Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds

Frédéric E. Theunissen,[1] Kamal Sen,[4] and Allison J. Doupe[2,3,4]

[1]Department of Psychology, University of California, Berkeley, California 94720-1650 and Departments of [2]Psychiatry and [3]Physiology, and [4]Sloan Center for Theoretical Neuroscience, University of California, San Francisco, California 94143-0444

The stimulus–response function of many visual and auditory neurons has been described by a spatial-temporal receptive field (STRF), a linear model that for mathematical reasons has until recently been estimated with the reverse correlation method, using simple stimulus ensembles such as white noise. Such stimuli, however, often do not effectively activate high-level sensory neurons, which may be optimized to analyze natural sounds and images. We show that it is possible to overcome the simple-stimulus limitation and then use this approach to calculate the STRFs of avian auditory forebrain neurons from an ensemble of birdsongs. We find that in many cases the STRFs derived using natural sounds are strikingly different from the STRFs that we obtained using an ensemble of random tone pips. When we compare these two models by assessing their predictions of neural response to the actual data, we find that the STRFs obtained from natural sounds are superior. Our results show that the STRF model is an incomplete description of response properties of nonlinear auditory neurons, but that linear receptive fields are still useful models for understanding higher level sensory processing, as long as the STRFs are estimated from the responses to relevant complex stimuli.

Key words: Natural sounds; auditory cortex; spectro-temporal; receptive field; field L; reverse correlation

Neuroscientists have successfully used the concept of a receptive field (RF) (Hartline, 1940) to characterize the stimulus features that are being encoded by sensory neurons both at the periphery and at higher levels of processing. The RF summarizes the encoding characteristics of a particular sensory neuron by showing the feature that will elicit the maximal response. This description has resulted in an understanding of the hierarchical computation underlying the feature extraction that occurs for example in the visual system, where the simple center-surround RFs of visual ganglion cells (Kuffler, 1953) become RFs for edge and bar detection in complex cells of V1 (Hubel and Wiesel, 1962). A similar representation is used in the auditory system, in frequency tuning curves, where the spatial dimensions have been replaced by a spectral dimension. More recently, visual and auditory neurophysiologists have also added the dimension of time to static RFs, obtaining spatial-temporal RFs in the visual system (DeAngelis et al., 1995; Cai et al., 1997; Ringach et al., 1997a; De Valois and Cottaris, 1998), and spectral-temporal RFs in the auditory system (Aertsen and Johannesma, 1981a,b; Aertsen et al., 1981; Clopton and Backoff, 1991; Eggermont et al., 1983a,b; Kim and Young, 1994; Kowalski et al., 1996a; Nelken et al., 1997; deCharms et al., 1998) (in both visual and auditory systems referred to as STRFs). The STRF shows which temporal succession of acoustical or visual features would elicit the maximal neural response. Recent research in the auditory cortex has suggested that auditory and visual cortical STRFs have remarkably similar time varying shapes (deCharms et al., 1998).

The underlying assumption that allows one to reduce all aspects of encoding by a neuron to an STRF is that the response to a novel time-varying stimulus that was not used in the estimation of the STRF can be predicted from the stimulus–response data used in the STRF estimation by simple linear interpolation or extrapolation. The STRF of a neuron can therefore be rigorously defined as the best linear model that transforms any time-varying stimulus into a prediction of the firing rate of a neuron. It is also this linear model of neural encoding that allows one to easily obtain STRFs from experimental data. If the visual or auditory spatial-temporal dimensions are sampled uniformly and randomly, then one can estimate the STRF simply by averaging the stimuli before each spike. This procedure is called the reverse correlation method (Boer and Kuyper, 1968). However, the uniform and random sampling requirement implies that one needs to use a stimulus ensemble that is the equivalent of white noise in both the spatial/spectral and temporal dimensions. If such an ensemble is not used, then the spike-triggered average stimulus will, in general, not be equal to the STRF (Aertsen et al., 1981; Eggermont et al., 1983a; Ringach et al., 1997b).

Therefore, even though describing neural coding in terms of RF has been a successful approach, the underlying linear assumptions and the use of white noise to obtain the STRF raise important issues. First, white noise has been shown to be a poor stimulus in higher sensory areas, because it elicits very few spikes, making the calculation of the STRF difficult for practical reasons (as the calculation requires large amount of data) and the results of questionable validity for methodological reasons (as the neurons might not be driven into their full dynamical range). Second, because sensory neurons exhibit varying degrees of linearity, it is essential to estimate how much of the response of the neurons can

Received July 8, 1999; revised Dec. 17, 1999; accepted Dec. 17, 1999.

This work was supported by research grants from the Alfred P. Sloan Foundation (to F.E.T., K.S., and A.J.D.) and the National Institutes of Health (NS 34385 to A.J.D. and MH 11209 and MH 59189 to F.E.T.). We thank Steve Lisberger for critical reading of an earlier version of this manuscript.

Correspondence should be addressed to Frédéric E. Theunissen, University of California, Berkeley, Department of Psychology, 3210 Tolman, Berkeley, CA 94720-1650. E-mail: fet@socrates.berkeley.edu.

be explained with the linear model implicit in the STRF. For example, it is known that certain neurons in higher auditory areas in primates (Rauschecker et al., 1995), bats (Ohlemiller et al., 1996), and birds (Scheich et al., 1979; Margoliash, 1983, 1986) are not well driven by simple or white noise stimuli and seem to be specifically tuned to the sounds of the animal's vocalizations. These neurons show some degree of nonlinearity, because they do not respond well to components of the vocalization presented alone but respond strongly to the complete vocalization. Could we nonetheless explain some of the encoding properties of such complex auditory neurons with the linear STRF model for a subset of specialized sounds? And if so, how would we calculate STRFs from such stimulus ensembles that in general will show correlations in time and frequency? Finally, how does one verify that the linear model is in fact capturing a significant fraction of the encoding of such neurons?

In this work, we address these issues by demonstrating how STRFs can be calculated from ensembles of natural sounds and by using simple methods to quantify the accuracy of the linear assumption. We further examine the degree of linearity by comparing the similarity and predictive power of STRFs obtained from natural sounds to those obtained from an ensemble of pure tone pips. The auditory neurons we examine lie in the songbird forebrain auditory fields L1, L2a, L2b, and L3, which are homologous both in their anatomical location in the chain of acoustical processing and in their functional properties to primary and secondary auditory cortical areas in mammals (Muller and Leppelsack, 1985; Fortune and Margoliash, 1992; Vates et al., 1996). Understanding how neurons in these high-level auditory areas process the sounds of the conspecific's vocalizations, in particular conspecific song, is behaviorally relevant, because songbirds must be able to recognize song both to identify the singer and for young males to be able to copy a tutor song (Marler, 1970; Konishi, 1985). In addition, auditory processing in field L (Lewicki and Arthur, 1996) must ultimately contribute to the extreme selectivity of song-selective auditory neurons found further along in the auditory hierarchy (in the song system; Nottebohm et al., 1976). Such neurons respond highly nonlinearly only to the specific spectral and temporal context of the bird's own song (Margoliash, 1983, 1986; Margoliash and Fortune, 1992; Lewicki, 1996; Theunissen and Doupe, 1998).

MATERIALS AND METHODS

Electrophysiology. All physiological recordings were done in urethane-anesthetized adult male zebra finches in acute experiments. Extracellular waveforms were obtained using tungsten electrodes that were inserted into the neostriatum of the bird at locations that were previously marked with stereotaxic measurements. The extracellular waveforms were transformed into spike trains by windowing the largest action potential. Single units or small multiunit clusters were recorded in this manner. At the end of the experiment, the bird was killed, and the location of the recordings was verified histologically in Nissl-stained brain sections. A complete description of similar experimental procedures can be found in Theunissen and Doupe (1998). The location of the sites was classified into anatomical subregions of Field L as described in Fortune and Margoliash (1992). The data presented here were obtained from five birds and 20 recording sites (five from area L1, two from L2a, five from L2b, and eight from L3).

Stimuli. Two stimulus ensembles were used: a random tone pip ensemble and a conspecific song ensemble. The random tone ensemble was made of 20 different 2 sec trains of tone pips. The frequency, amplitude, length, and intertone interval were random. The distribution of frequencies was made to match the average power spectrum of zebra finch song. The tone length and the intertone interval had Gaussian distributions with the same mean and SD as the syllable length and intersyllable interval of a large ensemble of zebra finch songs recorded in our laboratory. The tone pip loudness was varied randomly in logarithmic steps (0.125, 0.25, 0.5, 1.0). The loudest tones had an intensity that was similar to the peak intensity of the songs (80 dB SPL). The tones had onset and offset half-cosine ramps of 25 msec. The conspecific song ensemble consisted of 21 songs from different adult male zebra finches obtained in our colony including the bird's own song. Ten spike train response trials were obtained for each of the 21 songs and each of the 20 2 sec train of tone pips. The intertrial interval was 6 sec, and the trials from different stimuli were interleaved in random order.

STRFs in the auditory domain. The STRF is defined as the optimal linear filter that transforms a representation of a time-varying stimulus into a prediction of the firing rate of the neuron as estimated by the poststimulus time histogram (PSTH) of the neuron. This optimal linear filter can be simply obtained from the spike-triggered average stimulus (STA) if (1) the stimulus space used in the experiment encompasses all stimuli that are capable of producing significant neural responses, (2) if the sampling in the stimulus space is done in a random and uniform manner, and (3) if the multiple spatial dimensions used to represent the stimulus are independent of each other. If these three conditions are satisfied, the STRF will be proportional to the STA. This method for estimating the STRF is called the reverse correlation (Boer and Kuyper, 1968). We explain below why these three conditions can cause significant difficulties for the estimation of STRFs for high-level auditory neurons and how one can circumvent some of these constraints.

The first significant difficulty is the choice of a spatial representation of sound. The definition of the spatial dimensions of the stimulus in the visual domain is unambiguous because the natural decomposition of an image into pixels yields a representation in which the spatial dimensions are clearly independent of each other. In the auditory domain, the spatial dimensions are usually taken to be the amplitude values of each of the narrow-band signals obtained from a decomposition of the sound into frequency bands (a spectrographic representation of sound). This general form for the auditory spatial dimensions makes sense physiologically, because sounds are decomposed into narrow-band signals by frequency channels along the length of the cochlea. However, as we will elaborate below, such spatial dimensions will not be independent of each other when an invertible spectrographic representation of sound is used. For this reason (violation of condition 3), even when white noise sound is used, which would satisfy both the complete (condition 1) and random (condition 2) sampling of the auditory space, the spike-triggered average spectrogram will only be an approximation of the STRF. In addition, in both visual and auditory modalities, one would like to be able to investigate neural properties with stimuli other than white noise and therefore not be constrained by the random and uniform sampling condition (condition 2).

STRFs in the auditory domain have been defined in two manners that can be related to each other. Realizing that an STRF based on the spike-triggered average spectrogram would depend not only on the statistical properties of the stimulus (condition 2) but also on the nature of the spectrographic representation (condition 3), Aertsen and Johannesma (1981a) originally defined an invariant STRF as the second-order Volterra kernel between the sound pressure waveform and the PSTH. This particular definition of the STRF is therefore independent of any choice of spectrographic representation and could theoretically be estimated from stimulus ensembles with different statistical properties. The second-order kernel that relates sound pressure waveform to the PSTH can also be rewritten as a first-order Volterra kernel that relates the Wigner function of the sound-pressure waveform to the PSTH (Eggermont and Smith, 1990; Eggermont, 1993; Kim and Young, 1994; Nelken et al., 1997). The Wigner function yields a scale-free time-frequency representation of the sound (see also Appendix). The second definition of the auditory STRF is the more common one: it is the first-order Volterra kernel between a spectrographic representation of sound and the PSTH (Eggermont et al., 1983a; Escabi et al., 1998; Shamma et al., 1998). As shown in the Appendix, the second definition of the STRF (called in this section the spectrographic STRF) can be thought of as a filtered version of the invariant STRF (see also Klein et al., 2000). This definition of the STRF is favored because it illustrates the spectral-temporal patterns that elicited neural responses in an explicit manner. However, depending on the choice of the spectrographic representation, one can obtain many spectrographic STRFs for a given neuron. Moreover, a spectrographic STRF runs the risk of losing information as a result of the filtering operation. To be able to relate the spectrographic STRF to the invariant STRF, without any a priori assumptions, the spectrographic representation of sound must be invertible (see Appen-

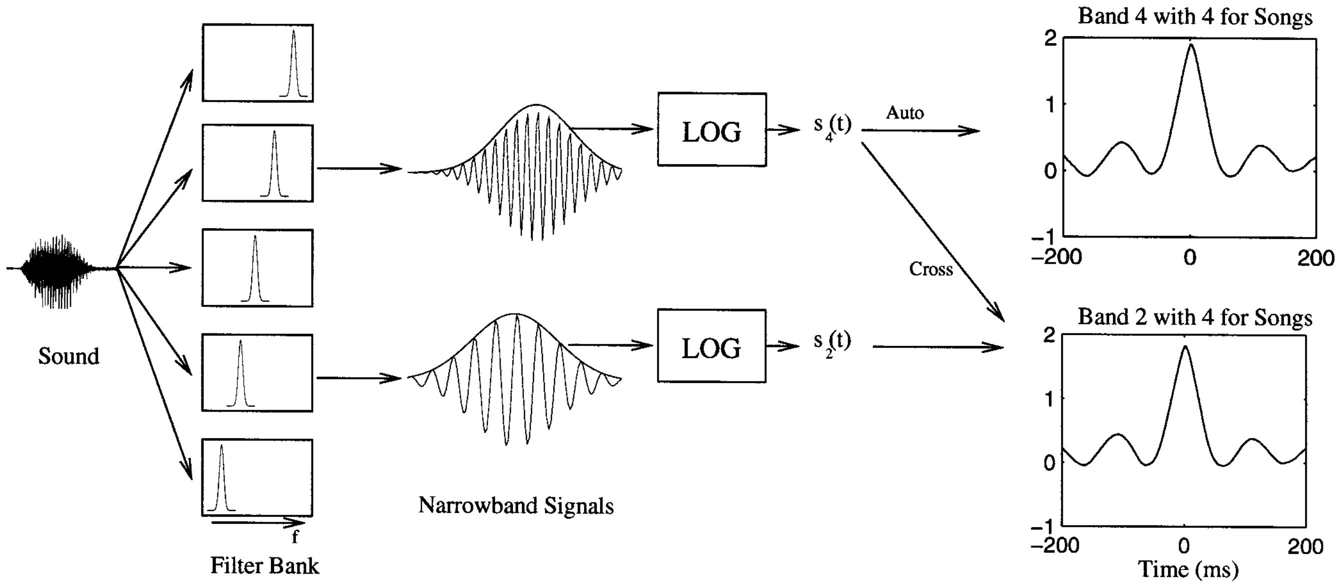

Figure 1. Schematic illustrating the spectrographic decomposition and the calculation of the stimulus autocorrelation matrix. The sound is decomposed into frequency bands by a bank of Gaussian filters. The result is a set of narrowband signals with time-varying amplitude and phase. Our representation of sound is based on the time-varying amplitude envelopes. Although the time-varying phase is discarded, the relative phase across frequency bands is preserved because of the large overlap between adjoining filters. The time-varying amplitude envelope or its log is what is usually represented in a spectrogram. Our representation of sound and the spectrograms shown in this paper are based on the log of the amplitude envelopes. The stimulus autocorrelation function is then found by cross-correlating the log-amplitude envelope of a particular band with the log-amplitude envelope of all the other bands, including itself. The autocorrelation for the entire ensemble is done by averaging the correlation at each time point for all stimuli in a given ensemble. Here we show the results of two of such pairwise correlations: the correlation of band 4 (centered at 1250 Hz) with band 2 (centered at 500 Hz) and of band 4 with itself, for the song ensemble.

dix). That is, it should be possible to recover the original sound from the spectrogram, except for a single absolute phase. Noninvertible representations could be used if a priori knowledge of the neural response properties tells us that any two different stimuli that are not separately recoverable from the noninvertible representation of sound yield the same neural responses. In such cases the invariant STRF can still be recovered from the spectrographic STRF.

An invertible spectrographic representation of sound requires the use of overlapping frequency bands, as explained in this paragraph and in more mathematical detail in the Appendix. The fine temporal structure of a sound is given by the relative phase of the narrow-band signals obtained in the decomposition of the sound into frequency bands. The phase of these narrow-band signals is thrown away in a spectrographic representation, where only the amplitude envelopes of the narrow-band signals are preserved. However, the relative phase of the narrow-band signals can be recovered from the joint consideration of the amplitude envelopes, as long as there is sufficient overlap among the frequency bands (Cohen, 1995; Theunissen and Doupe, 1998). It is for this reason that, in a complete spectrographic representation, where frequency filters overlap, the spike-triggered average of even a white noise stimulus will only be an approximation of the spectrographic STRF. In a spectrographic representation with nonoverlapping frequency bands, the spike-triggered average spectrogram would be equal to the spectrographic STRF but, in general, because such a spectrographic representation is noninvertible, one might not be able to obtain the invariant STRF.

An invertible spectrographic representation of sound. In this study, we used the spectrographic definition of the STRF (from here on called the STRF) and an invertible spectrographic representation that is illustrated schematically in Figure 1. The sounds are represented by a set of functions of time $s_{\{i\}}(t)$, where $s_i(t)$ is taken to be the log of the amplitude envelope of the signal in the frequency band i (the brackets indicate that we refer to the entire set of time-varying functions). The frequency bands were obtained with Gaussian filters of 250 Hz width (SD). We used 31 frequency bands spanning center frequencies between 250 and 8000 Hz. In this manner, the center frequencies of neighboring bands are separated by exactly 250 Hz or the equivalent of 1 SD. It is this large amount of overlap that allows this representation to be invertible. We extracted the amplitude envelope of each frequency band using the analytical signal as explained in Theunissen and Doupe (1998) and characterized the sound by the difference between the log value of the amplitude and the mean log amplitude for that band. We have shown that the particular time-frequency scale used in our spectrographic representation of sounds is the most efficient at representing the spectral and temporal structure of songs that is essential in eliciting the response of song-selective neurons in the zebra finch (Theunissen and Doupe, 1998). The additional log transformation was used because it improved our results significantly, as we will discuss in Results. Note that, in the auditory domain, the STRF is a linear transformation between a nonlinear representation of the stimulus and the PSTH. The calculation of the amplitude envelopes that make up the spectrogram is a nonlinear operation (although in our case it is invertible), and additional nonlinearities such as the log transform are often used. This model, therefore, includes known or determined static nonlinearities between the stimulus and the response.

Calculation of the STRF for any stimulus ensemble. In this section we elaborate on the details of the analytical solution of the STRF for any sound ensemble, which involves correcting for the correlations in the stimulus ensemble. The STRF is defined as the multidimensional linear Volterra filter $h_{\{i\}}(t)$, such that:

$$r_{\mathrm{pre}}(t) = \sum_{i=1}^{nf} \int h_i(\tau) s_i(t - \tau) d\tau,$$

where $r_{pre}(t)$ is the predicted firing rate, $s_{\{i\}}(t)$ is the multidimensional representation of the time-varying stimulus, and nf is the total number of spatial dimensions or frequency bands in our case. $h_{\{i\}}(t)$ is found by requiring that $r_{\mathrm{pre}}(t)$ be as close as possible to $r_{est}(t)$, the estimated firing rate obtained from a PSTH, in the mean square error sense. One finds that in the frequency domain, the set of $h_{\{i\}}$ can be obtained by solving a set of linear equations for each frequency w. This set of equations is written in vector notation as:

$$\mathbf{A}_{\mathrm{w}} \cdot \vec{H}_w = \vec{C}_w.$$

$\mathbf{A}_w$ is the autocorrelation matrix of the stimulus in the frequency domain:

$$\mathbf{A}_w = \begin{pmatrix} \langle S_1^*(w)S_1(w)\rangle & \langle S_1^*(w)S_2(w)\rangle & \cdots & \langle S_1^*(w)S_{nf}(w)\rangle \\ \langle S_2^*(w)S_1(w)\rangle & \langle S_2^*(w)S_2(w)\rangle & & \\ \vdots & & & \\ \langle S_{nf}^*(w)S_1(w)\rangle & & & \langle S_{nf}^*(w)S_{nf}(w)\rangle \end{pmatrix}, \text{ where}$$

$S_i(w)$ is the Fourier transform of $s_i(t)$, * denotes the complex conjugate, and $\langle\ \rangle$ indicates that we estimated the cross moments by averaging over samples. $\vec{H}_w$ is the Fourier transform of the set of $h_i(t)$:

$$\vec{H}_w = \begin{pmatrix} H_1(w) \\ H_2(w) \\ \vdots \\ H_{nf}(w) \end{pmatrix}.$$

$\vec{C}_w$ is the cross-correlation between the spike trains and the stimulus amplitude envelopes in each band:

$$\vec{C}_w = \begin{pmatrix} \langle S_1^*(w)R(w)\rangle \\ \langle S_2^*(w)R(w)\rangle \\ \vdots \\ \langle S_{nf}^*(w)R(w)\rangle \end{pmatrix},$$

where $R(w)$ is the Fourier transform of $r_{est}(t)$. $\vec{C}_w$ is the Fourier transform of the spike-triggered average spectrogram. To solve for $h_{\{i\}}$, the matrix $\mathbf{A}_w$ is inverted for each frequency:

$$\vec{H}_w = \mathbf{A}_w^{-1} \cdot \vec{C}_w.$$

The inverse Fourier transform is then applied to $\vec{H}_w$ to obtain the STRF in the time domain, $h_{\{i\}}(t)$. As long as an invertible representation of sound is used, and the stimulus autocorrelation matrix, $\mathbf{A}_w$, is also invertible, the spectrographic STRF found with this procedure can be directly related to the invariant STRF. As we will discuss below and show in Results, the autocorrelation matrices of the stimulus ensembles used in this work were not invertible. This will happen when stimulus ensembles do not sample all possible sounds. In such cases, the number of independent dimensions needed to represent the stimulus ensemble is less than the total number, *nf*, used in the initial representation of the stimuli. However, an estimated STRF can still be obtained by performing a singular value decomposition (SVD) of the autocorrelation matrix. The SVD method is used to determine numerically the smallest number of independent spatial dimensions that are needed to represent the stimulus ensemble (Press et al., 1992). The division by the autocorrelation matrix is then performed only for the subset of sounds spanned by these independent dimensions. The estimated STRF obtained in this way can therefore only be equated to the invariant STRF if the neural responses to the regions of sound that are not sampled are insignificant.

Parameters used in our numerical calculations. The number of frequency values, w, used to represent the stimulus autocorrelation matrix and the stimulus–response cross-correlation in the frequency domain depends on the number of time points used in the time window in the estimation of the crossmoments. We used a sampling rate of 1 kHz (the amplitude envelopes of our frequency bands are effectively band limited to frequencies <500 Hz; Theunissen and Doupe, 1998) and looked at the cross moments in a 400 msec time window, yielding 200 frequency values, w. Therefore, solving for $h_{\{i\}}$ required inverting 200 31 × 31 complex matrices. [Note that there are two "frequencies": the first one corresponds to the "spatial" dimension in our spectrographic representation of sound: the 31 frequency bands (ranging from 250 to 8000 Hz). The second is the 200 frequency values w (ranging from 0 to 500 Hz) that characterize the temporal modulations in and across each of the 31 bands in our representation of sound.]

Figure 1 shows how the stimulus autocorrelation is calculated in the time domain. For each frequency band, the autocorrelation of the log of the amplitude envelope is obtained: these autocorrelations are the diagonal terms in the stimulus autocorrelation matrix. The off-diagonal terms are obtained by cross-correlating the amplitude envelope in one frequency band with the envelope in another band. In this calculation all the stimuli in one ensemble are used, and the results are averaged. We therefore estimated our stimulus autocorrelation matrix from ~40 sec of sound (2 sec per stimulus times 20 stimuli). Figure 5 shows the entire autocorrelation matrix for three stimulus ensembles: white noise, tone pips, and zebra finch song. The particular correlation properties of these stimulus ensembles will be explained in Results.

The estimation of the STRFs by this generalized cross-correlation method required two additional numerical techniques. First, because of limited data, the spike-triggered average is inherently noisy. The noise corresponds to the expected deviations from the mean that are obtained in any averaging procedure. Although this noise was relatively small in our case, it is amplified in the calculation of the STRF in the division by the autocorrelation matrix of the stimulus. This is particularly true at high-amplitude modulation frequencies in which the autocorrelation of the stimulus is very small. Therefore, to prevent the amplification of high-frequency noise, the cross-correlation between stimulus and spike trains was lowpass-filtered. The frequency cut-off was found systematically by investigating which amplitude modulation frequencies in the spike-triggered average had significant power, using a jackknife resampling technique in the frequency domain (Efron, 1982). In brief, we calculated the jackknife estimate of the SE of the real and imaginary part of $\vec{C}_w$. For each element in the vector, we then estimated at what cut-off frequency, w, the magnitude of each of the $\langle S_w^* R_w\rangle$ elements was less than two times the SE. This cut-off frequency was used to lowpass filter $\vec{C}_w$.

The second numerical consideration deals with the inversion of the stimulus autocorrelation matrix. Because each of the amplitude envelopes of our stimuli was band limited to low frequencies and because the temporal correlations were similar in all frequency bands, the autocorrelation matrices are singular. In other words, for both our song ensemble and our random tone ensemble, the 31 × 400 dimensions used to represent their second-order structure are redundant. As mentioned above, in such situations one can find the independent dimensions by using the SVD method. The number of independent dimensions or eigenvectors of the autocorrelation matrix was estimated by ignoring all the eigenvectors that had eigenvalues that were smaller than the maximum eigenvalue of all the frequency matrices for a given stimulus ensemble, multiplied by a tolerance factor. The tolerance factor was found by testing a range of values and choosing the one that gave the best predictive power to the STRF model. The predictive power was estimated by calculating the coherence between the actual response and the predicted response to novel stimuli from the same ensemble, as explained below. The coherence is a function of the frequency and is given by Marmeralis and Marmeralis (1978):

$$\gamma^2(w) = \frac{\langle R(w)R_{pre}(w)^*\rangle\langle R(w)^*R_{pre}(w)\rangle}{\langle R(w)R(w)^*\rangle\langle R_{pre}(w)R_{pre}(w)^*\rangle}.$$

The best R_{pre} is the one that gives the largest coherence averaged over all frequencies. Although the number of significant eigenvectors and therefore the tolerance value is a property of the stimulus ensemble, we found that the best tolerance value changed slightly for each neuronal site. We attribute this effect to variances in the noise in the cross-correlation between spike train and stimulus ($\vec{C}_w$). By slightly varying the number of eigenvectors used in the estimation of the STRF, this noise can be selectively filtered out (because the eigenvectors with the smallest power represent the higher amplitude modulation frequencies, which are the most corrupted by noise).

Goodness of fit. To be able to compare our results with previous published work, we also calculated the goodness of the STRF model by calculating the cross-correlation coefficient between the predicted firing rate and actual firing rates:

$$CC = \frac{\langle (r_{pre}(t) - \overline{r_{pre}(t)})(r_{est}(t) - \overline{r_{est}(t)})\rangle}{\sqrt{\langle (r_{pre}(t) - \overline{r_{pre}(t)})^2\rangle\langle (r_{est}(t) - \overline{r_{est}(t)})^2\rangle}}.$$

The predicted firing rate was obtained by convolving the STRF with the stimulus, rectifying, and scaling the result to minimize the square error between the predicted and estimated firing rates.

For each zebra finch song or 2 sec train of tone pips, we calculated a STRF that was based on all the other stimuli in the same ensemble but that specifically excluded the stimulus–response data being tested. The CC calculated from such a stimulus ensemble-matched STRF was used to estimate how well the STRF model obtained from a particular stimulus ensemble fitted the stimulus–response function of any stimulus that was not part of the ensemble but that had identical second-order statistical properties. The CCs obtained from these calculations are referred to as CC-matched.

For each song and train of tone pip stimuli, we also calculated the predicted firing rate obtained using the STRF calculated from all the stimuli in the other stimulus ensemble. For each neuronal site, the STRF obtained from all the songs was used to predict the response to tone pips, and vice versa. The CCs obtained from these calculations are referred to as CC-switched.

The estimated firing rate was obtained by smoothing the PSTH with a hanning window. The window that gave the maximal cross-correlation value for the matched ensemble was used. The width of the window was similar to the width of the STRF and had values between 6 and 96 msec (width at half maximum). This estimation of the cross-correlation coefficient has a negative bias because of the noise in the PSTH from the limited number of samples ($n = 10$ trials for each stimulus per neuronal site). We generated a bias-corrected cross-correlation and calculated SEs by using the jackknife resampling method after using the z-transform on our pseudosamples of correlation coefficients (Miller, 1974; Hinkley, 1978). The bias-corrected values were slightly higher, as expected (mean biased, 0.42; mean unbiased, 0.48). Note that this cross-correlation comparison can still miss information in the spike trains that would be lost by the averaging procedure and therefore still represents a lower bound of the goodness of the fit.

RESULTS

To begin to understand the hierarchy of acoustical processing in the avian auditory forebrain and in particular to determine the stimulus–response function of complex field L auditory neurons when stimulated with conspecific songs, we compared neural responses obtained using a random tone pip ensemble to those obtained with a large ensemble of conspecific songs. The conspecific song ensemble was chosen because we are interested in understanding the processing of complex natural sounds. The random tone pip ensemble was designed to have the same overall power density as the song ensemble and similar temporal structure. Because the tone pips were independent of each other, any zebra finch song could be generated by a linear combination of a particular subset of such tone pips. By designing the tone ensemble in this manner, we were able to compare the stimulus–response function of neurons obtained from two quite different stimulus ensembles that nonetheless sampled a similar spectral-temporal region of auditory space.

Average properties of the stimulus ensembles and mean responses

The random tone pip ensemble was designed so that the succession of tone pips would have a similar temporal structure to the succession of syllables in a zebra finch song. For the ensemble of 20 songs used in this study, the mean syllable length was 95 msec with a SD of 37 msec. The mean intersyllable interval was 37 msec with a 21 msec SD. We used these values and assumed a Gaussian distribution to generate tone pips with similar temporal structure as the songs. The frequency of the tone pips was also randomly chosen so that the song ensemble and the tone-pip ensemble would have the same average power spectrum. The average power density curve for the 20 songs used in this study is shown in Figure 2.

Figure 3 illustrates the responses of a neuronal site in L2b to the two types of stimuli. This neuronal site responds selectively to certain lower-frequency tone pips and seems to respond to most of the syllables of the spectrally complex zebra finch song shown. The mean firing rate of this neuron was much higher for the song stimuli than for the tone-pip stimuli. This pattern was true for all of our neuronal sites, as shown in Figure 4 ($p = 0.0001$). The maximal firing rate was more similar between ensembles than the average rate, although it was still significantly higher for the song ensemble ($p = 0.03$). Both ensembles elicited higher responses from most neurons than continuous white noise, although other broadband stimuli [such as the spatially modulated stacks or ripples that have been used by other investigators (Schreiner and Calhoun, 1994; Kowalski et al., 1996a)] elicited rates that were as high as the song stimuli (data not shown).

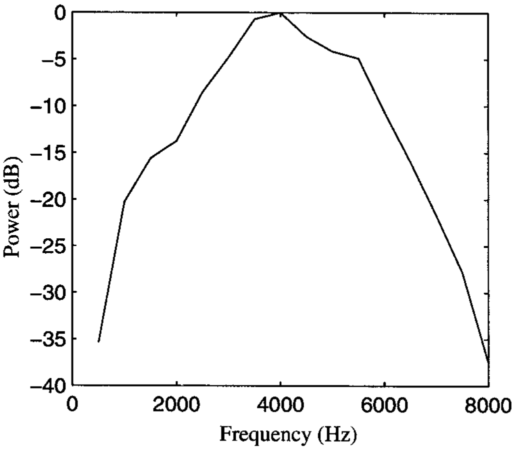

Figure 2. Power spectrum of the ensemble of 20 zebra finch songs used in this study. The curve shows the mean power density as a function of frequency. The same power density was used to generate the ensemble of tone pips.

Spectral-temporal properties of the stimulus ensembles

As exhibited in Figure 3, the neuronal response in these high auditory areas is reliably correlated with specific acoustical patterns in the song or tone pip ensemble. A useful first-order description of this neural encoding is the STRF, a linear model that relates the spectral-temporal acoustical patterns of the stimulus to the probability of spiking. To correlate such spectral-temporal patterns with spike activity, one needs to transform the one-dimensional sound pressure waveform into a two-dimensional time-frequency representation of sound, of which many are possible. As described in detail in Materials and Methods, Appendix, and Figure 1, we used an invertible spectrographic representation of sound, that is, one from which the original sound can be recovered.

If a white noise stimulus is used, the STRF can be estimated by the reverse-correlation method by averaging the stimulus before each spike: the spectral-temporal structure that is related to high probabilities of spiking will add, and the other spectral-temporal patterns will average to zero. This approach has two problems, however. First of all, and most significantly, for most neurons in higher level auditory areas white noise does not elicit many spikes. Secondly, and on a more technical point, white noise that is completely free of correlations cannot be achieved for an invertible spectral-temporal representation such as the one used here (see Materials and Methods).

To overcome these shortcomings, we used an extended version of the reverse-correlation method, in which the spike-triggered averaged stimulus is normalized by the correlations present in any particular stimulus ensemble. This normalization corrects for the fact that for all stimuli, but much more significantly for natural stimuli such as zebra finch songs, specific spectral and temporal patterns are correlated with each other (so-called second-order structure). To correct for these correlations, one has to effectively perform a weighted average of the stimulus around each spike, in a mathematical operation that involves a deconvolution in time and a decorrelation in frequency of the spike-triggered average stimulus, as explained in Materials and Methods.

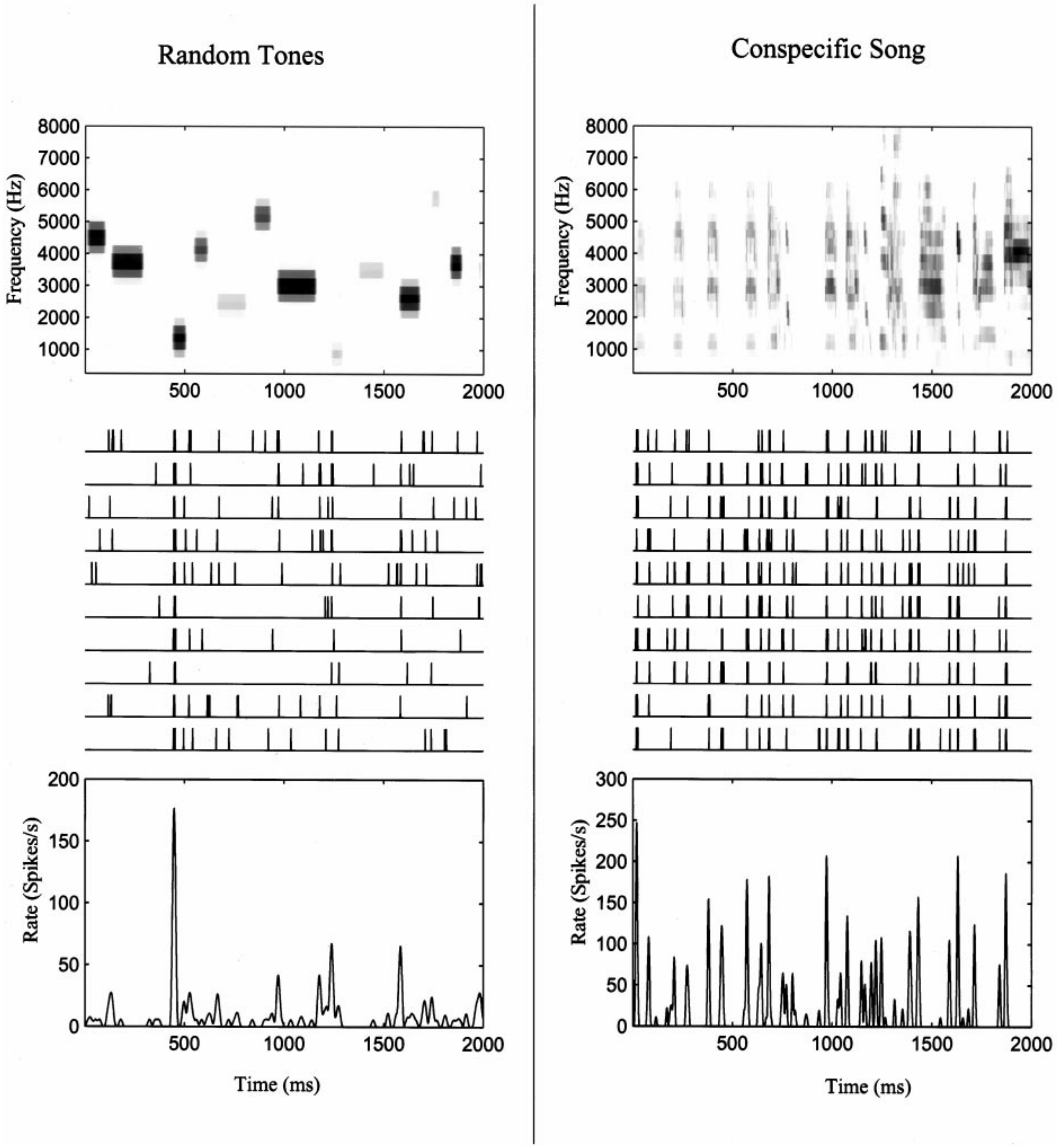

Figure 3. Samples of stimuli (*top panels*), neuronal responses (*middle panels*), and mean spike rate (*bottom panels*) for the two stimulus ensembles: a song-like sequence of random tone pips (*left column*) and a zebra finch song (*right column*). The *top panels* show the spectrographic representation of the stimuli using the frequency decomposition described in Materials and Methods. The *middle panels* show 10 spike train recordings obtained in response to 10 presentations of the stimuli for a neuronal site in area L2b of one of the birds used in the experiment. The *bottom panel* shows the estimated spike rate obtained by smoothing the PSTH with a 6 msec time window.

This normalization procedure depends on the specific second-order structure of the stimulus ensemble. The second-order structure is described by the autocorrelation function of the stimulus. Because our stimulus representation is based on the amplitude envelopes of the decomposition of the sound into 31 frequency bands (our spectrographic representation of sound as explained in Materials and Methods), the autocorrelation consists of a matrix of 31 × 31 functions that show the correlation of the amplitude envelopes in each band with the amplitude envelopes of all the other bands (see Materials and Methods and Fig. 1).

The entire autocorrelation matrix in the time domain is shown in the top panels of Figure 5 for our two stimulus ensembles and for white noise. To be able to simply equate the spike-triggered spectrogram to the STRF, the stimulus autocorrelation matrix in the time domain should consist of delta functions on the diagonal and null functions everywhere else. In such a case, the stimulus autocorrelation matrices in the frequency domain $\mathbf{A}_w$ (see Materials and Methods) are all equal to the unity matrix, and the STRF is therefore exactly equal to the spike-triggered average spectrogram. As shown in Figure 5, white noise approaches this ideal, but because our frequency bands overlap, the correlations spread out slightly to nondiagonal terms. It is because of this

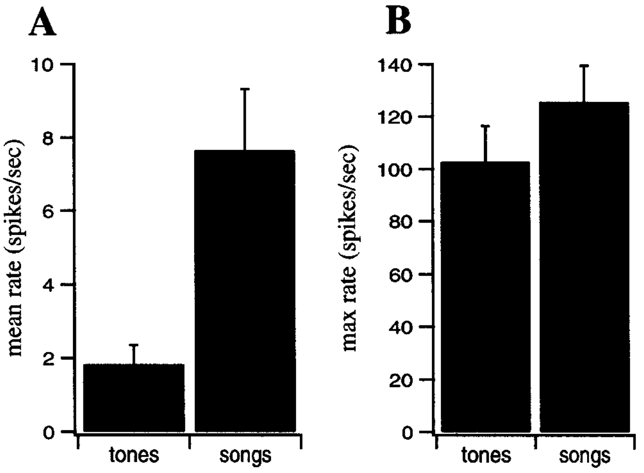

Figure 4. Mean (*A*) and maximal (*B*) firing rates of our Field L neurons relative to background, obtained using the tone ensemble and the song ensemble.

spread that, even when white noise is used, the STRF obtained by just taking the spike-triggered spectrogram is an approximation.

The second-order statistical structure of the two stimulus ensembles used in this paper are shown in the middle and right panels of Figure 5. There is more similarity between the random tone and white noise autocorrelation matrices than between the song and white noise autocorrelation matrices. As is the case for the white noise ensemble, the off-diagonal terms of the autocorrelation of the random tone ensemble quickly become small. The random tone autocorrelation is different, however, in its temporal correlations, as exhibited by a much wider central peak in the diagonal terms; in other words, the amplitude envelopes in each frequency band vary slowly in time. In the Fourier space, these temporal correlations are shown by the fact that power in amplitude modulations is concentrated at the lower frequencies, as shown in the middle bottom panel of Figure 5 and in Figure 6. These temporal correlations are similar to those found in the song ensemble. This is to be expected because we designed our tones to have the same average duration and gap-time as the syllables in our ensemble of zebra finch song. The random tone pip ensemble also differs from white noise because tones in each frequency band are presented in isolation. This leads to small negative correlation values in the off-diagonal terms. Finally, the song autocorrelation matrix shows both the temporal and spatial (across frequency bands) correlations that are expected from the acoustical properties of the zebra finch song. In zebra finch song, the temporal correlations are similar in all frequency bands and also synchronized across bands: the off diagonal terms have correlation functions similar to those of the diagonal terms. The spatial and temporal correlations can also be seen in the two-dimensional Fourier spectra of the stimulus ensembles, shown in the bottom panels of Figure 5. Both the tone-pip stimuli and the song ensemble have most of their power at low temporal frequencies and low spatial frequencies. We will come back to this point in the next section.

The normalization procedure leading to the true STRF requires the division of the spike-triggered average stimulus by the autocorrelation matrix of the stimulus. This division effectively increases the contribution of infrequent spectral-temporal patterns in the weighted spike-triggered average. One problem arises when particular spectral-temporal patterns are not sampled at all, because the normalization procedure would require a division by 0. This is the case, for example, with an ensemble of zebra finch songs: zebra finch songs span only a restricted region of the stimulus space of all possible sounds. To deal with this problem, we used an SVD of the correlation matrix (see Materials and Methods). This method allows one to assess the true dimensionality of the stimulus space by determining the regions of the space where the probability of occurrence of specific spectral-temporal patterns is significantly different from zero. The subspace that is not sampled is then simply ignored in the calculation. The level of significance was chosen by defining a noise tolerance threshold that is expressed as a fraction of the power density along the stimulus dimension where the probability of occurrence of spectral-temporal patterns was maximal (see Materials and Methods). The regions of space where the probability of occurrence was below this level of significance were then ignored. We chose the tolerance threshold that gave the STRF that best predicted neural responses to novel stimuli. We found the mean optimal tolerance factor to be 0.001 (range, 0.00005–0.005) for the song ensemble and 0.003 (range, 0.0001–0.005) for the tone ensemble.

The nature of the stimulus subspace being sampled significantly can be examined by looking at the space covered by the independent dimensions that are needed to describe that subspace. When the correlations have stationary statistics, this subspace is shown in the two-dimensional Fourier spectra of the stimulus, as illustrated in the bottom panels of Figure 5. In our case, the temporal correlations were stationary but we did not assume stationarity in our spatial correlations. Nonetheless, the bottom panels of Figure 5 illustrate approximately the spectral and temporal patterns that are sampled by our two stimulus ensembles, as well as white noise. In addition, we show in Figure 6 the number of dimensions needed to represent the stimulus subspace as a function of the temporal frequencies (describing the power of the temporal modulations in the amplitude envelopes in each of the frequency bands). The number of dimensions depends both on the power at a particular amplitude modulation frequency and on how correlated the amplitude modulations in one frequency band are with those in a different frequency band. At the tolerance values used here, the autocorrelation matrix of a white noise ensemble obtained with the same number of data points would have the maximum number of 31 dimensions for all frequencies. The bottom panels of Figure 5 and Figure 6 show that both of our stimulus ensembles sample mostly the low-frequency end of the temporal spectrum. In addition, in zebra finch song, the presence of energy in one frequency band is correlated with the presence of energy in many other frequency bands, yielding high spatial correlations. In the tone-pip ensemble, the presence of energy is one frequency band is positively correlated with energy in the neighboring bands and negatively correlated with energy in frequency bands further away. Because the spatial correlations are stronger in the song ensemble than in the tone pip-ensemble, fewer dimensions are needed for the song ensemble (Fig. 6). Also the large spatial extent of the correlations in the song ensemble implies that only the very low spatial frequencies are being sampled, as shown in the bottom right panel of Figure 5. The tone pip ensemble also samples a restricted region of the lower spatial frequencies, although a somewhat larger one than the song ensemble.

Therefore, our two stimulus ensembles sample a limited region of the spectral-temporal space that could be occupied by any sound, as illustrated by their spectra shown in the bottom panels of Figure 5. As explained in Materials and Methods in mathe-

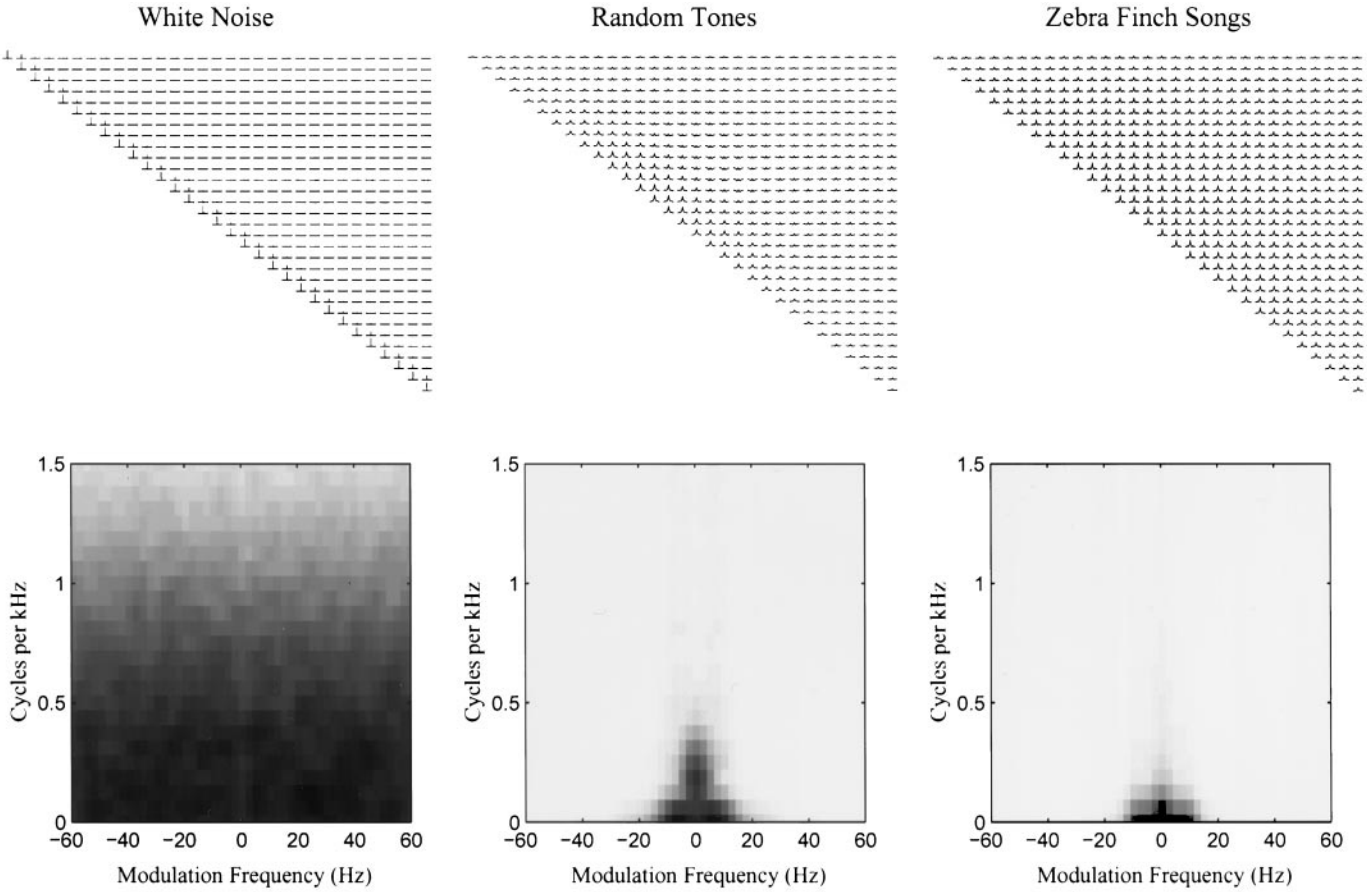

Figure 5. Stimulus autocorrelation matrices (*top panels*) and two-dimensional power spectra (*bottom panels*) for a white noise ensemble and the random tone-pip and song ensembles used in this paper. The diagonal corresponds to the autocorrelation of the log of amplitude envelope of each frequency band with itself. The off diagonal terms correspond to the cross-correlation between the log of amplitude envelopes across frequency bands. The bands are ordered in increasing frequency from *left* to *right* and *top* to *bottom*. The *top left corner* corresponds to the autocorrelation of the amplitude envelope in the 250 Hz band. Only the *top right* side of the matrix is shown because it is symmetric along the diagonal (with the time axis inverted for the *bottom left* entries). The time window of each autocorrelation function is from −200 to +200 msec as shown in Figure 1. The ideal white noise signal would have null functions in the off-diagonal terms and delta functions in the diagonal terms. The white noise signal approaches this ideal. The random tone pip ensemble is also closer to white noise than the song ensemble but still has the spectral and temporal structure that is a result of our design (see Materials and Methods and Results). The *bottom panels* show the two-dimensional power spectra of the stimulus ensemble. These two-dimensional power spectra are obtained by taking the square of the two-dimensional Fourier transform of the stimulus in their spectrographic representation. The two-dimensional power spectra illustrate the temporal and spectral correlations of the stimulus in the Fourier domain. The *x*-axis shows the frequency of temporal modulations, and the *y*-axis the frequency of the spectral modulations found in spectrograms of the different sounds. The two-dimensional power spectra are symmetric around the origin and therefore only the top quadrants are shown. The two-dimensional power spectra can be obtained directly from the autocorrelation matrix of the stimulus (*top row*), although the reverse is only true if the correlations along the spatial (spectral) dimension are stationary.

matical detail, by sampling such a limited stimulus, we are obtaining an estimation of the STRF that would be valid only for stimuli that sample identical stimulus space, unless the neural response to sounds outside the stimulus space being sampled is nonsignificant. Although we did not investigate all possible sounds, we found that for most neurons, the tone pip and the song ensemble elicited relatively strong responses. Much smaller responses were found, in many cases, with sustained white noise. Therefore, by limiting the stimulus space as we did, we suggest that we are effectively performing a better sampling of the acoustical patterns that are relevant to the neurons. Stimuli that include higher spatial and temporal frequencies (such as white noise) elicit much smaller neuronal responses because, in white noise, the specific sound patterns that excite the neuron occur at much smaller power levels, as they are embedded in other sounds. In addition, the neurons could exhibit nonlinear stimulus–response properties that could further depress or eliminate their responses to white noise.

Another major goal of this work was to compare for each neuron the STRF obtained with songs to the one obtained with simpler synthetic stimuli, similar to stimuli used in previous estimations of auditory STRFs using reverse correlation methods (deCharms et al., 1998). For the comparison to be valid, the space of sounds sampled by the synthetic stimulus ensemble should at least encompass all sounds present in the song. In theory, one might have chosen white noise, because it samples all acoustical space. However, white noise turned out to be a poor stimulus because it is unable to elicit sustained firing rates in most of the neurons. We therefore used a synthetic ensemble of tone pips that sampled a spectral-temporal space that was, by design, more restricted than white noise and similar to the one of zebra finch songs. The tone pips in our experiments had significant power in

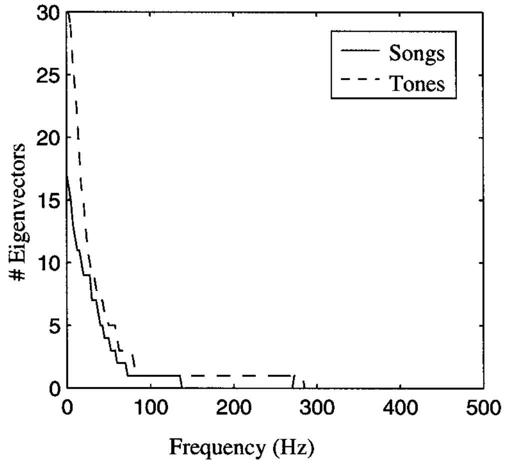

Figure 6. The nature of the second-order structure in the stimulus ensembles can be assessed by estimating the number of independent dimensions (or eigenvectors) of the autocorrelation matrix as a function of the frequency of amplitude modulation, *w*. The random-tone pip and the song ensemble have most of their second-order structure in the low range of amplitude modulations (see Results for details).

a similar (slightly larger) region of temporal and spatial frequencies. In fact, we designed our tone-pip ensemble so that any song could be well approximated by a linear combination of such trains of tone pips, although the reverse will not necessarily be true. If the neurons were linear, the STRF found with tone pips would then either be identical to the one found with the song ensemble or would potentially include spectral-temporal patterns that are not found in the song ensemble. Moreover, in linear neurons, the tone pip-generated STRF should predict the response to song as well as the song-generated STRF. We will show that neither of these theoretical predictions were true in our data, suggesting that our neurons are nonlinear and that higher order properties of sound besides their average spectral-temporal spectra or higher order characterizations of neural responses (such as adaptation) need to be considered to fully understand the neural response.

Calculation of the STRF

To demonstrate that our numerical approach can generate STRFs that have been properly corrected for the correlations present in the stimulus ensemble, we generated artificial neuronal data based on an STRF obtained from a simple neuron in L2a with a tone pip ensemble. The time-varying firing rate of a model neuron was then obtained by convolving the STRF with the spectrographic representation of each stimulus in our two ensembles. The firing rate was scaled to obtain an average of 10 spikes/sec which approximately matches the average firing data of the actual neurons studied here. To generate trials of neuronal data from our linear model neuron, we assumed Poisson firing statistics with time-varying firing rate. We generated 10 trials of artificial neuronal data for all the stimuli in our tone ensemble and in our song ensemble. From these artificial data, we calculated the STRFs with our numerical methods, in exactly the same way that we did with our real neuronal data. Figure 7 shows how the spike-triggered averages obtained from our two ensembles are radically different but that, once normalized by the stimulus autocorrelation, the estimated STRF obtained from the song and tone ensemble are in fact very similar. Moreover, these STRF are almost identical to the one that was used to generate the data. This simulation demonstrates that our method can compensate for the different correlations found in our two ensembles. Note, however, that if we had used a STRF that exhibited sensitivity to spectral temporal patterns that we had not sampled in our stimulus ensemble (such as high-frequency amplitude modulations), we would not have been able to recover the original STRF.

An example of the results of the STRF calculation on real data is illustrated in Figure 8 for the neuron of Figure 3. The spike-triggered spectrograms are shown on the left panels of Figure 8*A*, and the normalized STRFs are shown on the right panels for the two ensembles. As in the artificial data, the spike-triggered average spectrograms for the two ensembles differ significantly. In the case of actual data, however, it is unclear whether the differences are attributable simply to the additional spectral and temporal correlations that are found in the song ensemble or whether the differences reflect a different linear stimulus–response function discovered by the use of a natural sound ensemble. However, after we normalized the spike-triggered average spectrograms by the autocorrelations of the stimulus ensembles, we obtained the properly derived STRFs. The STRFs obtained for the two ensembles are, in the case of this particular neuronal site, much more similar than the spike-triggered averages. In both cases, the main excitatory region is around 1.5 kHz ~15 msec after the onset of the sound. Thus, most of the differences in the spike-triggered averages were attributable to stimulus properties and not to the neural responses. This example demonstrates that for presumably simpler, more linear neurons, the linear stimulus–response function found with the simple tone pip ensemble approximately matches the stimulus–response function found when complex natural sounds are used.

The tolerance level chosen for the normalization by the stimulus autocorrelation can have a drastic effect on the STRF. Figure 8*B* illustrates the effect of the tolerance level on the estimation of the STRF for the song ensemble for the example neuron of Figures 8*A* and 3. The bottom right panel of Figure 8*B* shows the integrated coherence, a measure that quantifies the quality of the STRF model by measuring its ability to predict responses to novel stimuli (see Materials and Methods). The coherence is maximal for a particular tolerance value, in this case 0.0005. This tolerance value also yields the STRF that is by visual inspection the closest to the one that is obtained with the tone-pip ensemble.

The STRF for song ensembles and the tone ensemble

We used the same procedure exemplified for the neuron of Figures 3 and 8 for all the neural sites that we recorded, obtaining a song-based STRF and a tone pip-based STRF at each site. Figure 9*A* illustrates three examples of STRFs in which the result obtained from the tone pip ensemble was markedly different from the one obtained from the song ensemble. These particular neuronal sites also had more complex STRFs than the one in Figure 8 and are more characteristic of our data set.

From the STRF one can also extract the more traditional characterizations of auditory processing such as spectral tuning and amplitude modulation rate tuning. The spectral tuning can be found by projecting the STRF along the frequency axis for the time values that show maximal and minimal firing probabilities (Fig. 9*B*, *solid* and *dashed lines*, respectively). The modulation rate tuning can be obtained by projecting the STRF onto the time axis for the best excitatory frequency and measuring its power as a function of temporal modulation frequencies. These procedures are illustrated in Figure 9*B* for the neuronal site of Figures 3 and 8. For the entire data set, we found that the peak excitatory spectral tuning ranged from 375 to 5125 Hz with tuning width

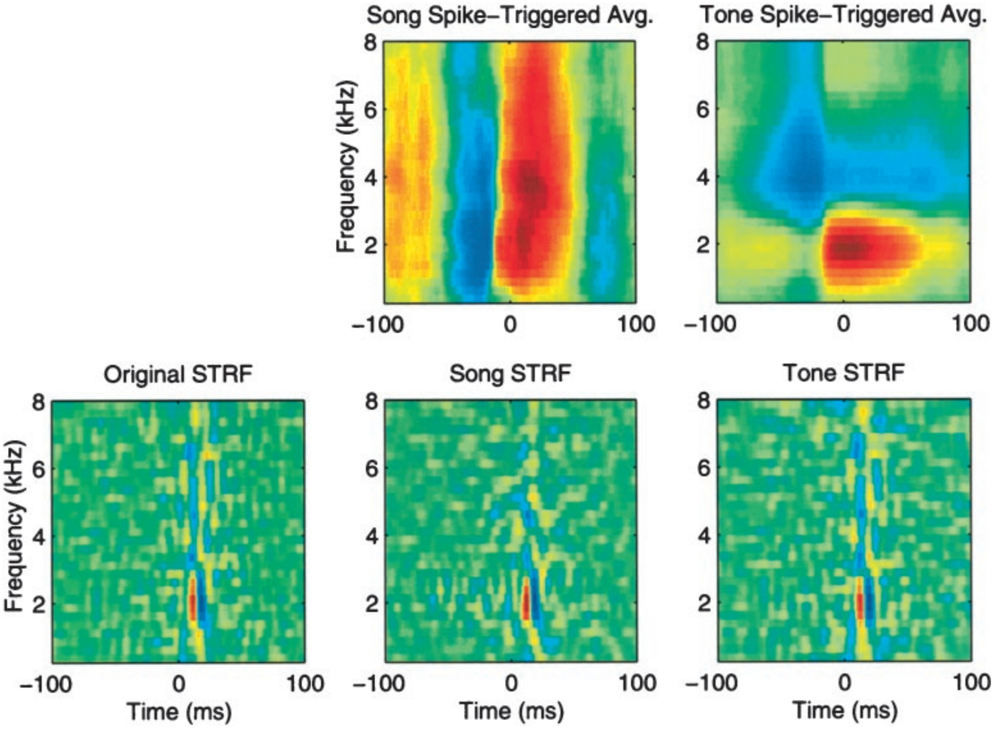

Figure 7. Validation of the STRF calculation and illustration of the normalization procedure. The "Original STRF" shown in the *bottom left* panel of the figure was used to generate artificial neural response data for the song ensemble and tone ensemble. The model neuron was linear and had Poisson firing statistics with a similar average firing rate to the actual neurons in our data set. These artificial data were then used to obtain estimates of the STRF with our methodology. The *top row* shows the spike-triggered average spectrograms that are obtained from these artificial data by averaging the stimulus spectrogram around each spike in a 400 msec window. These spike-triggered average spectrograms are only equal to the STRF of the neuron if a white noise stimulus (in the time/frequency representation of choice) is used. When correlations are present in the stimulus, either in the time dimension or across the frequency bands of the spectrogram, the spike-triggered average needs to be normalized. This normalization involves a deconvolution in time and decorrelation in frequency. The STRFs obtained from the spike-triggered averages by the normalization procedure are shown on the *bottom right* panels. As expected, a similar STRF estimate is obtained for both ensembles and these estimates are very close to the Original STRF that was used to generate the data.

factors (Q values at half maximum) ranging from 0.4 to 4.1. These values are similar to those found in avian forebrain by other groups (Zaretsky and Konishi, 1976; Muller and Leppelsack, 1985; Heil and Scheich, 1991). The peaks of the amplitude modulation transfer function ranged from 10 to 40 Hz. Amplitude modulation transfer functions have not been examined in detail in the avian forebrain (Heil and Scheich, 1991), but these values are similar to those found in mammalian auditory cortex (Schreiner and Urbas, 1988). In general, however, the STRF can give more information than just the spectral and amplitude modulation tuning, because it shows the specific spectral-temporal pattern that optimally excites the neuron. Such patterns can only be described in terms of their spectral and modulation rate tuning if the frequency and time dimensions are independent. In contrast, the STRF of a neuron can also show, for example, whether the neuron responds well to frequency sweeps or to other more complex spectral-temporal patterns. The neuronal sites shown on the top (N1) and bottom of Figure 9*A* (N3) are examples of more complex STRFs. A detailed analysis of the functionality that can be derived from the STRFs and the mapping of these functions throughout the avian auditory forebrain will be presented in a future publication.

Prediction of neural responses from the STRF

An additional advantage of describing the neural encoding with an STRF, as opposed to a more classical description based on tuning curves, is that the STRF can be used directly to predict the response of the neuron to any dynamical stimuli. This is a very important feature, because it allows validation of the description of the stimulus–response properties of the neurons. In this work, we focused our analysis on quantifying the goodness of fit of the predictions obtained from the STRF model to the actual data. We also analyzed the significance of differences that we found in these fits using two different stimulus ensembles to derive the STRFs.

If a neuron encodes the stimulus parameters of our choice (which can include a nonlinear transformation) in a linear fashion, then the STRF model will be able to completely predict the deterministic part of the response to any stimulus. Discrepancies between the predicted and the actual response would then be attributed to random neuronal noise. A problem arises immediately, however, when one finds very different STRFs for the same neuron with two different stimulus ensembles, as was the case for most of our data. In such cases, one finds two models to characterize the neural encoding, and will obviously make two different response predictions. Are the differences between these models significant in the sense that the goodness of their prediction is significantly different? Both to answer this question and to test the validity of the underlying linear assumption, we used the STRFs to obtain a prediction of the firing rate of each neuron to test stimuli taken from the same ensemble. This firing rate prediction was obtained by convolving the STRF with the test stimulus in its spectrographic representation. To prevent over-fitting, the particular stimulus–response data being tested were always excluded from the data used for estimation of the STRF (see Materials and Methods). The prediction could then be compared directly to the actual data, and the validity of the linear model could be assessed.

We quantified the quality of the prediction by calculating an unbiased correlation coefficient between predicted and actual spike rate (see Materials and Methods). The correlation coefficients from our entire data set are shown in Figure 10, *D* and *E*. When the STRF model was used to predict the response to a new stimulus that is nonetheless from the same type of ensemble used to derive the STRF (i.e., songs or tones), we found average correlations of 0.51 (range, 0.08–0.71) for the random tones and

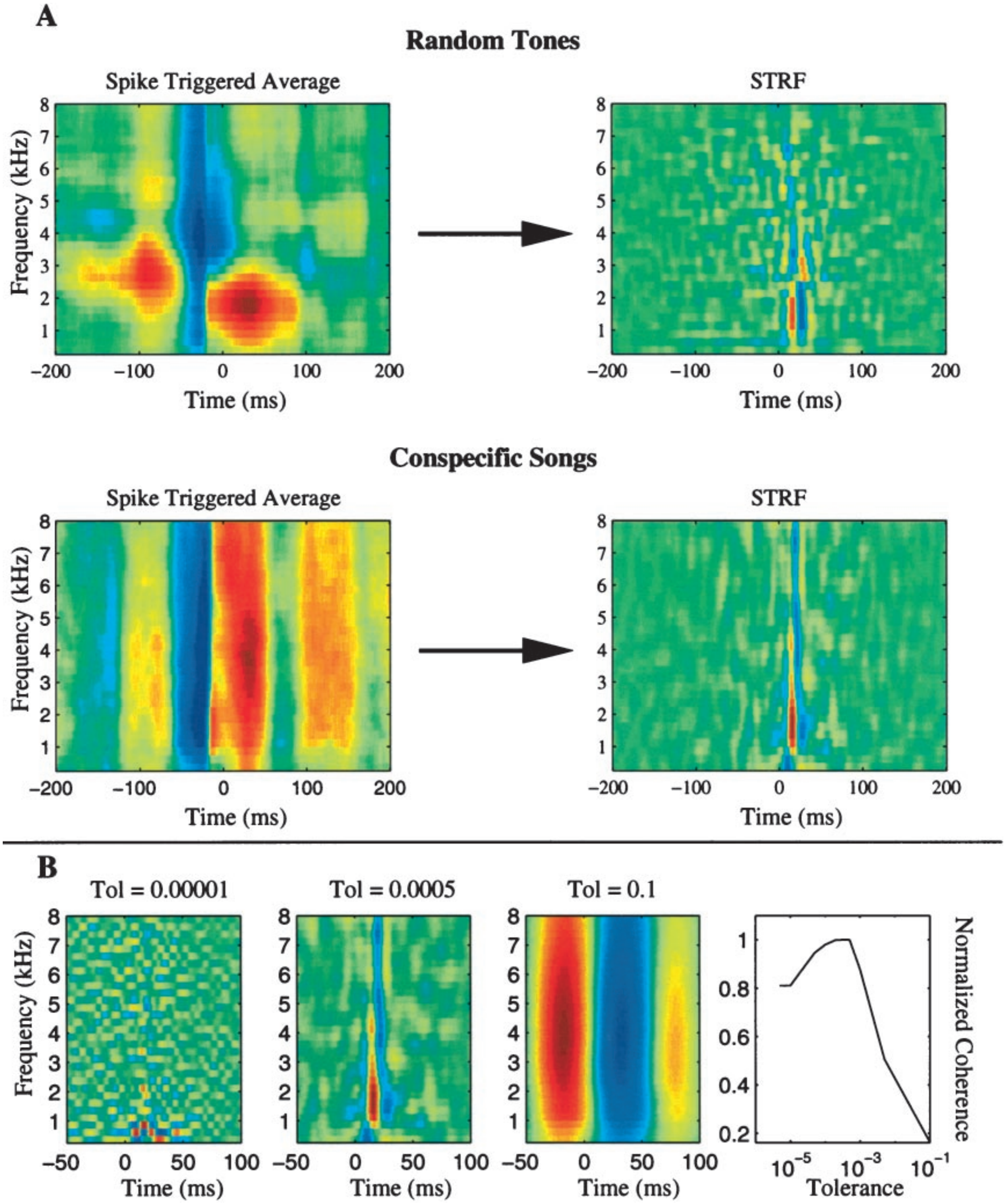

Figure 8. STRF calculation for a real neuron. The figure illustrates the STRF calculation explained in Results and in Figure 7 for the neuronal site of Figure 3. The calculation is based on all the responses that were obtained for the song ensemble and tone ensemble (10 trials for each of the 21 songs and 20 random tone sequences). As shown in *A*, for this particular neuron, the STRFs obtained from both ensembles are similar in that they exhibit similar areas of excitation. On the other hand, the spike-triggered average spectrograms were remarkably dissimilar. Most of the differences in the spike-triggered average were therefore attributable only to the statistical properties of the stimulus and not to the stimulus–response properties of this neuronal site. *B* shows examples of three song-based STRFs for this same neuron obtained with different noise tolerance levels, as explained in Materials and Methods. The time axis (*x*-axis) has been expanded from *A*. The normalized coherence between predicted response and actual response as a function of the tolerance level is plotted in the *rightmost* panel. The best predictions were obtained with a tolerance value of 0.0005.

0.45 (range, 0.12–0.66) for the song ensemble (Fig. 10*E*, "CC-matched"). These two distributions of correlation coefficients were very similar (p = 0.07, Wilcoxon signed rank test). In both cases, there was a wide range of fits showing that the linear model is a good approximation for some neurons and a much poorer approximation for others. However, both the response to tone pips and songs could be modeled with similar effectiveness as long as the STRF used in the fitting was obtained with the same type of stimulus ensemble.

The picture changed radically when the STRFs were switched, so that the predicted response to a stimulus from one type of ensemble was generated with the STRF obtained using the other ensemble. Figure 10*A–C* contrast the actual data with the predicted responses obtained with the matched STRF and with the

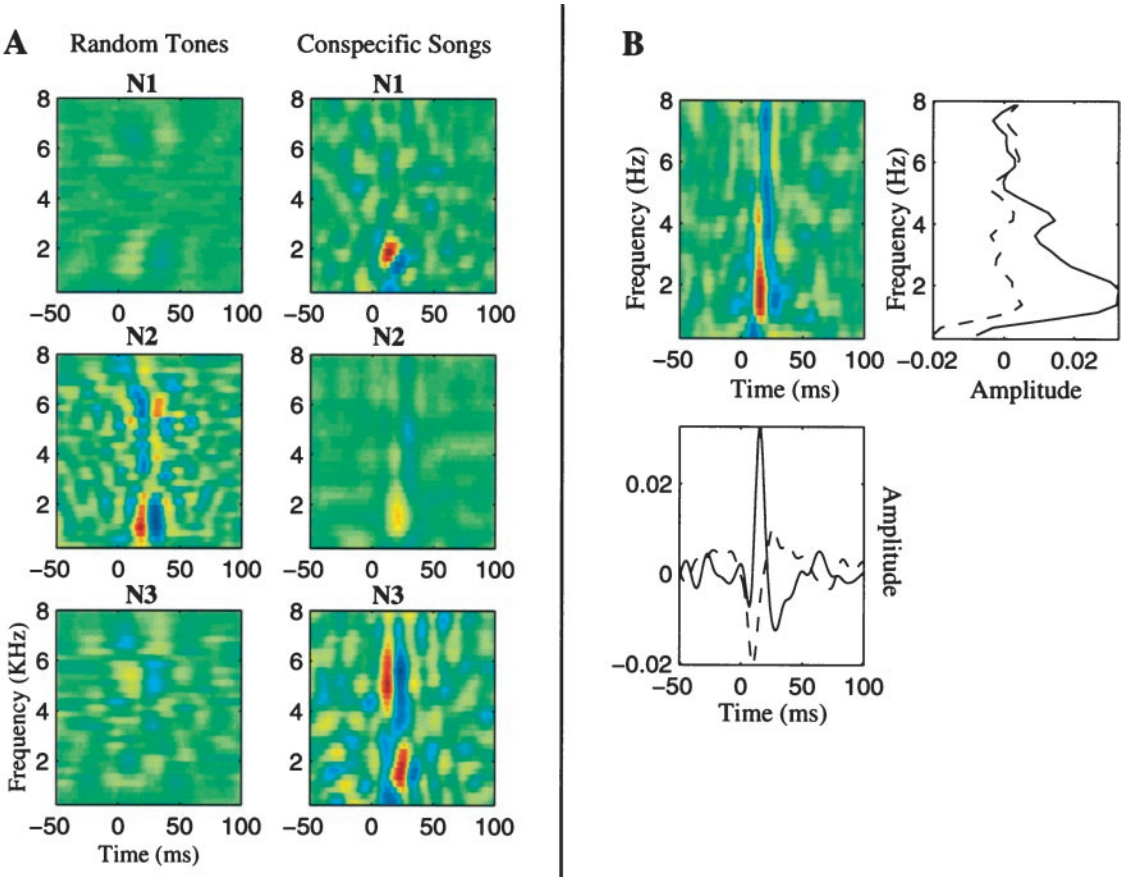

Figure 9. STRFs derived using the random tone pip and song ensemble for three neuronal sites (N1–N3) with complex stimulus response properties (*A*) and decomposition of an STRF onto its spectral axis and temporal axis (*B*). In *A*, each row corresponds to the two STRFs that were obtained for a particular site. The *left column* corresponds to the STRF calculated from the random tone pip ensemble, and the *right column* corresponds to the STRF calculated from the song ensemble. These three examples were chosen to illustrate cases in which the STRF obtained from the random tone pip ensemble was different from the one obtained from the song ensemble. These particular sites were in L1 (N1 and N3) and L3 (N2). The STRFs differ both in amplitude and in shape. The *top* neuronal site can be described as being sensitive to a moving spectral edge. Similar STRFs have been found in some cortical neurons (deCharms et al., 1998). The *bottom* neuronal site is sensitive to a temporal combination of a low-frequency sound followed by a high-frequency sound. Both of these complex spectral-temporal responses became evident only when the song ensemble was used. In *B*, the STRF of the neuron of Figure 8 is projected onto its spectral axis (*top right*) and temporal axis (*bottom left*). Such analyses can be used to extract the spectral and amplitude modulation tuning of the neurons as explained in Results. The *solid line* is the projection corresponding to the largest excitatory point, and the *dotted line* is the projection corresponding to the largest inhibitory point.

predicted responses obtained with the switched STRF. When the STRFs were switched, the prediction was a significantly poorer fit of the original stimulus response. The decrease in the quality of the prediction for all our data is shown in panels D and E, where the cross-correlation coefficients for the switched case are compared to those found in the matched case. The average switched correlation was 0.28 (range, −0.03–0.55) for the tone ensemble responses predicted using the song STRF and 0.19 (range, −0.25–0.50) for the song ensemble responses predicted using the tone STRF. These two distributions are not only significantly different from the corresponding distributions in the matched case (p = 0.0001), but also significantly different from each other (p = 0.0003). The prediction of the response to the song ensemble using the STRFs obtained with tone pips resulted in a larger deterioration of the fit than did the prediction of tone responses with the song-derived STRFs (Fig. 10*E*, "CC switched"). Based on our results with these two ensembles, we conclude that the STRFs obtained from the song ensemble are overall better models of the stimulus–response function of these neurons.

The error in our estimate of the quality of the prediction (SE of CCs for a particular neuron) is attributable in part to noise in the estimation of the STRFs from a limited sample set. The noise in the STRFs will also lead to a biased estimation of the CC (lower values than actual). Although we corrected the bias in the CCs for the noise in the PSTH, we did not correct for the bias caused by the noise in the STRFs. An estimate of the magnitude of this bias can be obtained from the SE in the CCs. The SE was obtained with the jackknife resampling technique and is shown in the error bars of Figure 10*D* for each data point. By comparing the error in the CCs obtained with song STRFs to those obtained with tone STRFs, we can show that the difference in the quality of the predictions of song responses by song- and tone-derived STRFs is not attributable simply to differences in the noise-generated bias in the CCs (as might have been true, for instance, if all tone responses were less vigorous than song responses, leading to poorer signal-to-noise ratios of tone STRFs). The SE in the CCs was similar for all predictions obtained with song and tone STRFs (mean CC SE of song STRF on song = 0.041 vs mean CC SE of tone STRF on song = 0.037; mean CC SE of tone STRF on tone = 0.06 vs mean CC SE of song STRF on tone = 0.053). Therefore, we conclude that the small differences of signal-to-noise ratio of the tone and song-derived STRFs cannot

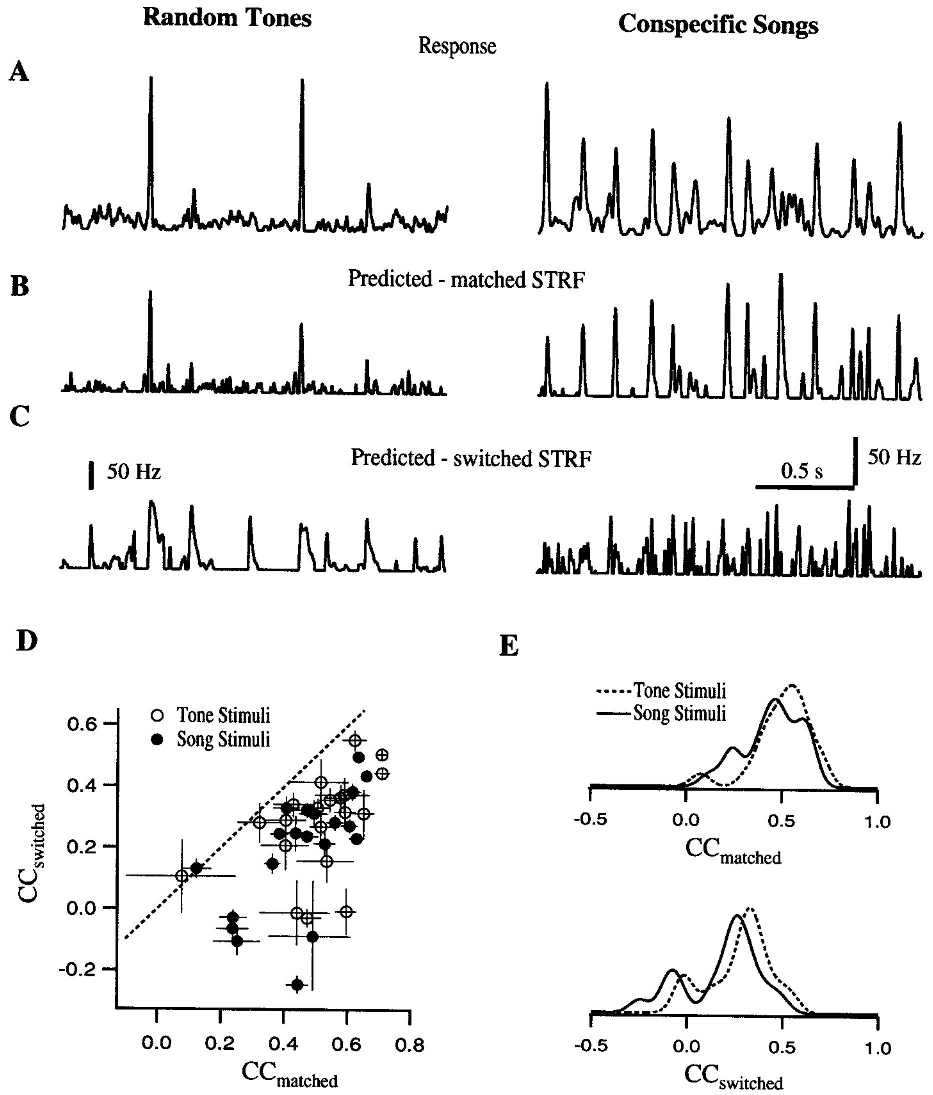

Figure 10. Comparison of the predictions obtained from the two STRFs calculated for each neuronal site. *A* shows the actual firing rate of a particular neuronal site in response to random tone-pip stimuli (*left*) and to a song (*right*). *B* shows the firing rate predicted using the STRF calculated with the corresponding stimulus ensemble. *C* shows the predicted response calculated after switching the STRFs. This example illustrates that the STRFs obtained from the different ensembles can give radically different results. *D*, Scatter plot of the correlation coefficients (*CC*) between the predicted and the estimated firing rates, for switched STRFs versus matched STRFs. For each neuronal site, the CCs are calculated for the songs and tone pips. The *solid dots* show the CCs for the prediction to song stimuli, obtained either with song-STRFs (CC matched on the x-axis) or with the tone STRFs (CC switched on the y-axis). The *open dots* show the predictions for tone stimuli with the matched and switched filters. If the STRFs generated with the two different stimulus ensembles were identical, the points would lie on the $x = y$ line shown in dots. Most points are *below* the line, although some are close to the line and some are closer to 0, showing the range of differences. *E*, The data in *D* are replotted by projecting all the points onto the x-axis (*top plot*) and onto y-axis (*bottom plot*), thus showing the individual distributions of CC-matched and CC-switched for the predictions to song and tone-pip stimuli. The distributions are obtained by convolving the raw data with a smoothing kernel of 0.05. The distribution of CC-matched for the two types of stimuli are not significantly different, but the CC-switched for the response to songs predicted using the tone STRFs are significantly smaller than the CC-switched for tone pips. In addition, all the CC-switched are smaller than or equal to the CC-matched.

explain the significant difference in song and tone prediction seen in our results.

In all the data presented so far, we calculated the STRFs from the log of the amplitude envelopes of our spectrographic decomposition of sound. When the same calculations were done without using the log transformation, the results were similar for the tone ensemble (mean CC = 0.52) but much worse with the song ensemble (mean CC = 0.26). By including this additional preprocessing step of using the log transform in our methodology, we illustrate how static nonlinearities can easily be included in the calculation of STRFs. This also demonstrates that nonlinearities in the neural encoding are different for different stimulus ensembles. In this particular case we found that the relationship between sound intensity and probability of firing was more linear for the tone-pip stimuli than for the song stimuli. Although we did not systematically investigate all possible nonlinear transformations, we found that the log relationship worked well for the natural ensemble.

DISCUSSION

Complex sensory cortical neurons are well known to exhibit many nonlinearities, and these nonlinear effects become even more evident to the experimenter when complex stimuli with natural statistics are used, both in the auditory (Schwarz and Tomlinson, 1990; Calhoun and Schreiner, 1998) and in the visual modalities (Gallant et al., 1998). Because of this effect, the response to complex natural sounds including vocalizations often cannot be explained from neural responses to simpler sound stimuli (Rauschecker et al., 1995). Moreover, sensory systems may be optimized for the encoding of natural stimuli (Rieke et al., 1995; Dan et al., 1996). One would like to be able to systematically study how such complex nonlinear neurons encode stimuli and how they mediate natural behaviors. For example, for very nonlinear neurons, referred to as combination-sensitive neurons, a systematic decomposition of the ethologically relevant sound into small parts has been revealed to be very useful (Suga, 1990; Margoliash and Fortune, 1992). However, for less specialized auditory neurons, such as the majority of those found in auditory cortex of nonspecialized mammals or in the avian forebrain homolog of auditory cortex, this approach will fail. Such auditory neurons respond to a very large ensemble of sounds, and their stimulus–response function cannot be explored systematically by attempting to relate their firing rate to identified stretches of sounds, such as the syllables in an animal's vocalizations. To be able to extract the stimulus–response encoding function of these nonspecialized complex auditory neurons, the concept of an STRF has much to offer. First, the estimation of the STRF involves finding the spectral-temporal sound patterns that excite the neuron in a systematic manner. Second, the STRF is a model of the stimulus–response function of the neuron that hypothetically could be used to model the response to any stimulus. Finally, the STRF description encompasses classical characterizations of higher auditory areas such as spectral tuning and amplitude modulation tuning.

Calculating STRFs for complex neurons is a problem, however, because the STRF is a linear model and the complex neurons of interest exhibit significant nonlinearities. When the STRFs are calculated with broad band stimuli, the spectral-temporal features that would elicit large responses from complex neurons do not occur in isolation, but instead occur with small power and are embedded in large amounts of what the neurons evaluate as extraneous sound. For nonlinear neurons, such broad-band stimuli yield no responses or responses that are much smaller than expected from the neuronal encoding model based on STRFs obtained with narrow band stimuli. Moreover, the simple reverse correlation method that has been used so far to estimate STRFs relies on the use of such random broadband stimuli. For these reasons and because STRFs derived from such stimulus ensembles are poor predictors of neuronal responses to natural sounds, it has been suggested that the concept of a linear STRF is of limited use in understanding the auditory system (Aertsen and Johannesma, 1981b; Eggermont et al., 1983a; Nelken et al., 1997). On the other hand, the approach has had much greater success in explaining the responses of neurons in A1 when the STRF is obtained from complex synthetic stimuli with limited spectral and temporal modulation bandwidth and/or with methods that do not rely on the reverse correlation (Kowalski et al., 1996b; Escabi et al., 1998; Shamma et al., 1998; Versnel and Shamma, 1998). Our study with natural sounds supports these studies and extends the use of the STRF to the analysis of the neural response of complex auditory neurons.

In our analysis, we extended the reverse-correlation method so that we could derive an STRF from any stimulus ensemble. This extension involves a normalization of the spike-triggered average stimulus by the correlations (or second order statistics) of the spectral-temporal patterns of the stimulus ensemble. By doing so, we were able to estimate STRFs for ensembles that had limited spectral and temporal modulation bandwidth and specific spectral-temporal correlations. This was particularly useful because we found that auditory neurons found in the field L region of the zebra finch responded strongly to the sounds of conspecific song. Our analysis involves describing the second-order statistics of relevant stimulus ensembles and estimating for each a “stimulus ensemble-dependent” STRF that is nonetheless correctly normalized for the stimulus structure. We could then compare the STRFs obtained with different stimulus ensembles to obtain a first-order description of the stimulus–response function of our neurons.

We found that zebra finch songs have limited spectral and temporal modulation bandwidths and very high correlations in the amplitude envelopes across frequency bands. For a stimulus ensemble consisting of these songs, we showed that the estimated STRFs yielded spectral and amplitude modulation tuning that were in accord with previous work using simple synthetic sounds. We also quantified the goodness of the linear assumption underlying the STRF model by using the STRF to predict the neural response to novel stimuli and comparing that to the actual response. We found that the quality of the linear approximation of the response characteristics of neurons in L1, L2, and L3 varied over a wide range, working well for certain neurons but not for other, presumably more nonlinear, neurons. In the few other experiments in which a direct estimation of the validity of the classic STRF model has been made, the quality of the model (measured, as we did, by the mean and range of the correlation coefficients between predicted and actual response) was similar to what we found here in the “matched” case, where the STRFs used to predict the response were obtained from stimuli of the same type as the test stimuli (Eggermont et al., 1983b; Kowalski et al., 1996b). In this respect our neurons are as linear (or nonlinear) as other auditory neurons both in the periphery (Eggermont et al., 1983b) and in primary auditory cortex (Kowalski et al., 1996b).

We then tested the generality of the STRF model and further examined the linearity of our neurons by also calculating a second STRF for each neuron, using an ensemble of tone pips that had similar spectral and temporal modulation bandwidth. The tone-pip stimulus ensemble was designed so that it would include all sounds in the zebra finch song but with clearly different statistical structure. This stimulus ensemble also elicited relatively high firing rates from field L neurons, although it yielded much smaller mean rates than the song ensemble. By comparing the results obtained from the two STRFs, we showed that many neurons in the avian auditory forebrain respond in a nonlinear fashion to the particular spectral and temporal statistics of species-specific vocalizations, because the STRF obtained from the song ensemble was significantly different from the STRF obtained from simpler tone pips. Moreover, when we tested how well the linear model would generalize by using the STRF calculated from one ensemble to predict the responses to stimuli from a different ensemble, we found that the quality of the prediction decreased significantly. In particular, the STRF obtained from tone pips is a poor model for the responses found to natural sounds. This represents another way in which to quantify the non-linearity of the neurons. Although this effect had been observed in a qualitative manner (Eggermont et al., 1983b; Nelken et al., 1997), it had not been quantified, because STRFs from natural sounds or images have not been estimated previously.

Despite this nonlinearity, we found that by calculating the STRFs from natural sounds, we could predict the response to new songs as well as the tone-derived STRFs could predict the response to simple stimuli. In fact, the prediction to the natural sounds with the song STRFs for the forebrain auditory neurons studied here is as good as the prediction that has been found in visual thalamic neurons (which are presumably more linear) to natural scenes, obtained with a classic STRF (Dan et al., 1996). These results show that the limitations of the STRF model are not only caused by its underlying linear assumption, but also by the fact that STRFs have previously been estimated with nonoptimal stimulus ensembles. Finally, the fact that song STRFs predicted tone responses better than the tone-pip STRFs predicted song responses raises the possibility that, in general, a better fit for the stimulus–response functions of neurons will be obtained with a stimulus ensemble with richer spectral-temporal correlations such as those found in natural sounds. Figure 11 exemplifies how a particular nonlinear stimulus response function could lead to all the effects that we have described in our discussion about the neurons in our data set.

Our methodology and results also suggest ways in which the STRF framework could be further extended. We have shown that by calculating STRFs from multiple stimulus subspaces one can greatly improve the fit of the actual stimulus–response function.

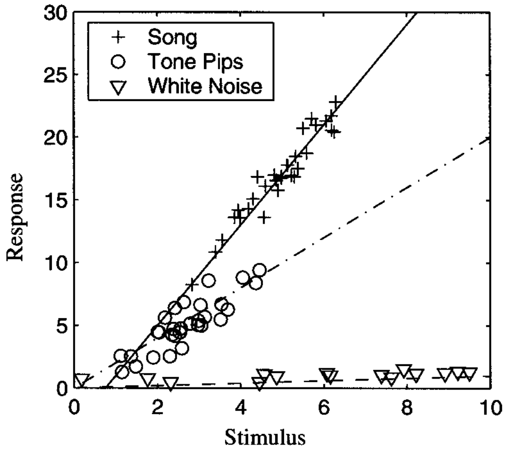

Figure 11. For illustrative purposes, we show the distribution of stimulus response points for a hypothetical nonlinear neuron that exhibits the properties of the neurons in our data set. The multidimensional space (31 frequency bands × 400 points in time or 1600 dimensions in our case) used to represent the stimulus is collapsed onto the one-dimensional *x*-axis, and neural response is shown on the *y*-axis. White noise stimuli sample the entire space, whereas tone pips and songs only sample a region of that space. The hypothetical neuron has poor responses to white noise but significant responses to tone pips and songs, with songs being the preferred stimuli. A linear fit for each stimulus subset shows that relatively good stimulus–response predictions can be found, although the fit obtained from one ensemble is not a good model for the stimulus–response function found in a different ensemble. In particular, the fit to the white noise data is a very poor predictor. The fit to the song ensemble is better at predicting the response to tones than vice versa. The data in our ensemble exhibit similar patterns (albeit with on average much smaller correlations) in multidimensional space where the line is replaced by a hyperplane. An STRF is the set of coefficients that define such a hyperplane.

One might therefore model the nonlinear stimulus–response function by effectively doing a piece-wise linear fit in multiple subspaces of stimuli. Such a description of the stimulus–response function of a neuron would go hand in hand with the description of the relevant statistics (in this case second order correlations) of the different stimulus ensembles. This is crucial, because the experimenter is now only sampling a limited subspace of all possible stimuli. For completeness, one would have to ensure that either the sampling of sound stimuli is based on ethological distributions and/or on a broad investigation of all sounds capable of eliciting significant responses. Finally, we have also shown how to improve on this first-order model of the STRF by including particular static nonlinearities in the calculation (see Results). The nonlinearities can also be studied by examining the systematic deviations of the data from the predicted response obtained with the STRF model (Shamma et al., 1998). Our methodology combined with such potential extensions could allow for a better characterization of high-level sensory processing: nonlinear effects are clearly important, but at the same time the linear model can explain a large fraction of the response and will undoubtedly be part of a more complete model. Our approach is general and could also be applied at various levels of the visual system to validate the effectiveness of the classical STRF and to further understand the response of complex visual cortical neurons to natural images.

APPENDIX

Invertible spectral-temporal representation signals

All of the possible time-frequency representations can be derived from the Wigner distribution of a signal, which, in a mathematical sense, can be thought of as the canonical distribution (Cohen, 1995, their Chapter 9). The Wigner distribution satisfies the requirements of probability theory that ensure that it has the properties of a generalized joint-distribution function (Cohen, 1995, their Chapter 8). Moreover the signal can be recovered from the Wigner distribution up to a constant phase factor because the Wigner distribution is unique and invertible (Cohen, 1995, pp 127, 128). The Wigner distribution of a signal, $s(t)$, is given by:

$$W(t, w) = \frac{1}{2\pi} \int s^*(t - \tau/2)s(t + \tau/2)e^{-j\tau w}d\tau. \qquad (1)$$

All other time-frequency representations of $s(t)$, $G(w, t)$ can be written as a two-dimensional filtered version of the $W(t, w)$:

$$G(t, w) = \int\int g(t' - t, w' - w)W(t', w')dt'dw', \qquad (2)$$

where $g(t, w)$ is the two-dimensional filter. In practice, Equation 2 is written in the Fourier domain, where the convolution can be replaced by a product. In that formulation, the two-dimensional Fourier transform of $g(t, w)$ is called the kernel of $G(t, w)$ and is written as $\Phi(\vartheta, \tau)$:

$$g(t, w) = \frac{1}{4\pi^2} \int\int \Phi(\theta, \tau)e^{j\vartheta t + j\tau w}d\vartheta d\tau$$

$$\text{and} \quad \Phi(\theta, \tau) = \int\int g(t, w)e^{-j\vartheta t - j\tau w}dtdw \qquad (3)$$

Using Equations 1 and 3, Equation 2 can then be rewritten as (Cohen, 1995, their Chapter 9):

$$G(t, w) = \frac{1}{4\pi^2} \int\int\int e^{-j\vartheta t - j\tau w + j\vartheta u}\Phi(\vartheta, \tau)s^*(u - \tau/2)s(u + \tau/2)dud\tau d\vartheta \qquad (4)$$

The spectrographic class of time-frequency representations is given by:

$$G(t, w) = \left| \frac{1}{\sqrt{2\pi}} \int s(\tau)k(\tau - t)e^{-jw\tau}d\tau \right|^2,$$

where $k(t)$ is the time window used in the short-time Fourier transform. It can be shown that the kernel of a spectrographic representation is given by the Wigner distribution of $k(t)$ (Cohen, 1995, p 141). A spectrographic representation is therefore invertible (except for the fixed phase) if the Wigner distribution of $k(t)$ is not equal to zero anywhere where the Fourier transform of the Wigner distribution of the signal is nonzero. Therefore, to be generally invertible, a spectrographic representation must have $k(t)$ with a nonzero Wigner distribution everywhere (for a mathematically equivalent proof, see also Cohen, 1995, p 108). The Gaussian window used in our work satisfies this constraint: writing $k(t)$ as:

$$k(t) = (\alpha/\pi)^{1/4}e^{-\alpha t^2/2},$$

one finds that $W(t, w)_k$ is equal to

$$W(t, w)_k = (1/\pi)e^{-\alpha t^2 - w^2/\alpha}.$$

The remaining issue deals with the transformation from the continuous theoretical framework to the discrete representation that is used in practice. Samples in time and frequency must be taken at intervals that are small enough to capture all significant spatial frequencies, where space is the time-frequency grid. The amplitude envelopes obtained in a spectrographic representation of sound are effectively band-limited: the upper frequency is given by the bandwidth of the time window used to calculate the spectrogram (Flanagan, 1980; Theunissen and Doupe, 1998). A fine frequency grid must also be used to ensure that all frequencies are represented. We used these properties to define our sampling rate in time and the overlap of our frequency bands in frequency. We have shown empirically in our previous work that the entire signal (except for an absolute fixed phase) can effectively be recovered from our specific decomposition (Theunissen and Doupe, 1998).

Generalized STRF

Because there are many time-frequency representations, there will be correspondingly many possible definitions of the STRF. For theoretical reasons, an STRF based on the Wigner distribution has many appeals. First of all, it is equivalent to the second-order Volterra term in the functional expansion between the sound pressure waveform and the neural response, and in this sense it is the natural extension of the first-order expansion that is used for phase-locked units. Second, it is based on a unique and invertible time-frequency representation. For these reasons, the STRF in auditory research was originally defined based on this mathematically appealing definition (Aertsen et al., 1981; Kim and Young, 1994). However, the STRF based on the Wigner distribution, called here invariant STRF or $\mathrm{STRF_I}$, has a distributed representation that is often difficult to interpret without some sort of smoothing procedure (Kim and Young, 1994). Recently researchers have, therefore, adopted a more general definition of the STRF that is based on other time-frequency representations, called here spectrographic STRF or $\mathrm{STRF_S}$, effectively performing the desired smoothing operation. Indeed, because other representations are in fact filtered versions of the Wigner distribution, it can be shown that any $\mathrm{STRF_S}$ can be obtained from $\mathrm{STRF_I}$ by a two-dimensional filtering (or smoothing operation) (Klein et al., 2000). The two-dimensional filter is the inverse of the two-dimensional filter that is used to transform the Wigner distribution into the general time-frequency representation. The proof goes as follows: The response $r(t)$ is found by convolving the STRF with the time-frequency representation of the signal $s(t)$:

$$r(t) = \int\int STRF_S(t - t', w')G(t', w')dt'dw', \qquad (5)$$

in terms of $G(t, w)$, or:

$$r(t) = \int\int STRF_I(t - t'', w'')W(t'', w'')dt''dw'', \qquad (6)$$

in terms of $W(t, w)$.

Replacing Equation 2 in Equation 5,

$$r(t) = \int\int STRF_S(t - t', w')\left[\int\int g(t'' - t', w'' - w')\,W(t'', w'')dt''dw''\right]dt'dw' \qquad (7)$$

Therefore,

$$STRF_I(t - t'', w'') = \int\int g(t'' - t', w'' - w')STRF_S(t - t', w')dt'dw'$$

or

$$STRF_I(t, w) = \int\int g(t' - t, w - w')STRF_S(t', w')dt'dw'. \qquad (8)$$

Therefore, $\mathrm{STRF_S}$ will only be theoretically equivalent to $\mathrm{STRF_I}$ if one is able to invert the filtering operation shown in Equation 8. Just as for the transformation between G and W, this operation will be possible only if the kernel of G is nonzero at all Fourier values where $\mathrm{STRF_I}$ has significant power. Because $\mathrm{STRF_I}$ is a priori not known, a rigorous implementation would require a time-frequency distribution with nonzero kernel. In this manner, one would be able not only to find $\mathrm{STRF_I}$, but also to compare the STRFs found by different research groups.

REFERENCES

Aertsen AM, Johannesma PI (1981a) The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol Cybern 42:133–143.

Aertsen AM, Johannesma PI (1981b) A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. Biol Cybern 42:145–156.

Aertsen AM, Olders JH, Johannesma PI (1981) Spectro-temporal receptive fields of auditory neurons in the grassfrog. III. Analysis of the stimulus-event relation for natural stimuli. Biol Cybern 39:195–209.

Boer ED, Kuyper P (1968) Triggered correlation. IEEE Trans Biomed Eng 15:169–179.

Cai D, DeAngelis GC, Freeman RD (1997) Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. J Neurophysiol 78:1045–1061.

Calhoun B, Schreiner C (1998) Spectral envelope coding in cat primary auditory cortex: linear and non-linear effects of stimulus characteristics. Eur J Neurosci 10:926–940.

Clopton BM, Backoff PM (1991) Spectraltemporal receptive fields of neurons of cochlear nucleus of guinea pig. Hear Res 52:410–422.

Cohen L (1995) Time-frequency analysis. Englewood Cliffs, NJ: Prentice Hall.

Dan Y, Atick JJ, Reid RC (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J Neurosci 16:3351–3362.

DeAngelis GC, Ohzawa I, Freeman RD (1995) Receptive-field dynamics in the central visual pathways. Trends Neurosci 18:451–458.

deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280:1439–1443.

De Valois RL, Cottaris NP (1998) Inputs to directionally selective simple cells in macaque striate cortex. Proc Natl Acad Sci USA 95:14488–14493.

Efron B (1982) The jackknife, the bootstrap and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Eggermont JJ (1993) Wiener and Volterra analyses applied to the auditory system. Hear Res 66:177–201.

Eggermont JJ, Smith GM (1990) Characterizing auditory neurons using the Wigner and Rihacek distributions: a comparison. J Acoust Soc Am 87:246–259.

Eggermont JJ, Aertsen AM, Johannesma PI (1983a) Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. Hear Res 10:167–190.

Eggermont JJ, Aertsen AM, Johannesma PI (1983b) Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. Hear Res 10:191–202.

Escabi MA, Schreiner CE, Miller LM (1998) Dynamic time-frequency processing in the cat midbrain, thalamus, and auditory cortex: spectrotemporal receptive fields obtained using dynamic ripple stimulation. Soc Neurosci Abstr 24:1879.

Flanagan JL (1980) Parametric coding of speech spectra. J Acoust Soc Am 68:412–419.

Fortune ES, Margoliash D (1992) Cytoarchitectonic organization and morphology of cells of the field L complex in male zebra finches (*Taenopygia guttata*). J Comp Neurol 325:388–404.

Gallant JL, Connor CE, Van Essen DC (1998) Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing [corrected and republished article originally printed in NeuroReport 1998 Jan 5;9:85–90]. Neuroreport 9:2153–2158.

Hartline HK (1940) The receptive fields of optic nerve fibers. Am J Physiol 130:690–699.

Heil P, Scheich H (1991) Functional organization of the avian auditory cortex analogue. I. Topographic representation of isointensity bandwidth. Brain Res 539:110–120.

Hinkley DV (1978) Improving the jackknife with special reference to correlation estimation. Biometrika 65:13–21.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol (Lond) 160:106–154.

Kim PJ, Young ED (1994) Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory nerve fibers. J Acoust Soc Am 95:410–422.

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comp Neurosci, in press.

Konishi M (1985) Birdsong: from behavior to neuron. Annu Rev Neurosci 8:125–170.

Kowalski N, Depireux DA, Shamma SA (1996a) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophysiol 76:3503–3523.

Kowalski N, Depireux DA, Shamma SA (1996b) Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. J Neurophysiol 76:3524–3534.

Kuffler SW (1953) Discharge patterns and functional organization of mammalian retina. J Neurophysiol 16:37–68.

Lewicki MS (1996) Intracellular characterization of song-specific neurons in the zebra finch auditory forebrain. J Neurosci 16:5855–5863.

Lewicki MS, Arthur BJ (1996) Hierarchical organization of auditory temporal context sensitivity. J Neurosci 16:6987–6998.

Margoliash D (1983) Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. J Neurosci 3:1039–1057.

Margoliash D (1986) Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. J Neurosci 6:1643–1661.

Margoliash D, Fortune ES (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. J Neurosci 12:4309–4326.

Marler P (1970) A comparative approach to vocal learning: song development in white-crowned sparrows. J Comp Physiol Psychol 71:1–25.

Marmeralis P, Marmeralis V (1978) Analysis of physiological systems. The white noise approach. New York: Plenum.

Miller RG (1974) The jackknife: a review. Biometrika 61:1–15.

Muller CM, Leppelsack H-J (1985) Feature extraction and tonotopic organization in the avian forebrain. Exp Brain Res 59:587–599.

Nelken I, Kim PJ, Young ED (1997) Linear and nonlinear spectral integration in type IV neurons of the dorsal cochlear nucleus. II. Predicting responses with the use of nonlinear models. J Neurophysiol 78:800–811.

Nottebohm F, Stokes TM, Leonard CM (1976) Central control of song in the canary, *Serinus canarius*. J Comp Neurol 165:457–486.

Ohlemiller KK, Kanwal JS, Suga N (1996) Facilitative responses to species-specific calls in cortical FM-FM neurons of the mustached bat. NeuroReport 7:1749–1755.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Solution of linear algebraic equations. In: Numerical recipies in C, pp 23–105. Cambridge, UK: Cambridge, UP.

Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268:111–114.

Rieke F, Bodnar DA, Bialek W (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proc R Soc Lond B Biol Sci 262:259–265.

Ringach DL, Hawken MJ, Shapley R (1997a) Dynamics of orientation tuning in macaque primary visual cortex. Nature 387:281–284.

Ringach DL, Sapiro G, Shapley R (1997b) A subspace reverse-correlation technique for the study of visual neurons. Vision Res 37:2455–2464.

Scheich H, Langner G, Bonke D (1979) Responsiveness of units in the auditory neostriatum of the guinea fowl (*Numida meleagris*) to species-specific calls and synthetic stimuli. II Discrimination of Iambus-like calls. J Comp Physiol [A] 132:257–276.

Schreiner CE, Calhoun BM (1994) Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. Audit Neurosci 1:39–61.

Schreiner CE, Urbas JV (1988) Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hear Res 32:49–64.

Schwarz DW, Tomlinson RW (1990) Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey (*Macaca mulatta*). J Neurophysiol 64:282–298.

Shamma SA, Depireux DA, Klein DJ, Simon JZ (1998) Representation of dynamic broadband spectra in auditory cortex. Soc Neurosci Abstr 24:402.

Suga N (1990) Cortical computational maps for auditory imaging. Neural Networks 3:3–21.

Theunissen FE, Doupe AJ (1998) Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVc of male zebra finches. J Neurosci 18:3786–3802.

Vates GE, Broome BM, Mello CV, Nottebohm F (1996) Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches (*Taeniopygia guttata*). J Comp Neurol 366:613–642.

Versnel H, Shamma SA (1998) Spectral-ripple representation of steady-state vowels in primary auditory cortex. J Acoust Soc Am 103:2502–2514.

Zaretsky MD, Konishi M (1976) Tonotopic organization in the avian telencephalon. Brain Res 111:167–171.