocalizations by single neurons and groups of neurons in the auditory midbrain. *Journal of Neurophysiology, in press*.

Sen, K., Theunissen, F. E., & Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *J Neurophysiol*, *86*, 1445–1458.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–4.

Shinn-Cunningham, B. G., Best, V., Dent, M. L., Gallun, F. J., Mcclaine, E. M., Narayan, R., & Sen, K. (2007). *Behavioral and neural identification of birdsong under several masking conditions*. Springer Berlin Heidelberg.

Singh, N., & Theunissen, F. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America*, *114*, 3394–3411.

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., & Gallant, J. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*, 289–316.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci*, *20*, 2315–2331.

Theunissen, F. E., & Shaevitz, S. S. (2006). Auditory processing of vocal sounds in birds. *Current Opinion in Neurobiology*, *16*, 400 – 407. Sensory systems.

Truccolo, W., Eden, U., Fellows, M., Donoghue, J., & Brown, E. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *Journal of Neurophysiology*, *93*, 1074–1089.

Wang, L., Narayan, R., Grana, G., Shamir, M., & Sen, K. (2007). Cortical discrimination of complex natural stimuli: Can single neurons match behavior? *Journal of Neuroscience*, *27*, 582–589.

Warland, D., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, *78*, 2336–2350.

Woolley, S., Gill, P., & Theunissen, F. (2006). Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *Journal of Neuroscience*, *26*, 2499–2512.

Woolley, S. M. N., Gill, P. R., Fremouw, T., & Theunissen, F. E. (2009). Functional Groups in the Avian Auditory System. *Journal of Neuroscience*, *29*, 2780–2793.

Zemel, R., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*, 403–430.

A Generating Samples from $p(s, q|n, \theta)$

To generate samples from the joint distribution, $p(s, q|n, \theta)$ described in section 2.5.5, we use a technique known as Gibbs sampling (Geman & Geman, 1984; Robert & Casella, 2005). Gibbs sampling works by iteratively sampling from the conditional distributions $p(s|q, n, \theta)$ and $p(q|s, n, \theta)$. First we initialize s to a matrix S_0; we take S_0 to be a matrix where each bin equals the vocalization period prior mean. Then we draw a sample, Q_0, from the conditional distribution $p(q|s = S_0, n, \theta)$ (note that our hierarchical model does not depend on n or θ so that $p(q|s = S_0, n, \theta) = p(q|s = S_0)$. From now on we will write this conditional distribution as $p(q|s)$). We then draw a sample, S_1, from the conditional distribution $p(s|Q_0, n, \theta)$. Iterating this process by drawing alternating samples

$$S_{i+1} \sim p(s|Q_i, n, \theta), \tag{38}$$

$$Q_{i+1} \sim p(q|s = S_{i+1}), \tag{39}$$

(where the notation $s \sim p(s)$ indicates that s is drawn from the distribution p) results in samples, (Q_i, S_i), which converge to samples drawn from the joint distribution $p(s, q|n, \theta)$ (Geman & Geman, 1984; Robert & Casella, 2005).

The hierarchical prior models a spectrogram, S_i, as a Gaussian vector with a mean, μ_q, and spectral covariance matrix, Φ'_q, that depend on the vector of vocalization states, q, as well as the previously defined (see methods) temporal covariance matrix, C_T. Such a Gaussian vector can be written as a linear transformation of an uncorrelated standard Gaussian vector. More precisely, the spectrogram can be written as

$$S_i = \Phi_q^{'\frac{1}{2}} Z_i C_T^{\frac{1}{2}} + \mu_q, \tag{40}$$

where $\Phi^{1/2}$ and $C_T^{1/2}$ are the matrix square roots of Φ and C_T and Z_i is sampled from an uncorrelated standard Gaussian distribution

$$Z_i \sim \prod_{f=1}^{F} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z_i(f,t)^2}{2}\right). \tag{41}$$

To determine $\mu_{q=0}$ and $\Phi'_{q=0}$ ($q = 0$ denotes a silent period) we construct a spectrogram, S',

composed of all silent periods extracted from the data set of birdsongs (determined as described in the methods) and determine the empirical mean and covariance of S'

$$\mu_0 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{F}\sum_{f=1}^{F}S'(f,n) \tag{42}$$

$$\Phi_0(f,f') = \frac{1}{N-1}\sum_{n=1}^{N}\left(S'(f,n)-\frac{1}{N}\sum_{n'=1}^{N}S'(f,n')\right)\left(S'(f',n)-\frac{1}{N}\sum_{n'=1}^{N}S'(f,n')\right), \tag{43}$$

where N is the total number of time-bins in the data set. The same is done to construct $\mu_{q=1}$ and $\Phi'_{q=1}$ except using a spectrogram composed of all vocal periods. Although eqn. 40 depends on $C_T^{\frac{1}{2}}$ we will show that sampling only requires computation of the matrix $C_T^{-\frac{1}{2}}$. As discussed in section 2.5.2, we construct C_T^{-1} from a sparse matrix of Autoregressive coefficients (eqn. 14). As evidenced by eqn. 14, the bandwidth of the matrix C_T^{-1} does not grow with T allowing us to sample even for large values of T.

To draw a sample from $p(q|s=S_i)$ we first multiply S_i by $C_T^{-\frac{1}{2}}$, from eqn. 40

$$S_i \rightarrow Y_i = S_i C_T^{-\frac{1}{2}} = \Phi_q^{'\frac{1}{2}} Z_i + \mu_q C_T^{-\frac{1}{2}}. \tag{44}$$

This is done to create a collection of spectral vectors, $Y_i(.,t)$, that are conditionally independent given the latent variable q. At time t, the spectral vector $Y_i(.,t)$ is a Gaussian random variable drawn from a distribution whose mean and variance is determined by the value of q_t. Since q is a Markov process, Y_i forms a collection of spectral vectors that are emissions from a hidden-Markov model. We sample the $q(t)$ element of the vector q from the distribution $p(q(t)|q(t+1)...q(T),Y)$ using the forward filter-backward sample algorithm (de Gunst et al., 2001). The algorithm uses the relation $p(q(t)|q(t+1)...q(T),Y) \propto p(q(t)|q(t-1))p(Y(.,1)...Y(.,t),q_t=n)$ to compute $p(q(t)|q(t+1)...q(T),Y)$. This relation is helpful because $p(q(t)|q(t-1))$ is the known transition matrix and the forward probabilities $\alpha_n(t) = p(Y(.,1)...Y(.,t),q_t=n)$ can be computed recursively using the conventional forward algorithm (Rabiner, 1989). Given a method for computing the probabilities $p(q(t)|q(t+1)...q(T),Y)$, q(t) can then be sampled using inverse transform sampling.

Samples from $p(s|q=Q_i,n,\theta)$ are generated using a modified form of the Metropolis-Hastings (MH) algorithm known as Hybrid Monte-Carlo (HMC) (Roberts & Rosenthal, 2001; Duane et al., 1987). A step-by-step description for implementing the HMC and its convergence properties can be

found elsewhere (Ahmadian et al., 2010; Roberts & Rosenthal, 2001). We simply note that a key step in efficiently sampling the distribution $p(s|q,n,\theta)$ is using the inverse Hessian matrix of the log posterior, $\log p(s|q,n,\theta)$, evaluated at the MAP to construct the proposal distribution (again see Ahmadian (2010)). As previously noted (see methods) taking the inverse of the Hessian would be computationally expensive (scaling like $O(d^3)$ where d=FT) if the Hessian was not banded. In our case, the Hessian of the GLM log-likelihood, J, and Hessians of our Gaussian distributions, $C^{-1}(q) = \Phi_q'^{-1} C_T^{-1}$, are banded, making the Hessian of the log-posterior, given by $J + C^{-1}(q)$, also banded. In addition to the matrix $C^{-1}(q)$, the HMC algorithm specifies the proposal distribution with two parameters, the number of 'leapfrog iterations', L, and proposal distribution jump size, σ, which must be set by the user. We set L=1 and $\sigma = 0.9$ because we find that these values lead to relatively quick convergence rates for spectrogram estimates with 111 time bins (0.3 s). At each step of the Gibbs sampler we run the HMC for 100 iterations and keep the last sample.

We reconstruct spectrograms using $E[s|n]$, which we approximate by averaging the conditional means $E[s|q = Q_i, n, \theta]$ because, by Rao-Blackwellisation (Doucet et al., 2000), this leads to a better estimate of $E[s|n]$ then averaging the samples S_i. For Gaussian priors, the MAP is often a good approximation for the posterior mean (Ahmadian et al., 2010)

$$\hat{s}_i \quad = \quad \arg\max_s p(s|q = Q_i, n, \theta) \approx E[s|q = Q_i, n, \theta], \tag{45}$$

therefore each time we sample from $p(s|q = Q_i, n, \theta)$ we also calculate and store the most probable spectrogram, $\hat{s}_i$, under this distribution. Using the approximation in eqn. 45, we compute $E[s|n]$ by averaging different values of $\hat{s}_i$

$$E[s|n] \quad \approx \quad \frac{1}{N_{samp}} \sum_{j=1}^{N_{samp}} \hat{s}_i. \tag{46}$$

We average together 100 independent Gibbs sampler chains. Each chain was created by iterating the Gibbs sampler 100 times, burning the first 50 iterations and keeping the last 50 samples. Each chain was created on a separate machine using Columbia University's Hotfoot cluster.

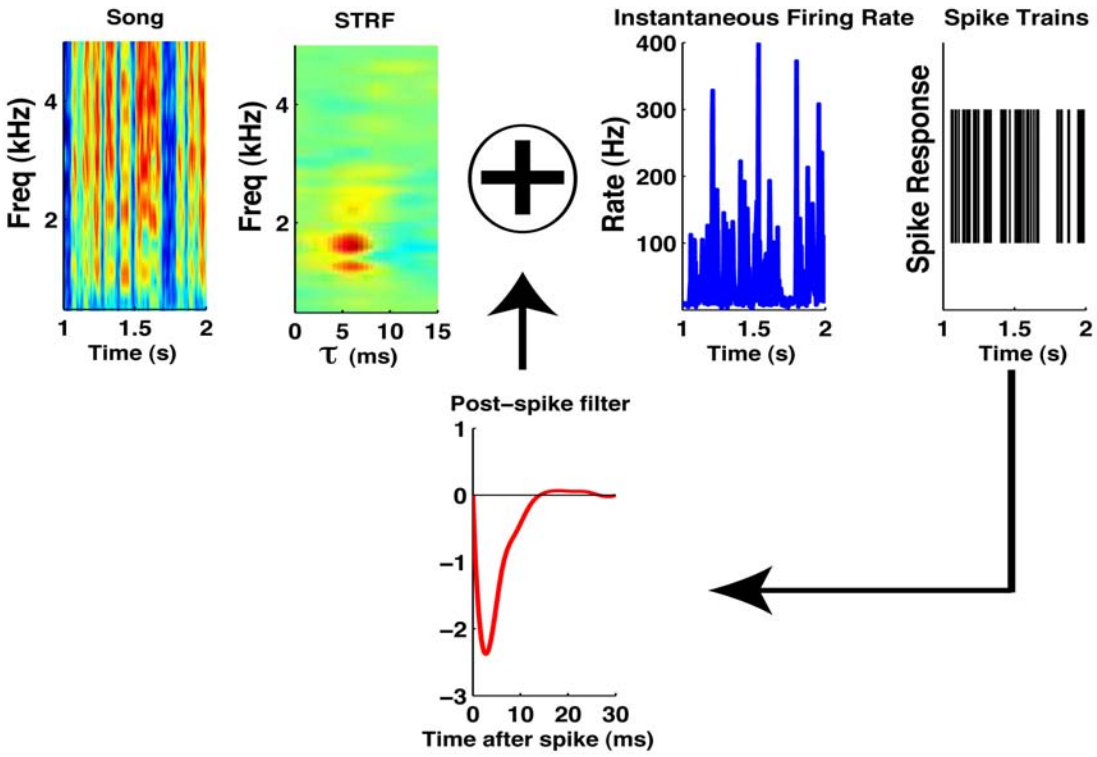

Figure 1: Encoding model and parameters. In the encoding model, each neuron is modeled with a spectrogram filter (STRF) and post-spike filter that captures stimulus-independent spiking properties. The stimulus is temporally convolved and frequency multiplied with the STRF and then exponentiated to obtain the instantaneous firing rate used for generating spikes. The spikes are convolved with the post-spike filter and used in the model as a feedback signal that affects future spike generation.

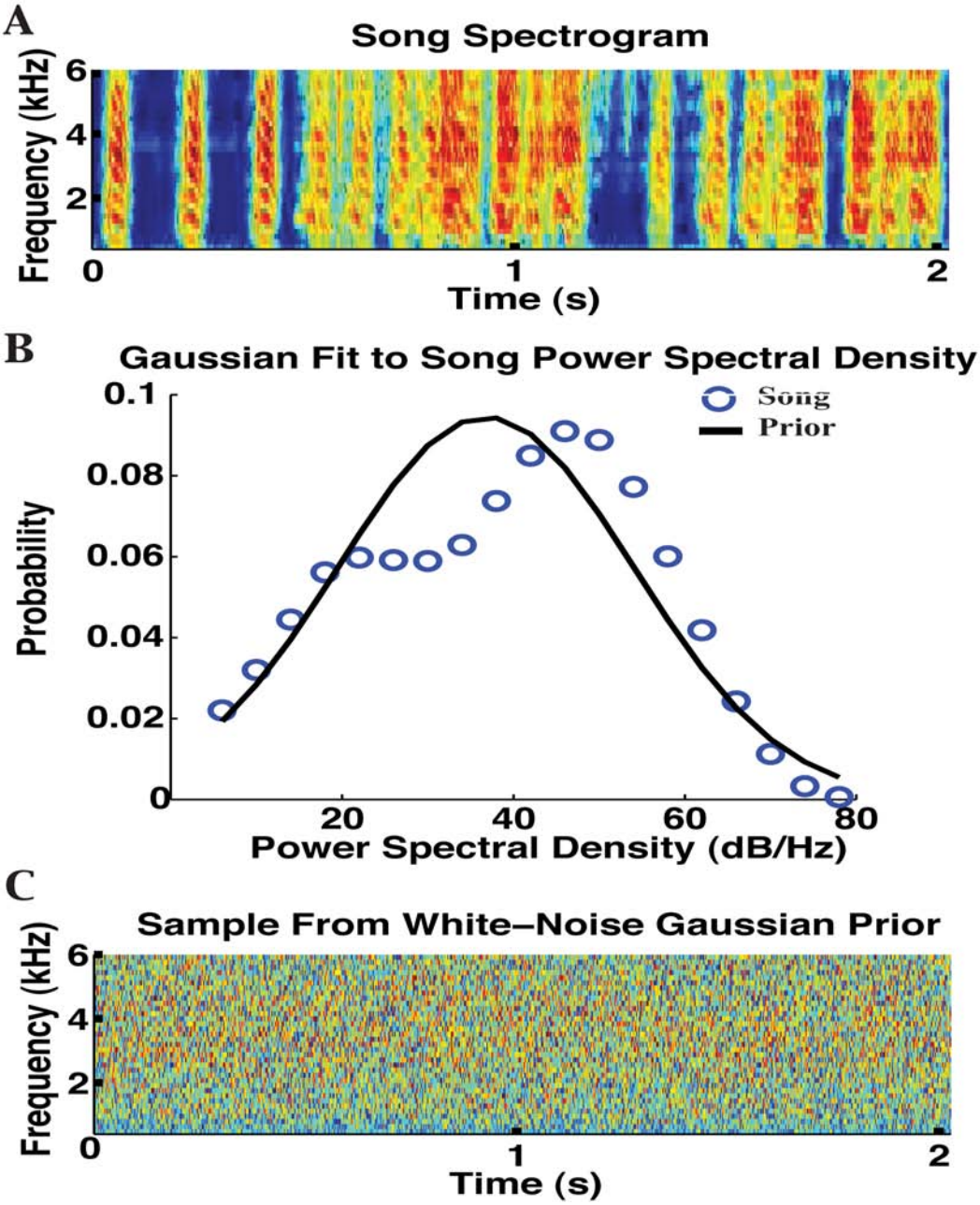

Figure 2: Least informative prior: uncorrelated Gaussian distribution. **A**, An example spectrogram with power spectral density indicated by color. **B**, Normalized histogram of power spectral density values across all songs and spectrogram bins (blue dots). The mean and variance of these power values is used to construct a Gaussian prior (black line) that confines estimated values of power spectral density to regions found in actual song spectrograms. **C**, To visualize the information provided by the prior, a sample spectrogram drawn from this prior is plotted. This prior does not provide information on spectrotemporal correlations in spectrograms, as demonstrated by this sample.

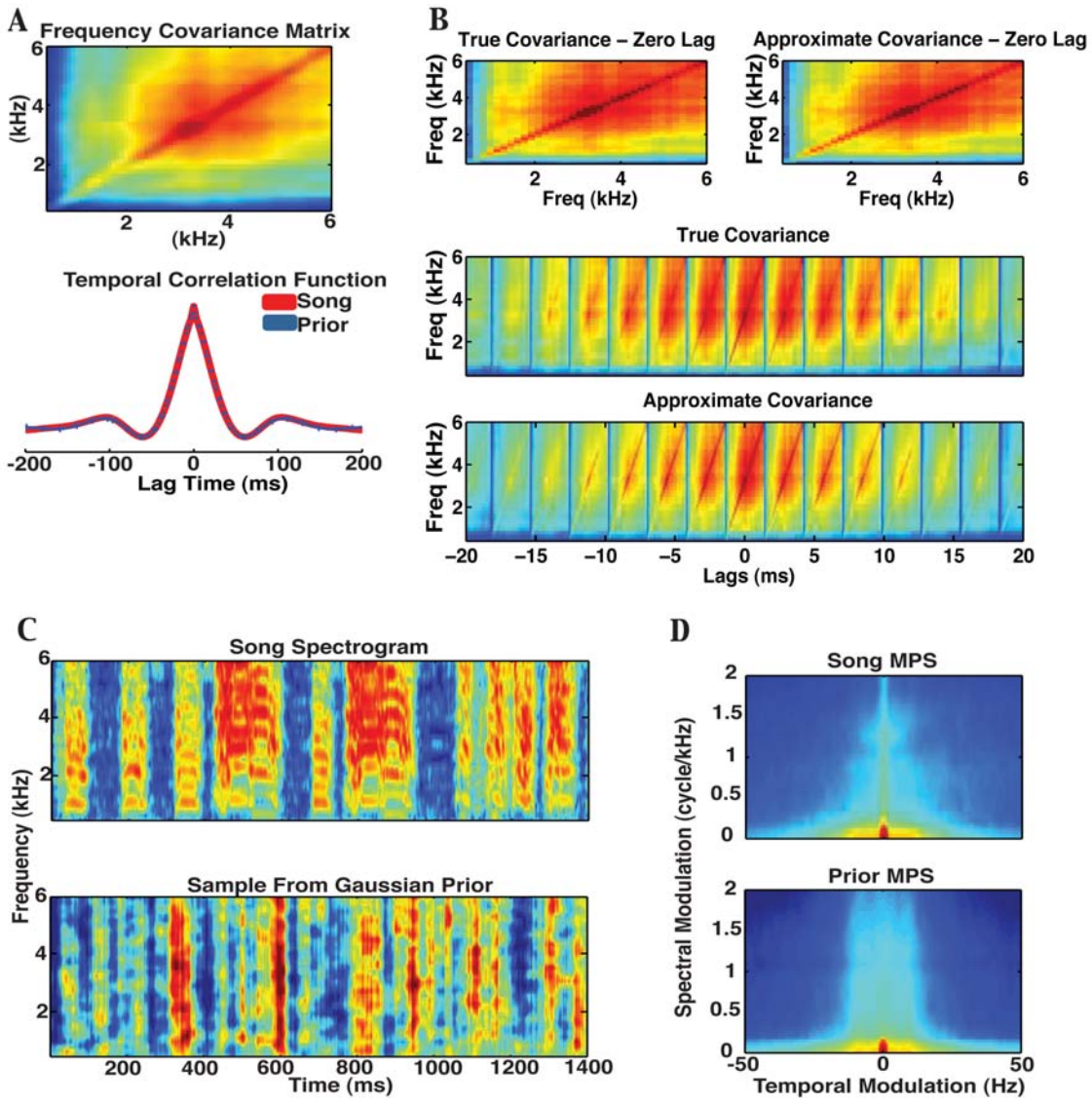

Figure 3: Spectrotemporaly Correlated Gaussian prior. **A**, The spectrotemporal covariance matrix is modeled as separable in frequency and time. The frequency component is the spectral covariance matrix (upper panel). The temporal component is fully described by the temporal autocorrelation function in song spectrogram power density (bottom panel, red line). The prior uses an approximation to this function using an Autoregressive model (blue line). **B**, The full spectrotemporal covariance matrix is a concatenation of several spectral covariance matrices, like those shown in the upper panel, each corresponding to the covariance at a different temporal lag. The bottom panel labeled 'Approximate Covariance' plots the separable covariance matrix and the middle panel labeled 'True Covariance' plots the non-separable covariance matrix. **C**, (Top) An example spectrogram used in determining song statistics for constructing the Gaussian prior. (Bottom) Sample spectrogram drawn from the Correlated Gaussian prior. **D**, Two-dimensional power spectra, also called the modulation power spectra (MPS), for song spectrograms (top) and for the prior (bottom); the prior does a good job of capturing information about spectrotemporal modulations except at joint regions of high spectral modulations and temporal modulations near zero.

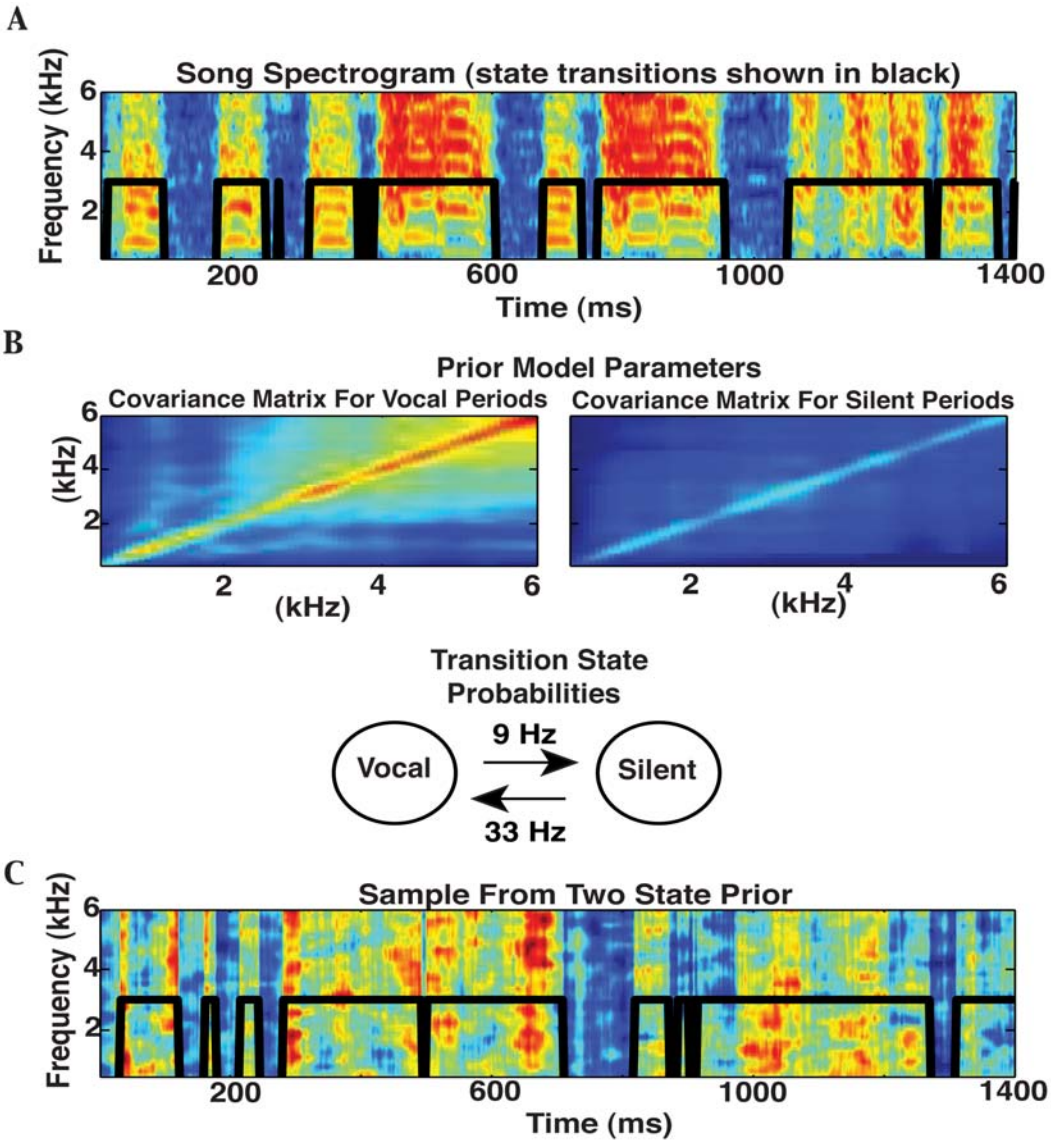

Figure 4: Most informative prior: hierarchical model with a two-state hidden variable that infers whether the spectrogram is in a vocalization or silent period. These periods have different statistical properties not captured by a single Gaussian prior. The state variable determines which spectral covariance matrix and mean the prior uses to inform reconstructions. **A**, Example spectrogram overlaid with vocalization and silent states (black line). **B**, (Top left) Spectral covariance matrix used during vocal periods. (Top right) Spectral covariance matrix used for silent periods. (Bottom) Prior information of transition rates between silent and vocal periods determined from song spectrograms. **C**, Sample spectrogram drawn from this prior; the sharp transitions in song statistics during vocal and silent periods better matches song spectrograms.

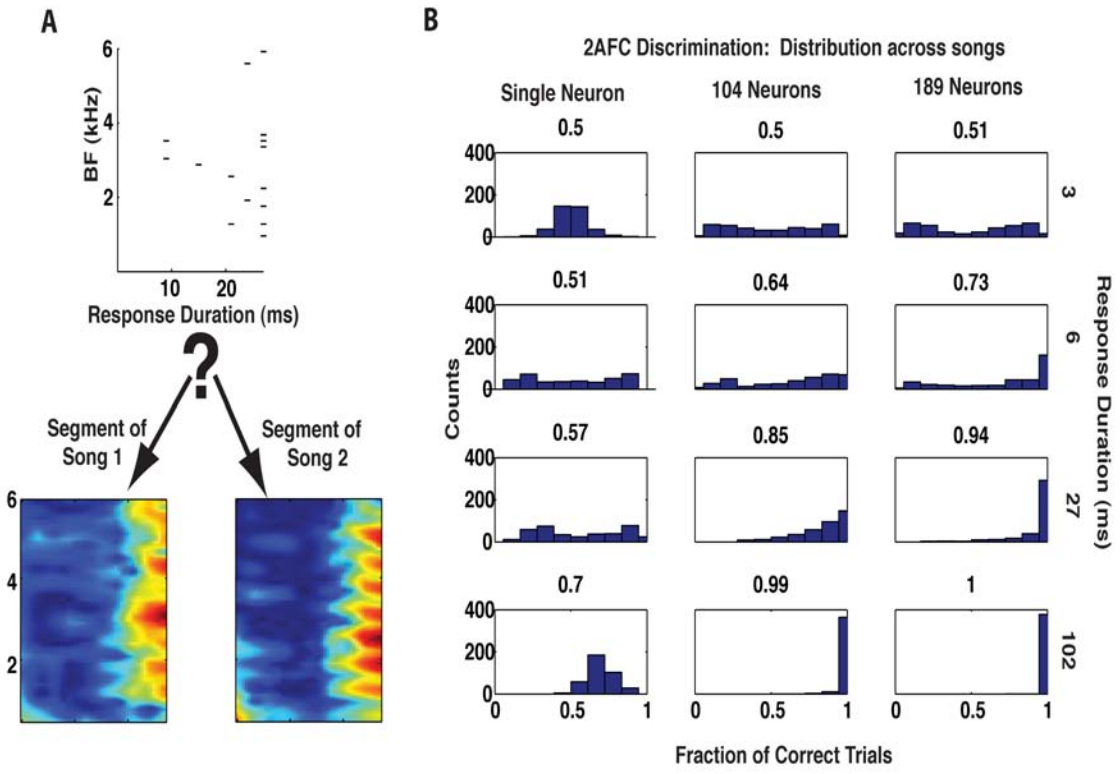

Figure 5: Conspecific song discrimination based on likelihood of spike-trains from multiple neurons. **A**, Spike trains from multiple neurons in response to presentation of song segment 1. Under a Two-alternative forced choice (2AFC) test, song discrimination is performed by choosing the song which leads to a greater likelihood of observing the given spikes. Spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches maximal value. Neurons with the same BF are plotted on the same row. **B**, 2AFC results as a function of response duration and the number of neurons used for discrimination. 2AFC was performed multiple times for each possible pairing of the twenty songs in the data set. Each panel shows the frequency of correct trials across all possible song pairings. Above each panel is reported the average of the histogram. On average, neurons performed at chance level when stimulus segments were only 3 ms in duration. Near perfect song discrimination can be achieved using 189 responses and response durations at least around 30 ms, or 104 neurons and durations of about 100 ms.

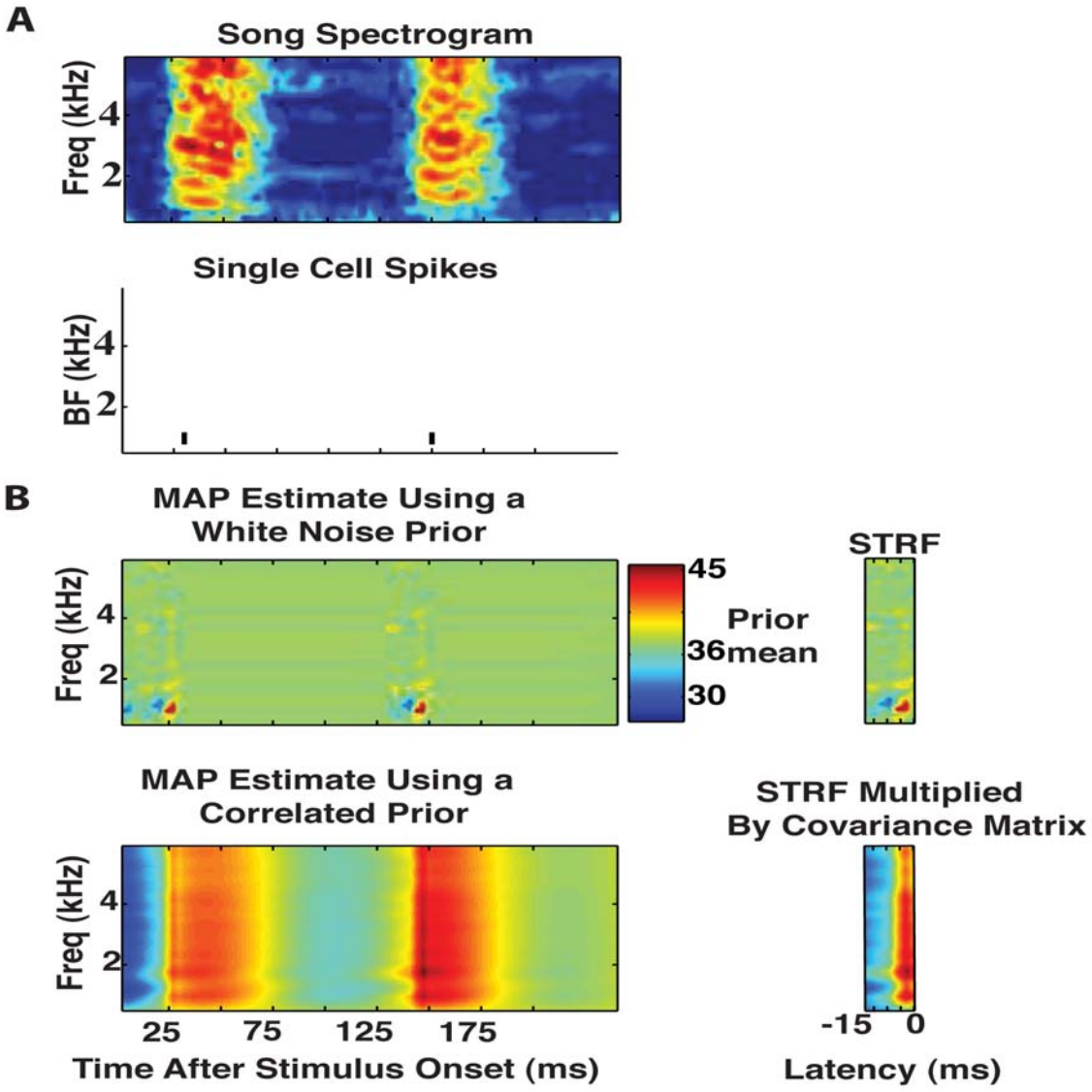

Figure 6: Single cell decoding of song spectrogram. **A**, (Top) Spectrogram of birdsong that elicited the two spikes shown immediately below. Spikes are plotted at the frequency at which this neuron's receptive field reaches maximal value. **B**, (Top, left) The most probable spectrogram from the posterior distribution (MAP Estimate) given the two spikes shown in A and using an uncorrelated prior. When a single spike occurs the MAP is determined by the neuron's spectrotemporal receptive field (STRF, shown to the right). In the absence of spikes, the MAP is determined by the prior mean. (Bottom, left) MAP estimate using the correlated Gaussian prior; when a spike occurs the MAP is determined by the neuron's STRF multiplied by the prior covariance matrix (shown to the right). Immediately after a spike the MAP infers spectrogram values using prior knowledge of stimulus correlations.

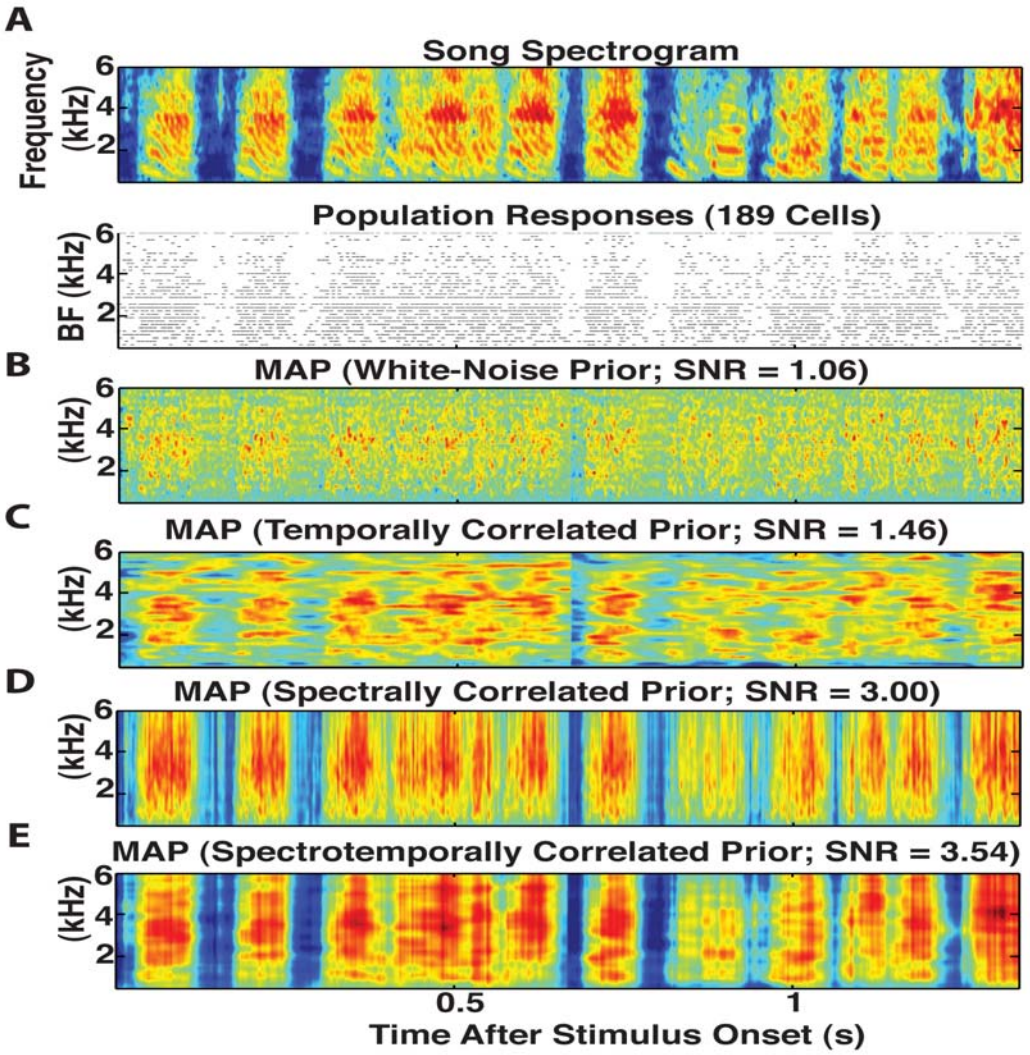

Figure 7: Population decoding of song spectrogram with varying degrees of prior information of song statistics. **A**, (Top) Spectrogram of birdsong played to 189 different neurons leading to the spike responses shown immediately below. Spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches its maximal value. Neurons with the same BF are plotted on the same row. MAP estimate given the responses in A, using an uncorrelated prior **B**, a prior with temporal correlations and no spectral correlations **C**, a prior with spectral correlations and no temporal correlations **D**, and a prior with spectral and temporal correlations **E**. Combining the spike train with spectral information is more important for reconstructing the original spectrogram than combining the spike train with temporal information. However, combining spikes with joint spectrotemporal information leads to the best reconstructions.

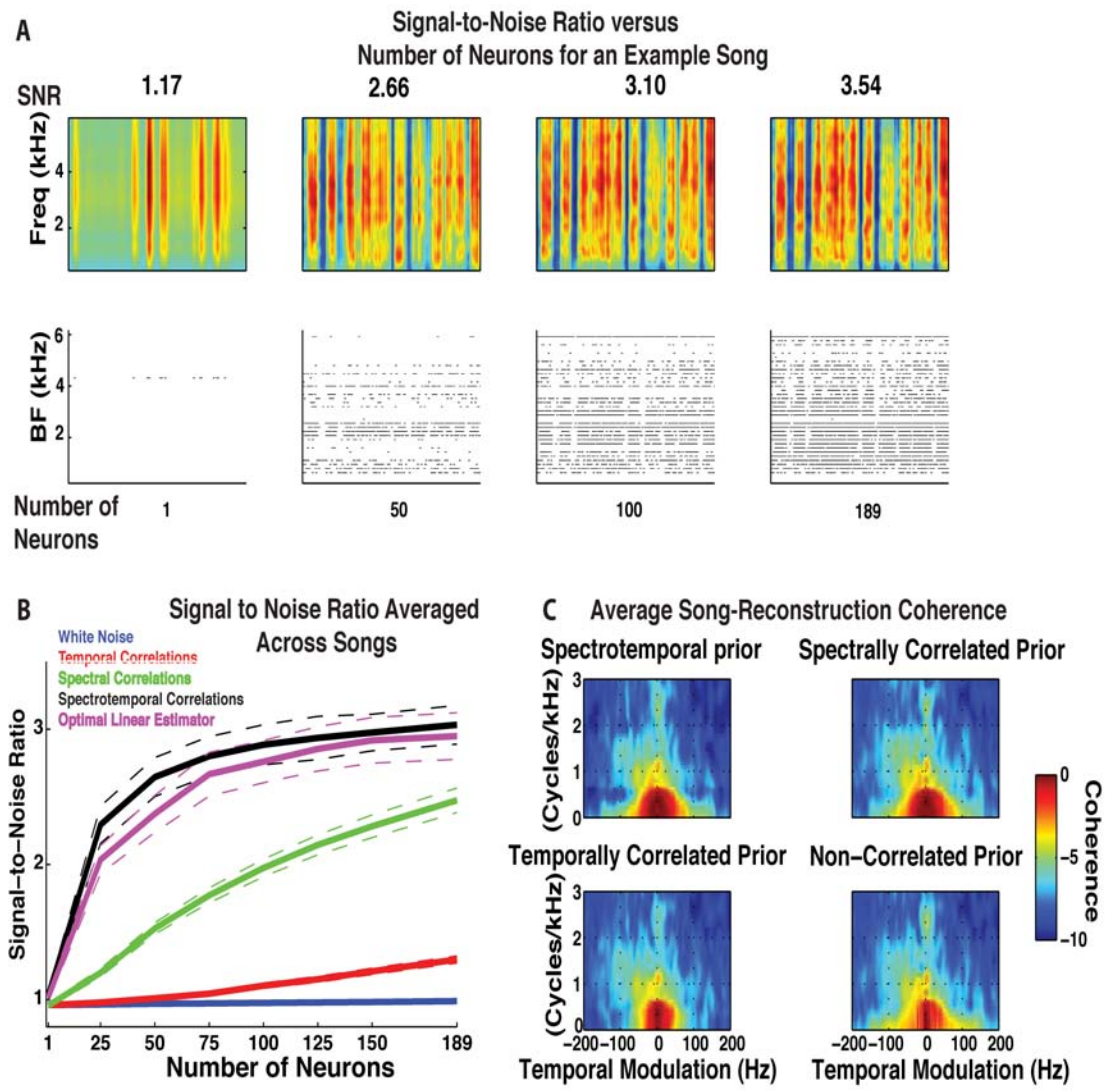

Figure 8: Decoding performance given different amounts of prior information and numbers of neurons. **A**,(upper row) Spectrogram reconstructions for an example song (Figure 7A) using a Gaussian prior with spectrotemporal correlations and using varying numbers of neuronal responses (plotted in the lower row). Above each reconstruction is the signal-to-noise (SNR) used to measure similarity between the reconstructed song and the song presented to the bird. **B**, Solid lines show the signal-to-noise (SNR) ratio averaged across all decoded songs while dashed lines show one standard error. The prior used for decoding is denoted by color. Spectral prior information leads to faster growth in the SNR than temporal information. For reference, the magenta line shows the growth in SNR for the commonly used optimal linear estimator (OLE). The OLE has access to both spectral and temporal correlations. **C**, Coherence between spectrograms and reconstructions under the four different priors. The horizontal axis reports temporal modulations and the vertical axis reports spectral modulations. All plots display the highest coherence at low spectral and temporal modulations. The primary effect of adding spectrotemporal prior information is to improve reconstructions at lower spectral and temporal modulations.

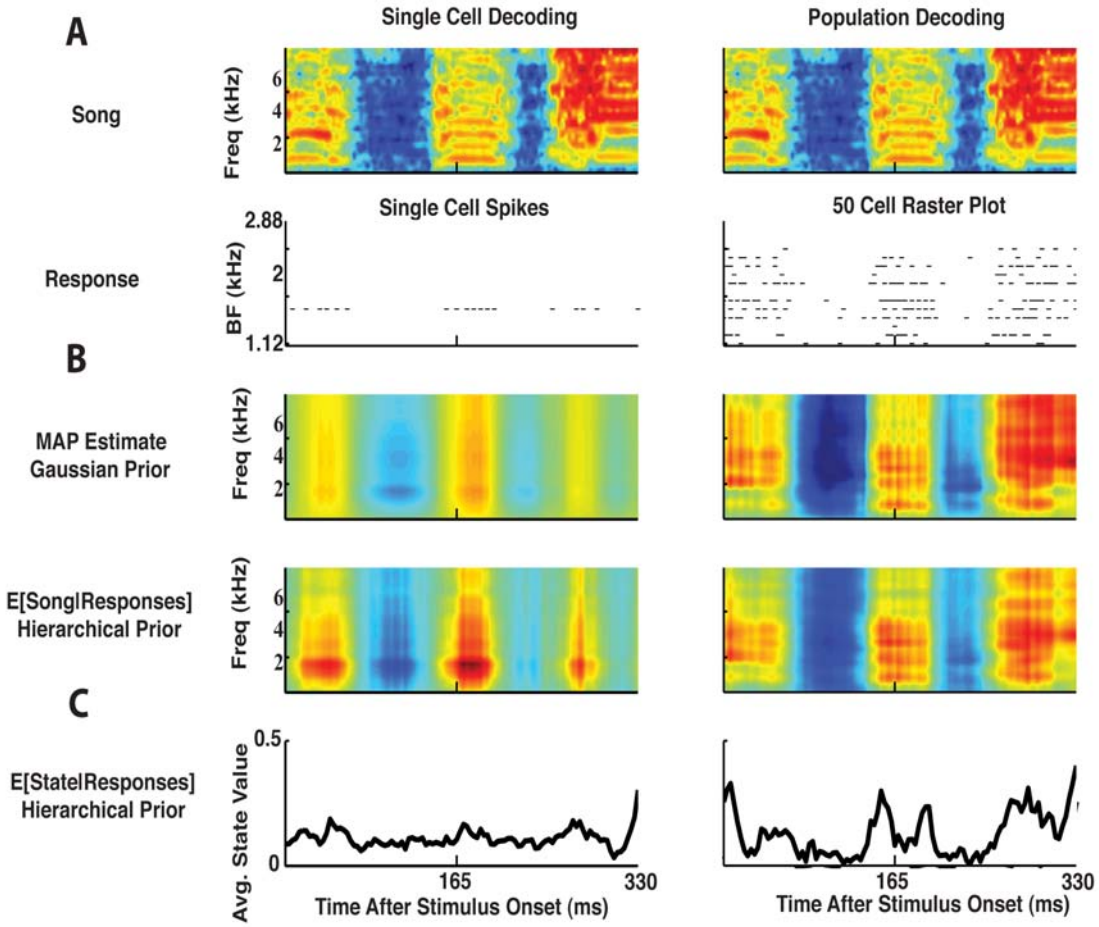

Figure 9: Single neuron and population decoding using a hierarchical prior. **A**, Song spectrogram along with a single cell's response to this song (left column) and the response of this cell plus forty nine other cells with nearby characteristic frequencies (right column). **B** MAP estimates using a single, correlated Gaussian prior (top row) are compared with estimates using the posterior mean and the hierarchical prior (bottom row); in both the single neuron and population decoding case, the estimate using a hierarchical prior looks similar to the MAP with a Gaussian prior. **C**, The expected value for vocalization state given responses; single cell responses do not yield enough information to accurately infer the spectrogram's vocalization state, however as the number of neurons used for inference increases the vocalization state becomes more pronounced.

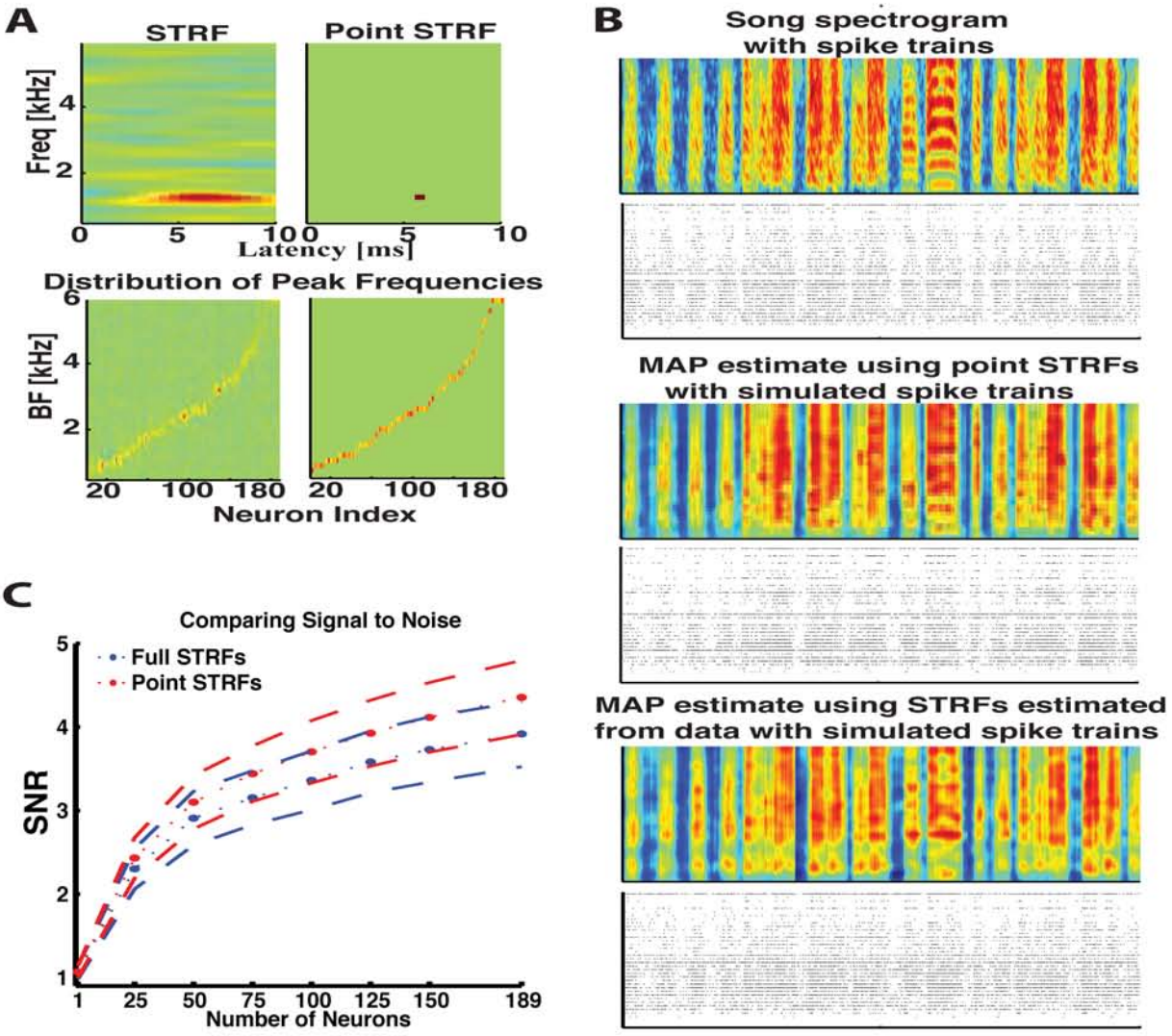

Figure 10: Spectral blur of STRFs causes a small loss of information for reconstructions. **A**, (upper, left panel) Example STRF and localized point STRF (upper, right panel) with equivalent peak frequency. (lower, left panel) Frequency vectors at latency where STRF obtains maximal value for the population of neurons used in this study. The equivalent plot for point STRFs (lower, right panel). Point STRF peak locations were randomly drawn from a distribution constructed using the peak locations of real STRFs. **B**, (first two rows) Song spectrogram and evoked responses of 189 real neurons. (middle rows) Reconstructed song spectrogram given simulated responses using a point STRF model. Simulated responses are shown immediately below the reconstruction. (bottom two rows) Reconstructed song spectrogram given simulated responses using full STRFs. Responses are shown immediately below the reconstruction. Reconstructions with full STRFs show slightly different spectral details but otherwise look *very* similar to reconstructions using point STRFs. **C**, SNR growth (plus and minus one standard error) as a function of the number of neurons used in decoding for point STRFs and full STRFs; on average the point STRFs have higher SNR than full STRFs.