J Comput Neurosci (2006) 21:5–20
DOI 10.1007/s10827-006-7059-4

Sound representation methods for spectro-temporal receptive field estimation

Patrick Gill · Junli Zhang · Sarah M. N. Woolley ·
Thane Fremouw · Frédéric E. Theunissen

Received: 10 June 2005 / Revised: 4 January 2006 / Accepted: 10 January 2006 / Published online: 22 April 2006

Abstract The spectro-temporal receptive field (STRF) of an auditory neuron describes the linear relationship between the sound stimulus in a time-frequency representation and the neural response. Time-frequency representations of a sound in turn require a nonlinear operation on the sound pressure waveform and many different forms for this non-linear transformation are possible. Here, we systematically investigated the effects of four factors in the non-linear step in the STRF model: the choice of logarithmic or linear filter frequency spacing, the time-frequency scale, stimulus amplitude compression and adaptive gain control. We quantified the goodness of fit of these different STRF models on data obtained from auditory neurons in the songbird midbrain and forebrain. We found that adaptive gain control and the correct stimulus amplitude compression scheme are paramount to correctly modelling neurons. The time-frequency scale and frequency spacing also affected the goodness of fit of the model but to a lesser extent and the optimal values were stimulus dependant.

Keywords Receptive field · Zebra finch · STRF · Reverse correlation · Auditory cortex

Action Editor: Israel Nelken

P. Gill · F. E. Theunissen (✉)
Biophysics Group, University of California at Berkeley,
3210, Tolman Hall, Berkeley, CA 94720
e-mail: theunissen@berkeley.edu

J. Zhang · S. M. N. Woolley · T. Fremouw · F. E. Theunissen
Department of Psychology and Neurosciences Institute,
University of California at Berkeley,
3210, Tolman Hall, Berkeley, CA 94720

Introduction

Auditory receptive fields

In sensory neurophysiology, one often describes the linear dynamic receptive field of a sensory neuron, which is the linear filter that, when convolved with the stimulus, produces the best approximation to the neuron's output (Ghazanfar and Nicolelis, 2001). In some modalities, one can find an appropriate elementary representation of stimuli for the estimation of dynamic receptive fields. For example, the time-varying bitmap is a natural choice for the receptive field estimation of visual ganglion cells. In audition, however, the sound pressure waveform is not an appropriate representation of sound for auditory neurons that do not phase lock to the sound pressure waveform (Aertsen and Johannesma, 1981; Eggermont et al., 1983a). Since linear transformations of a stimulus cannot improve the performance of a linear model, one must make a nonlinear transformation on the sound pressure waveform to be able to estimate a dynamic receptive field to such an auditory neuron. After this nonlinear transformation, receptive fields cannot be said to be linear without the caveat that the way the stimulus is represented is itself a nonlinear function of the stimulus in its most elementary form. The performance of a linear model can be influenced by the choice of nonlinear stimulus representation. In this paper we will examine the effects of choosing different representations of sound on the predictive power of linear neuron models.

Historically, synthetic sound stimuli that varied along a few parameters were used for the characterization of dynamic non-linear responses of higher level auditory neurons. For example, Phillips and Hall (1987) examined cells' responses to sinusoidally amplitude modulated (SAM) tones. In that study, two quantities were modified: the carrier frequency and the frequency of amplitude modulation. Schreiner and Calhoun (1994) and Calhoun and Schreiner (1998) recorded the output of cells reacting to sounds with spectral modulation (i.e. sounds with intensities constant in time but periodic in frequency). In both these cases, the simplest way to characterize these synthetic sounds is by the parameters used to generate them. Usually, these representations have a nonlinear relationship with the sound pressure waveform they generate and, if the space of parameters is small enough, it is possible to map out a cell's response characteristics without the need to use a linear model.

Since the natural auditory world of most organisms is not restricted to sounds fully characterized by a few parameters, there have been many studies on auditory cells' responses to more complex sounds (Eggermont et al., 1983b; deCharms et al., 1998; Theunissen et al., 2000; Depireux et al., 2001; Escabi and Schreiner, 2002). While synthetic sounds can be represented by the parameters which generated them (such as in Depireux et al., 2001), it is common to represent complex sounds (i.e. sounds that have no simple parameterization) with a time-frequency representation.

Time-frequency representations

Sounds can be described as pressure as a function of time, called the sound pressure waveform. In many cases, the perception of sound as well as the physiological responses of higher level auditory neurons are not best described by rapid (for humans up to around 20 kHz) changes in air pressure. A more ethological way of characterizing sound perception is to describe the time-varying spectral content of the perceived sound. These types of representations are called time-frequency representations, because they describe the spectral content of a sound as a function of time. Similarly, in the auditory system, the transformation between sound pressure waveform and time-frequency representation starts in the cochlea, where a few thousand frequency-selective primary auditory cells encode sound intensity in their frequency range. We expect higher cells (whose activities are influenced by the outputs of these primary cells) to be most easily described in terms of the time-frequency representations like those generated by the cochlea, rather than in terms of the sound pressure waveform.

Spectro-temporal receptive fields

While it is possible to map out a cell's response to a small-dimensional family of sounds (e.g. carrier and modulation frequency in SAM stimuli) through an exhaustive search, one cannot determine the stimulus-response map for more complicated sounds (such as natural sounds) by presenting every conceivable complex sound because of complex sounds' high dimensionality. One can, however, find the best linear approximation of a cell's stimulus-response function for a group of sounds. This linear approximation allows for distinct preferences to different frequencies at different latencies, and is therefore called the linear Spectro-Temporal Receptive Field (STRF) (Aertsen and Johannesma, 1981; Eggermont et al., 1983a; Klein et al., 2000; Theunissen et al., 2000; Depireux et al., 2001; Escabi and Schreiner, 2002). The STRF is a linear filter which, when convolved with the stimulus (in its time-frequency representation), predicts the neural response. The STRF also has an unambiguous meaning for cells which are intrinsically nonlinear. If the cell's true input-output relationship is characterized by a Volterra expansion (or, equivalently, a Wiener expansion, both are the functional generalizations of Taylor series), the STRF is the sum of the linear parts of all odd-order components of the Volterra expansion (Marmarelis and Marmarelis, 1978; Klein et al., 2000). Since the linear part of the odd-order components of the Volterra expansion depends on the stimulus, the estimated STRF also depends on the stimulus unless the cell has no odd-order nonlinearities.

For a linear STRF to describe the firing behaviour of a neuron, the following three properties are desirable. First, the response to the stimulus should be phase-locked to the stimulus and to nothing else. Second, the relationship between the firing rate and the convolution of a filter with the sound intensity should be linear. Third, sound intensity at a given frequency and latency should not modulate the gain of the STRF at different frequencies and latencies. Although the third source of nonlinearity cannot be implemented in general time-frequency representations, the second issue can be addressed by the following four considerations. First, auditory neurons' activity may be more amenable to modelling if the center frequencies of filters used in the time-frequency representation are spaced logarithmically (as in a wavelet representation) as opposed to linearly (as in spectrograms). Second, time frequency representations with filters with the incorrect bandwidths will mistake spectral sound modulations for temporal modulations or *vice versa*. Third, if cells use logarithmic or power law amplitude compression (in that their firing rate is proportional to the log of the intensity of sound or to the intensity of sound raised to a power less than 1), a time-frequency representation using a linear amplitude scale will perform more poorly. Fourth, if neurons adapt or encode the activity of adapting neurons, time-frequency representations of sound with built-in adaptive gain control will lead to STRFs with more predictive power than those without.

There are two ways in which comparing the performance of STRFs to different time-frequency representations furthers understanding of audition. First, it gives a better understanding of the computational mechanisms behind the stimulus-response function of an auditory cell. Second, it may give an insight into which kinds of time-frequency representations most closely match the internal representation of sound in the organism under study.

Methods

There are many STRF estimation techniques available (Willmore and Smyth, 2003). In this paper, we use a generalized reverse correlation technique that can estimate the stimulus-response transfer function of high-level sensory neurons using stimuli with non-stationary statistics, such as those found in natural sounds (Theunissen et al., 2001). We apply two different validation methods to measure the estimated spectro-temporal receptive fields (STRF) goodness of fit: a smoothed correlation and a measure of the mutual information between a STRF's prediction and a validation PSTH (Hsu et al., 2004a). The algorithms used here can be found in the STRFPAK software package (http://strfpak.berkeley.edu).

Mathematically, the STRF is defined as a spectro-temporal impulse response, $h(t, x)$ (where t is time and x is a spatial dimension, in this case frequency), that relates a spectro-temporal description of a stimulus, $s(t, x)$, to a time-varying neural response $r(t)$. If the neuronal system and the stimulus are assumed to be stationary, a linear estimate of the stimulus-response transfer function is obtained by:

$$\hat{r}(t) = \int \int h(\tau, x)\, s\,(t - \tau, x)\, d\tau\, dx + \bar{r}.$$

Here $h(t, x)$ is the STRF of the neuron; $\bar{r}$ is the mean neural response and $\hat{r}$ is the estimate of the actual neural response. The analytical solution of $h(t, x)$ which minimizes $\langle(\hat{r} - r)^2\rangle$ (angle brackets indicate taking the expectation value) is:

$$h = C_{ss}^{-1} C_{sr},$$

where C_{ss} is the auto-correlation of the stimulus s, and C_{sr} is the cross-correlation of the stimulus and the neural response. The inversion is performed using the pseudo-inverse of the auto-correlation matrix. This method is required because, when natural sounds are used, only a subset of the acoustical space is sampled. The pseudo-inverse also serves as a form of regularization that limits the number of parameters being fitted (Theunissen et al., 2000;, 2001).

In order to judge the quality of the estimated STRF, Hsu et al. (2004) provides two validation methods to measure the goodness of fit. The first measure is the correlation coefficient between the neural response and predicted neural response in time domain. It is calculated as follows:

$$\mathrm{CC} = \frac{\langle (r\,(t) - \bar{r})\,(\hat{r}\,(t) - \bar{\hat{r}})\rangle}{\sqrt{\langle (r\,(t) - \bar{r})^2\rangle\ \langle (\hat{r}\,(t) - \bar{\hat{r}})^2\rangle}}$$

The actual time varying rate $r(t)$ is not known but can be estimated from the spike trains convolved with a smoothing function. Since the calculation of the CC depends on the smoothing used to obtain $r(t)$ from the spike trains, we calculated the CC for a wide range of smoothing windows. One can then compute CCs either with a fixed window or the window which yields the highest CC. If one wishes to compare predictions of different cells, one should use the same smoothing window to estimate $r(t)$ for all cells, otherwise the CC does not properly track the performance of the STRF across cells. Here, we have chosen a Hanning window of width 21 ms to be the standard smoothing.

The second measure is the coherence between the neural response and predicted neural response in the frequency domain. The coherence is a function of frequency, ω, and is given by:

$$\gamma^2\,(\omega) = \frac{\langle R(\omega)\,\hat{R}^*(\omega)\rangle\,\langle R^*(\omega)\,\hat{R}(\omega)\rangle}{\langle R(\omega)\,R^*(\omega)\rangle\,\langle \hat{R}(\omega)\,\hat{R}^*\,(\omega)\rangle},$$

where capital letters $\hat{R}$ and R are used to indicate the Fourier transform of the estimated and actual responses respectively.

An overall goodness-of-fit estimate from the coherence is calculated by the following integration:

$$I = -\int_0^\infty \log_2(1 - \gamma^2\,(\omega))d\omega$$

Here I is expressed in bits per second, and is an estimate of the mutual information rate between r and $\hat{r}$. The estimation is accurate if the signals involved have Gaussian distributions. In the general case I can be taken as a measure of the integrated (or mean) coherence over frequencies. We will denote the validation measure the coherence-information validation.

Because neural data are inherently noisy, even a perfect model will have a CC and coherence that is less than one. To address the variability of noise across neurons, we calculate the coherence and correlation between our data with even-numbered trials and our data with odd-numbered trials to estimate how well an ideal model for the cell's activity could perform. One can then normalize the performance of the STRF by how well an ideal model might perform given the intrinsic noise in the neuron. The normalized CC will be called the CC ratio hereafter. This method is described in detail in Hsu et al. (2004).

Of the two methods, the coherence-information method is a better metric for prediction because it does not require the choice of a smoothing window (which can be somewhat arbitrary), and because prediction errors at one frequency do not mask well-predicted features at other frequencies, as is the case with the CC ratio. The coherence-information method does require the specification of a window length for its Fourier transforms, but as long as this window is longer than the longest expected PSTH features the length of this window is irrelevant. We chose a window length of 128 ms. Predictions will also be validated using CC ratios because of their widespread use.

Time-frequency representations

We varied four aspects of our time frequency representations: the spacing scheme and corresponding filter shape for our filter bank, the time-frequency scale, stimulus amplitude compression and adaptive gain control. The first aspect is the choice of logarithmic or linear filter frequency spacing. We contrasted sound decompositions that would result from filters with fixed frequency bandwidth, which are used extensively in sound spectrograms, to those that would result from filters with fixed octave bandwidth, which are better representations of the filtering occurring at the vertebrate cochlea (Von Békésy, 1960). Second, for both the fixed frequency bandwidth and the fixed octave bandwidth, we examined the effect of the scale of the time-frequency representation by using different filter bandwidths (in linear or octave units, respectively). The third aspect is the compression on the amplitude scale. We compared a linear amplitude scale (represented intensity proportional to stimulus power in Watts per square meter) to logarithmic and power law amplitude scales, which are more biologically realistic (Ruggero, 1992). The fourth aspect is adaptive gain control. In this study, we chose to use the adaptive gain control that was an intrinsic part of Richard Lyon's cochlear model (Lyon, 1982) as implemented by Malcolm Slaney (1988).

Spectrograms

Perhaps the most straightforward time-frequency representation of sound is the spectrogram. To generate a spectrogram from a sound pressure waveform, one creates a bank of band-passed filters whose impulse response consists of a carrier frequency (the center frequency of the filter) modulated by a short-time window. The Fourier transform of this time window is the frequency gain curve of the band-passed filter. Thus the duration of the window determines the temporal spread and spectral bandwidth of the filtered signal components (Cohen, 1995; Singh and Theunissen, 2003). In a spectrographic representation, the temporal duration of the filters is the same for all center frequencies. As a consequence, the spectral bandwidth is fixed in a linear frequency scale. Moreover, to uniformly tile the time-frequency space, the center frequencies are equally spaced. Thus, the duration of the short-time window modulating the carrier frequency, the spectral bandwidth, and the spacing of the center frequencies are interdependent and define the time-frequency scale of the spectrogram.

The experimenter chooses the time-frequency scale of a spectrogram. An inappropriate choice would result in what should be considered to be a spectral modulation to be represented as a temporal modulation, or vice versa. In theory, one could use a representation based on the Wigner distribution, which is scale independent (Klein et al., 2000; Theunissen et al., 2000). In practice, however, the mathematical complexities of the Wigner distribution make both the estimation and the interpretation of such STRFs difficult (Cohen, 1995). For those reasons, spectrograms or wavelets are used and the obtained STRF can be interpreted as a filtered version of the STRF that would be obtained from the Wigner distribution (Theunissen et al., 2000).

In an earlier work (Singh and Theunissen, 2003), it was discovered that setting filter carrier frequency and bandwidth at 125 Hz was a good choice, in that with these spectrograms there were few temporal modulations close to being interpreted as spectral modulations and vice versa. In this study we used filter banks with 62.5 Hz, 125 Hz and 250 Hz spacing to study the effect of changing the time-frequency scale around this "standard" of 125 Hz. The spectrogram with 62.5 Hz filter spacing will capture the most spectral features but the fewest temporal features, and the spectrogram with 250 Hz filter spacing will capture temporal features well at the expense of spectral resolution.

The shape of the time window and of the filter gain functions was chosen to be a Gaussian. The Gaussian window has the advantage of being symmetric in time and frequency and leads to well defined measure of temporal and spectral resolution (Singh and Theunissen, 2003).

Two other parameters important to the selection of the filter bank to be used for the generation of a spectrogram are the highest and lowest center frequencies to be used. It is desirable to include the full range of frequencies thought to be audible by the organism, but including more frequencies might lead to spurious correlations in the STRF estimations, which are likely to degrade STRF performance. We use a frequency range of 250 Hz to 8000 Hz which covers the audible frequency range for zebra finches (Okanoya and

Dooling, 1987; Zevin et al., 2004) and most song birds (Dooling, 1982).

Wavelets

Unlike successive filters in a spectrogram, the center frequencies of consecutive auditory filters on the mammalian or avian cochlea are not spaced linearly. They are approximately spaced logarithmically, like consecutive keys on a piano (Von Békésy, 1960). Correspondingly the bandwidth of the auditory filters measured by the physiological responses of the primary auditory nerve fibers to pure tones is approximately constant when measured in octave units (Ruggero, 1992; Gleich and Manley, 2000). If auditory processing in vertebrates is most naturally represented by a series of filter banks with logarithmically-spaced center frequencies and fixed octave bandwidths, we might expect this type of time-frequency representation to lead to better STRFs.

Here we will explore the ability of a time-frequency analysis called a Morelet continuous wavelet transform (CWT) to produce a good STRF (Mallat, 1989). Wavelets have successfully been applied for to audio compression (Painter and Spanias, 2000) and to a more limited extend to auditory research (Lewicki, 2002; Yao and Zhang, 2002). We use a Morelet wavelet filter bank instead of a potentially more biologically inspired decomposition (such as a Gammatone filter bank) because the time-frequency scale of the Morelet transform is easily parameterized.

As well as the top and bottom filter center frequency, one must choose the time-frequency scale of the Morelet filters to be used. The most conventional way to express the spacing of filters is to relate the number of filters used per octave. Morelet center frequencies are spaced logarithmically, so there will be on average a constant number of filters for every doubling of center frequency, called an octave. The full frequency range of interest spans five octaves (250–8000 Hz). We examine representations with 4, 8 and 16 filters per octave to determine which parameters lead to the best STRFs.

Linear, log and power law amplitude scales

Perception of sound intensity correlates with a compressive function of the power of the stimulus rather than the raw stimulus power. Although the psychoacoustical literature on the subject of sound loudness is complex, it is generally accepted that a power law with a coefficient around 0.6 relates loudness and sound pressure amplitude (Stevens, 1956). This power law is not as compressive as the log function but it will "flatten out" the distribution of amplitude. Similarly a compressive non-linearity is a common property of the auditory system found both at the level of the basilar membrane (Ruggero, 1992; Schlauch et al., 1998) and in many auditory neurons (Sachs and Abbas, 1974; Palmer and Evans, 1982; Phillips, 1990; Woolley and Casseday, 2004). It is also known that stimuli with an even log intensity distribution are better at driving neurons than stimuli with a flat intensity distribution (Escabi et al., 2003), suggesting that the logarithmic amplitude scale might be the most appropriate for sensory coding too. Here we will compare the predictive abilities of the STRF based on the linear representation of the amplitude envelope of the signals obtained from each filter in the filter bank to the STRF based on a logarithmic or power law representation of these amplitude envelopes.

Lyon's Cochlear model

If it is one's objective to decompose sound into a time-frequency representation in as similar a way to the cochlea as possible, one can improve on a Morelet filter bank by using a biologically inspired model of the first states of auditory processing. We used the relatively simple algorithm developed by Richard Lyon (1982). which has been implemented in a Matlab toolbox by Malcolm Slaney (1988; The toolbox is available on the web at http://rvl4.ecn.purdue.edu/ ∼malcolm/interval/1998-010/). Lyon's model includes not only the approximate logarithmic spacing of filter center frequencies (log at high frequencies and more linear at low frequencies), but also the rectification performed by inner hair cells and, optionally, adaptive gain control. Lyon's model also mimics the cochlear delay caused by the traveling of the sound as vibrations from the oval window up the vestibular canal. Thus the exact frequency profile of each Lyon's model filter is more biologically accurate than either spectrographic or wavelet representations.

Of particular interest is the ability of Lyon's model to implement adaptive gain control (AGC). AGC simulates the whole ear's desire to keep cochlear outputs at a set point. Briefly, AGC works by comparing a desired set point to the short-time-averaged mean of the output of a model channel added to the output of its nearest neighbours. The gain of this channel is modulated to push this average closer to the set point. Including the channel's neighbours results in spectral feature sharpening, and comparing outputs to the time-averaged mean results in temporal sharpening. The time-averaging function is the sum of four decaying exponentials with different time constants; the four stage adaptive gain mechanism loosely models physiological mechanisms that are known to change the gain of intensity sensitivity measured physiologically by factors of up to 100. Included in these physiological mechanisms are the electro-mechanical system of the outer hair cells, the middle ear reflex and depletion of neuro-transmitters in the inner hair cells. The four stages are implemented by using four time scales for the integration step ranging from 10 ms to 640 ms (Slaney, 1988).

We used four implementations of Lyon's model. The original Lyon model has filters bandwidths with a default Q factor (ratio of center frequency to bandwidth) of 8 for high frequency neurons to approximately match mammalian neuro-physiological data (Ruggero, 1992). We refer to that original model as the "original Lyon model". To compare Lyon's model with our best wavelet representation, we also tested a modified version of the original Lyon model that had a time-frequency scale with a Q factor of 4. This is the time-frequency scale that was best suited to describe the auditory response of avian midbrain and forebrain auditory neurons in our data set with the wavelet transformation. We also set the center-frequency spacing of the filters to be half of the filter bandwidth instead of the default spacing, which is one quarter of the filter bandwidth. We refer to this modified Lyon model as the Bird Lyon model because its parameters were coarsely optimized to obtain the best fits for avian auditory neurons. For both models, we ran simulations with and without AGC.

Lyon's model should mimic cochlear outputs better than any of the time-frequency outputs investigated so far. One might then expect Lyon's model to out-perform the other time-frequency representations.

Stimuli and neurophysiology

Two classes of sounds were used: songs from 20 different adult male zebra finches and 10 samples of synthetic noise that was limited in spectral and temporal modulations - modulation-limited noise (ml-noise; see below). Two-second samples of song from twenty zebra finches were recorded in our laboratory using standard procedures: adult male zebra finches are placed in an acoustically isolated sound recording booth. The sound inside the booth is monitored and saved to a computer when the 4 kHz component of a short time power spectrum crosses a used defined threshold. Samples or recorded sound are then screened for clean renditions of the bird's song. We found that 20 songs from 20 different individuals were sufficient to characterize the spectral and temporal modulations that typify zebra finch song (Singh and Theunissen, 2003). Before using the songs in our neurophysiological recordings, they were band-pass frequency filtered at 250 and 8000 Hz.

To generate the ml noise, we calculated the spectro-temporal log envelope function of the sound (i.e. its spectrogram in logarithmic units) as a sum of ripple sounds. Ripples are broadband sounds that are the auditory equivalents of sinusoidal gratings typically used in used in vision research. This sum of ripples can be written as:

$$S(t,x)=\sum_{i=1}^{N}\cos(2\pi\omega_{t,i}t+2\pi\omega_{x,i}x+\varphi_i)$$

where ϕ_{i} is the modulation phase and $\omega_{t,i}$ and $\omega_{x,i}$ are the spectral and temporal frequency modulations for the ith ripple component and $S\,(t,x)$ is the zero mean log envelope of the frequency band x. We used $N = 100$ ripples. The range of spectral and temporal modulation frequencies, $\omega_{x,i}$ and $\omega_{\mathrm{t,i}}$, was chosen to cover most of the range of spectral and temporal modulations found in zebra finch song: modulation frequencies were sampled randomly from a uniform distribution of modulations bounded by 50 Hz (temporal) and 2 cycles per kHz (spectral). The frequency of the carrier was sampled uniformly between 250 and 8000 Hz. The modulation phase was random, taken from a uniform distribution.

The actual envelope used to generate the sounds was then obtained from $S(t,x)$ by adding the DC level and modulation depth obtained from the log envelope of song. The envelope is written as:

$$S_{\mathrm{Norm}}(t,x)=A\,(x_{\mathrm{Peak}})+\frac{\langle\sigma\;(x)\rangle_x}{\sigma_S\,(x)}S\,(t,x)\,,$$

where: $A\,(x_{Peak})$ is the DC level of the log amplitude envelope at the frequency where it is the largest in song; $\sigma(x)$ is the standard deviation across time of the log envelope for a frequency, x, also calculated for song; and $\sigma_S(x)$ is the standard deviation obtained from the generated amplitudes $S\,(t,x)$. Therefore, the mean amplitude (and thus intensity) in each frequency band was constant and given by the peak of the log amplitude envelope in song; ml noise had a flat power spectrum between 250 and 8000 Hz with levels that matched the peak of the power spectrum of song found at $x_{\mathrm{Peak}} \sim 3.5$ kHz. The modulation depth (in log units) in each frequency band was set to the average modulation depth found in song across all frequency bands. The sound pressure waveform was then obtained by taking the exponential of the normalized log amplitude envelope, $S_{\mathrm{Norm}}\,(t,x)$, and using a spectrographic inversion routine (Singh and Theunissen, 2003). Peak power was balanced between song and ml-noise stimuli. Stimuli were presented at 70 dB SPL (peak).

In vivo electrophysiological recordings were obtained using standard extracellular recording procedures from urethane-anesthetized adult male zebra finches. Stimuli were presented free-field in a sound-attenuation chamber. The head-related transfer function (HRTF) was not taken into account, so our STRF estimates include the contribution of the HRTF. The speaker was placed 15 cm in front of the animal. The frequency response of the sound presentation, as measured by a microphone placed at the same location as the animal, was approximately flat between 500 Hz and 10 kHz (deviations within ± 5 dB). Ten trials of 20 songs and 10

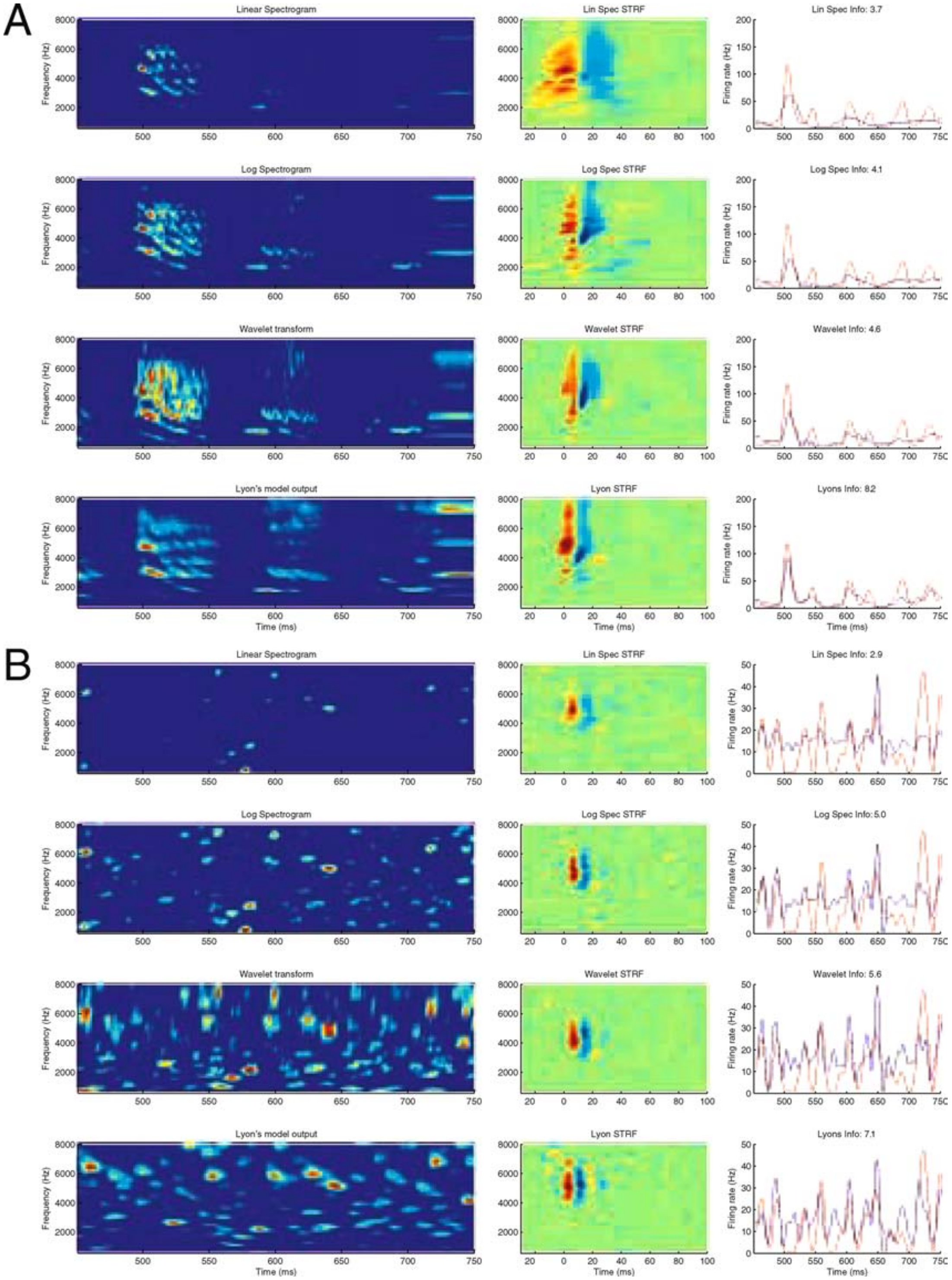

Fig. 1 Stimuli, STRFs and predictions using four time-frequency representations. (A) STRFs generated using different time-frequency representations and samples of their predictions. The left column contains different time-frequency representations of the same segment of conspecific song, the middle column shows the corresponding STRF and the right column shows the STRF's prediction to the stimulus segment and the goodness of fit of the STRF's prediction (expressed in bits per second). Note that the left column is only a sample from one stimulus and 20 stimuli were used to characterize the STRF. The time-frequency representations shown are, from top to bottom, a spectrogram with 125 Hz filter spacing, a spectrogram with 125 Hz filter spacing using logarithmic amplitude compression, a Morelet filter bank with 8 filters per octave, and the default Lyon model using adaptive gain control. The choice of time frequency representation affects the qualitative shape of the STRF to some degree, as well as the goodness of fit. (B) Columns and rows correspond exactly to those in (A) except that in (B) the stimulus used is modulation-limited noise

epochs of modulation-limited noise were presented to each cell. Single neuron responses from the avian auditory midbrain, the mesencephalicus lateral dorsalis (MLd), primary forebrain (Field L), and the secondary auditory forebrain region caudal mesopallium (CM) that has been implicated in the perception of learned sounds (Gentner and Margoliash, 2003) were examined. Spikes were collected from up to two brain regions simultaneously, at a sampling resolution of 1 ms. Spike arrival times from single neurons were obtained using a window discriminator. Responses from 91 MLd cells, 142 L cells and 35 CM cells were obtained. Recording locations were confirmed by identifying electrolytic lesions using standard histological procedures.

Figure 1 shows an exemplar for each of the two stimulus ensembles represented as a linear and a log amplitude spectrogram with 125 Hz filter spacing, an eight cycle-per-octave wavelet transform and Lyon's cochlear model with adaptive gain control. Because the modulation depth of ml-noise is matched to that of song (measured on a log scale), the sound appears sparse in a linear spectrogram. On the same figure the central panels show the STRF obtained from one example neuron from MLd. The right panels show an example of the prediction for that example stimulus obtained from the STRF and compared to the actual response (PSTH).

Results

To test the effect of the nature of the stimulus representation, the time-frequency scale, compression and adaptive gain control, we compared the ability of 22 time-frequency representations (three amplitude scales of three resolutions of log spectrogram, three amplitude scales of three resolutions of log wavelets, and four Lyon models) to generate a STRF with the best predictive power. All these representations were performed for both types of stimuli (conspecific song and modulation-limited noise). Thus, for each of the 268 neurons in our database, we generated $22 * 2 = 44$ STRFs, and compared their predictive power.

To exhaustively compare each STRF method, we would thus need $43 * (44/2) = 967$ comparisons of performance. For statistical integrity and in order to synthesize the results, we analyzed the results by initially only performing specific planned comparisons for groups of STRFs: we first compared amplitude compression schemes and wavelets to spectrograms. All statistical tests are Wilcoxon signed rank tests on the difference of CC ratios if they are used, or on the difference of information rates if the coherence method is used.

Log amplitude compression is better than linear or power law

For almost all neurons, we found that using a log compression on the amplitude envelope resulted in an improved performance. This effect is shown in Fig. 2 where we compare the predictive power using the log scale representation versus the linear and power law scales. Plotted together are results for both spectrograms and wavelets from all time-frequency scales. A paired Wilcoxon signed rank test on the ratio of performance with log and linear scales showed that for all brain areas, the log compression outperformed both linear and power law scales, $p < 10^{-7}$ in every individual brain area, regardless of whether the information coherence method or the CC ratio was used to validate predictions.

The average across all areas and methods of the information predicted was 2.2 bits/s for linear scales, 3.8 bits/s for power law scales and 4.4 bits/s for log scales. The percent increase (difference over mean) of log over power is 15%, and the percent increase of log over linear is 66%. The average across all areas and methods of the CC ratio was .255 for linear scales, .350 for power scales and .373 for log scales. The percent increase (difference over mean) of log over power is 6.5%, and the percent increase of log over linear is 38%. Given the strength of this effect and the scarceness of cells which are described best by other amplitude compression scales, we will consider only log amplitude representations for spectrograms and wavelets in the rest of the results section.

Spectrograms vs. wavelets

We compared the performance of the STRF predictions for a spectrographic versus a wavelet decomposition of the sound. For this comparison, for each neuron and stimulus type, we pitted the performance of the best of the three spectrographic representations against the performance of the best of the three wavelet transforms (see the methods section for more detail). The resultant scatter plot is shown in Fig. 3. Overall, the comparison is roughly diagonal, although when modulation-limited noise was used, a significant preference for spectrograms is evident: mean increases in predicted information from 5.75 bits/s to 8.06 bits/s (2.31 bits/s or 33%) in Mld, from 2.77 bits/s to 3.96 bits/s (1.19 bits/s or 35%) in Field L and from 1.68 bits/s to 2.65 bits/s (0.97 bits/s or 45%) in CM, all $p < 10^{-6}$. The change in CC ratio is of the same direction: CC ratios increase from .327 to .400 (.073 or 20%) in MLd, from .301 to .366 (.065 or 20%) in Field L and from .218 to .293 (.075 or 29%), all $p < 10^{-5}$. When conspecific song is used to generate STRFs, there is a slight preference for wavelet transforms over spectrograms in Field L and CM (mean increases of predicted information from 2.96 bits/sec

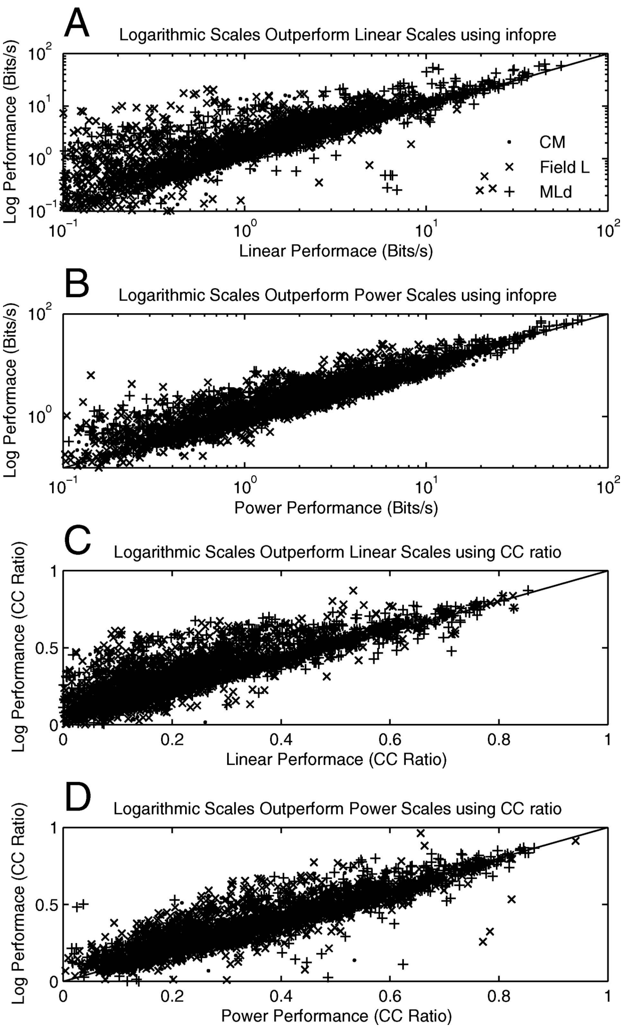

Fig. 2 Comparison of amplitude compression schemes. (A) Scatter plot of the performance of STRFs using logarithmic versus linear scales. The performance was measured with the coherence-information validation method (Infopre on the figure). Each cell is plotted 12 times here, since there are two stimulus types and six time-frequency choices (three scales of spectrograms and three scales of wavelets). B) Scatter plot like (A), but comparing the performance of log amplitude scales to power law amplitude scales. (C) and (D) are like (A) and (B) but show CC ratios. In (A, B, C and D), in all stimulus types, brain areas, logarithmic scales are preferred, with $p < 10^{-7}$

to 3.32 bits/sec (.36 bits/s or 12%) and from 2.39 bits/sec to 2.73 bits/sec (.34 bits/s or 13%), both $p < .05$, and mean CC ratio increases from .393 to .408 (.015 or 3.8%) and from .328 to .335 (.0078 or 2.3%), both $p < .02$), while MLd still favours spectrograms (mean predicted information improvements from 6.63 bits/sec to 6.71 bits/sec (.081 bits/sec or 1.2%), $p < .005$, mean CC ratio improvement from .487 to .511 (.023 or 4.7%), $p < .05$).

All points in Fig. 3 lie close to the diagonal or in favour of spectrograms; i.e. there are some conditions which substantially favour a spectrographic representation, but no conditions we investigated which substantially favour a wavelet representation.

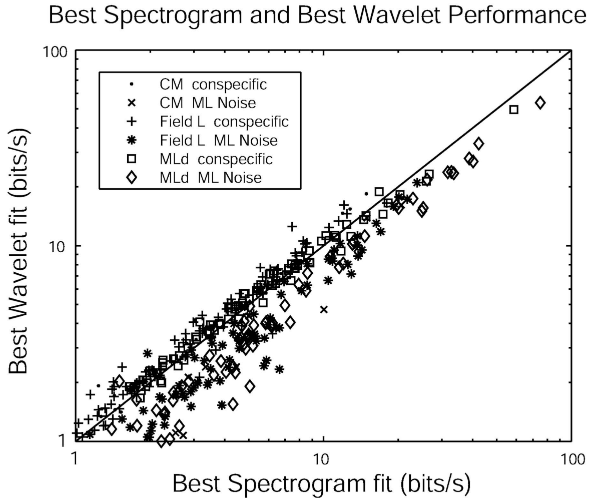

Fig. 3 Scatter plot of the performance of STRFs using the best wavelet versus the best spectrogram, quantified using the predicted information measure. The plot shows that overall there is no strong advantage to using a wavelet transform for any of the stimuli in any of the areas investigated, while spectrogram-based STRFs out-perform wavelet-based STRFs when modulation-limited noise is used

Best time-frequency scale

Three time-frequency scales were used for both the spectrograms and the wavelets. The time-frequency scale of the representation determines the temporal and spectral resolution with which the structure in the stimulus can be described. In all time-frequency representations, there is a trade off between the temporal and the spectral resolution. The optimal time-frequency scale will depend on the nature of the sound and on the response properties of the neurons.

Recall from the methods section the time-frequency scales of the various representations: the temporal, intermediate and spectral spectrograms had filter bandwidths of 250, 125 and 62.5 Hz; the temporal, intermediate and spectral wavelet transforms had 4, 8 and 16 filters per octave, and the Bird Lyon model and the original Lyon model had Q factors of 4 and 8 respectively.

Conspecific song

With conspecific song as the stimulus, in all cases the most temporal time-frequency representation significantly out-performed the most spectral time-frequency representation using the information-based validation method. When CC ratios are used, results are mixed because the PSTH smoothing done for the correlation coefficient (a 21 ms Hanning window, see Methods) obscures highly temporal features of predictions. Mean improvements and their significance levels using both validation methods are shown in Table 1.

Table 1 Improvements of most temporal over most spectral time-frequency scales for conspecific song, broken down by time-frequency representation class and brain area

	Spectrograms			Wavelets			Lyon		
	Temp	Spec	Diff	Temp	Spec	Diff	Temp	Spec	Diff
Conspecific, Predicted information									
MLd	6.68	5.40	1.28 21%^{3}	6.36	6.03	.033 5.3%^{3}	9.18	7.53	1.64 20%^{3}
Field L	2.90	2.55	.352 13%^{3}	3.25	2.83	.413 14%^{3}	4.31	3.56	.751 29%^{3}
CM	2.34	2.11	.227 10%^{1}	2.72	2.26	.462 19%^{3}	4.59	3.81	.780 19%^{3}
Conspecific, CC ratio									
MLd	0.469	0.502	−.034 −6.9%^{3}	0.472	0.462	.010 2.1%^{2}	0.525	0.535	−.010 −1.9%^{0}
Field L	0.371	0.376	−.005 −1.3%^{0}	0.396	0.380	.017 4.3%^{3}	0.439	0.427	.013 2.9%^{1}
CM	0.306	0.316	−.011 −3.5%^{0}	0.328	0.310	.017 5.4%^{2}	0.380	0.375	.004 1.2%^{0}

Entries contain the mean prediction strength (predicted information in bits/s above, CC ratio below) of the most temporal and most spectral spectrograms, wavelets and Lyon's models (with AGC), followed by the mean change (temporal minus spectral) and percent change (difference divided by mean). Superscripts indicate significance of the difference: $0 =$ not significant, $p > .05$, $1 = p < .05$, $2 = p < .001$, $3 = p < 10^{-5}$

Table 2 Improvements of most temporal over most spectral time-frequency scales for ml-noise, broken down by time-frequency representation class and brain area

	Spectrograms			Wavelets			Lyon		
	Temp	Spec	Diff	Temp	Spec	Diff	Temp	Spec	Diff
Ml-noise, Predicted information									
MLd	8.49	6.74	1.74 23%^{3}	6.65	8.06	−1.42 −19%^{3}	8.97	8.60	.371 4.2%^{1}
Field L	3.53	3.80	−.268 −7.3%^{1}	3.66	4.08	−.423 −11%^{3}	3.79	4.15	−.357 −9.0%^{3}
CM	2.59	2.38	.211 8.5%^{0}	2.59	2.86	−.296 −9.6%^{1}	2.86	2.92	−.058 −2.0%^{1}
Ml-noise, CC ratio									
MLd	0.380	0.392	−.012 −3.1%^{2}	0.370	0.389	−.020 −5.2%^{3}	0.432	0.445	−.013 − 2.9%^{1}
Field L	0.336	0.375	−.039 − 11%^{3}	0.342	0.373	−.031 − 8.6%^{3}	0.376	0.425	−.049 − 12%^{3}
CM	0.267	0.300	−.033 − 12%^{2}	0.261	0.285	−.024 − 8.9%^{1}	0.295	0.328	−.033 − 11%^{2}

Entries contain the mean prediction strength (predicted information in bits/s above, CC ratio below) of the most temporal and most spectral spectrograms, wavelets and Lyon's models (with AGC), followed by the mean change (temporal minus spectral) and percent change (difference divided by mean). Superscripts indicate significance of the difference: $0 =$ not significant, $p > .05$, $1 = p < .05$, $2 = p < .001$, $3 = p < 10^{-5}$

Modulation-limited noise

With modulation-limited noise as the stimulus, the best time-frequency scale depends more on the class of time-frequency representation. There is no clear pattern, suggesting auditory processing is less temporal when ml-noise is played.

Lyon's model and adaptive gain control

We tested whether Lyon's cochlear model, a more biologically inspired time-frequency representation than spectrograms and wavelets, improves STRF predictions and, in particular, whether including a form of adaptive gain control (AGC) matters. Lyon's model includes a biologically inspired model of gain control as described in the methods. STRFs to spectrograms and wavelet transforms typically fail to predict high-amplitude onset responses to sound after a period of silence (see Fig. 1A). With AGC, stimulus features with rapid onsets after a period of silence are emphasized, so perhaps a STRF to such a pre-emphasized stimulus will capture more onset behaviour. We tested both the Bird and the original Lyon models (see the methods section) with and without gain control. For all stimuli and in all regions, in terms of both the information coherence and the CC ratio, we found no time-frequency representation outperforms the Bird Lyon model.

Conspecific song

We found that the best representation for STRFs to conspecific song was the Bird Lyon model with AGC; comparisons with all other representations are shown in Fig. 4. The closest contender to the Bird Lyon model with AGC is the original Lyon Model with AGC, suggesting that the AGC is mostly responsible for the good performance of the Bird Lyon model with AGC. Using the coherence information metric, the Bird Lyon model with AGC outperformed all others in MLd, Field L and CM by at least 1.65 bits/s, $p < 10^{-8}$, .75 bits/s, $p < 10^{-9}$ and .78 bits/s, $p < .00002$ respectively. Using CC ratios yields statistically insignificant results in MLd and CM (i.e. the Bird Lyon model does not perform significantly bet-

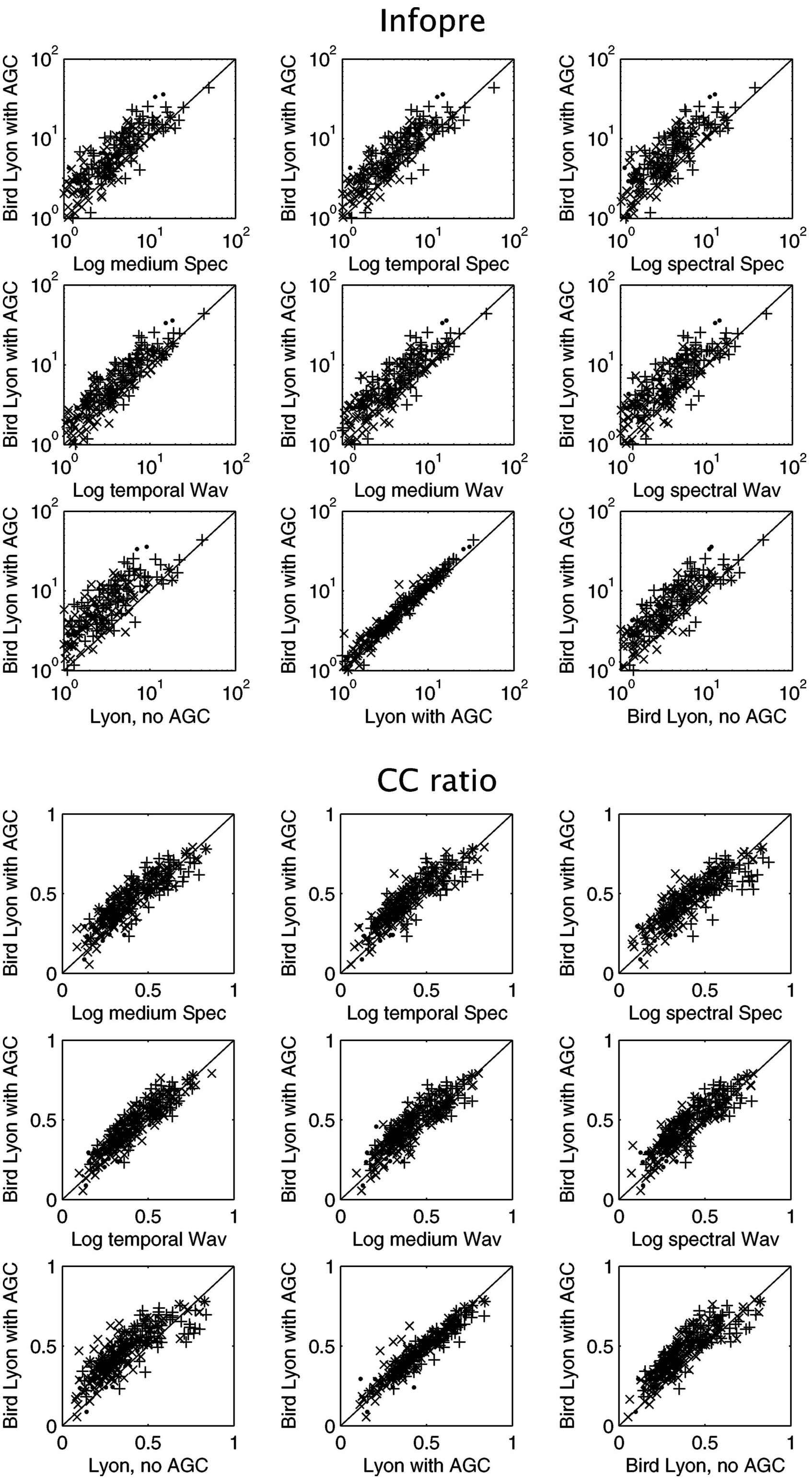

Fig. 4 Comparison of all other methods to the performance of the Bird Lyon representation with adaptive gain control, conspecific song used as the stimulus. The eighteen panels of this figure show scatter plots of the performance of the Bird Lyon model with AGC (y axis) against the performance of the spectrographic representation (top row), the wavelet representation (second row) and the original Lyon model or the Bird Lyon model without AGC (third row). MLd cells are denoted with a " + ", Field L cells with a "x" and CM cells with a "♦". In the top 9 panels the performance is quantified with the predicted information measure (Infopre) and units are in bits/s. In the lower 9 panels, the performance is quantified using the normalized correlation coefficient (CC Ratio). Points above the diagonal indicate the Bird Lyon with AGC representation yields a better-performing STRF. The only close contender is the original Lyon model using AGC, suggesting gain control is the major factor in this representation's success

ter than the regular Lyon model, but outperforms all others nevertheless) while in Field L the result is the same, but statistically weaker: the mean improvement in CC ratio is .013, $p < .02$, compared to its closest contender, the regular Lyon model. To give a feel for the effect size, we also report the increases in performance of the Bird Lyon model over a standard: the log temporal spectrogram. For conspecific song: in MLd, Field L and CM, predicted information rose from 6.68 bits/s to 9.18 bits/s (2.50 bits/s or 32%, $p < 10^{-8}$); from 2.90 bits/s to 4.31 bits/s (1.41 bits/s or 39%, $p < 10^{-8}$) and from 2.34 bits/s to 4.59 bits/s (2.25 bits/s or 65%, $p < 10^{-5}$). Using CC ratios, improvements in MLd, Field L and CM are from .469 to .525 (.056 or 11%, $p < 10^{-5}$); from .371 to .439 (.068 or 17%, $p < 10^{-8}$) and from .306 to .380 (.074 or 22%, $p < 10^{-5}$).

The bottom-center subplot of Fig. 4 is the closest to diagonal, suggesting that adaptive gain control is a key reason why the Bird Lyon model performs so well, and the time-frequency scale adjustment to the regular Lyon's model adds only marginal improvement.

Modulation-limited noise

STRFs generated using modulation-limited noise also had most predictive power when the Bird Lyon model with AGC was used; comparisons to all other stimulus types appear in Fig. 5. With modulation-limited noise, AGC helps somewhat less as there are no long periods of silence, so AGC modulates the amplitude of the time-frequency representation less than it would with conspecific song. Still, the Bird Lyon model yields the best STRFs in all areas no matter which validation metric is used. Mean coherence information improvements over the closest rival representation in MLd, Field L and CM are 1.20 bits/s, p “.0001, $p < 10^{-7}$, .23 bits/s, $p <.0001$, and .42 bits/sec, $p < .002$ respectively. Mean CC ratios are also better when the Bird Lyon model is used: improvements over all other methods in MLd, Field L and CM are at least .043, $p < 10^{-6}$, .018, $p < .0005$, and .037, $p < .0005$, respectively. As with conspecific song, we report the improvement of the Bird Lyon model over the temporal spectrogram. In MLd, Field L and CM, predicted information rose from 6.66 bits/s to 7.85 bits/s (1.19 bits/s or 16%, $p < 10^{-7}$); from 2.30 bits/s to 2.90 bits/s (0.6 bits/s or 13%, $p < 10^{-7}$) and from 1.63 bits/s to 2.24 bits/s (.61 bits/s or 32%, $p < .002$). Using CC ratios, improvements in MLd, Field L and CM are from .338 to .402 (.064 or 17%, $p < 10^{-6}$); from .261 to .323 (.062 or 21%, $p < 10^{-8}$) and from .205 to .267 (.062 or 26%, $p < .0005$).

Overall, when all the different stimulus representations are taken into account, two factors yield the largest and most systematic (across all neurons) effects: the amplitude compression and the adaptive gain control. We show the absolute increase in performance caused by these factors across brain regions and stimulus types in Fig. 6 where we compare the overall performance of STRFs using linear scale spectrograms, log scale spectrograms and the Bird Lyon model with adaptive gain control. Other factors have more subtle effects that depend more on the brain region and nature of the stimulus and are discussed below.

Summary and discussion

Determining the appropriate time-frequency representation is critical for the estimation of valid spectro-temporal receptive fields (STRFs) in audition. Although most auditory researchers using these techniques realize the importance of this step (Klein et al., 2000; Escabi and Schreiner, 2002; Elhilali et al., 2004; Machens et al., 2004), a systematic investigation of appropriate time-frequency representations has not been performed previous to this. We systematically compared a variety of common time-frequency representations for their ability to generate STRFs with high predictive power.

Although spectrograms are commonly used to analyze speech and other natural sounds, wavelet representations are more realistic representations of the transformations that are occurring at the auditory periphery. Similarly, wavelet transformations are also better at explaining certain psychoacoustic results. However, in terms of STRFs of higher-level auditory neurons, we did not find wavelets to give a strong advantage over spectrograms in any circumstance. On the other hand, many cells could only be modelled well with a spectrogram, and not with a wavelet transform. Moreover, wavelet transforms allow higher temporal modulations at higher center frequencies, meaning a preference for high frequency might be conflated with a preference for high temporal modulations. Thus our data indicate choosing a spectrogram over a wavelet representation might be a more prudent choice in the absence of data to the contrary.

As mentioned in the Methods section, for both the wavelet and spectrogram the time-frequency scale as determined by the width of the filters must be chosen. The filtering results in a loss of potential temporal or spectral resolution with a trade-off between them. Here again, one might expect that the optimal time-frequency scale might be determined by the filtering occurring at the periphery. However, we found a more complex picture. The optimal time-frequency scale depends both on the stimulus type and on the brain region being studied.

The best time-frequency representation in MLd was obtained with filter widths that are wider than those obtained from the auditory periphery. Therefore, finer temporal resolution than that obtained by the coding of the amplitude envelope at the auditory periphery can be found at higher levels in the auditory system, suggesting that the coding of the fine structure of the sound performed by auditory ganglion cells via phase-locking of the sound pressure waveform is integrated in the representations of sound by higher order neurons. Field L shows a greater variety of functional cell types, but overall is still best described through STRFs to the Bird Lyon model. AGC still gives a boost in predictions, but not as much as in MLd (see Fig. 6). Even though CM is the furthest away from the periphery, AGC provides a

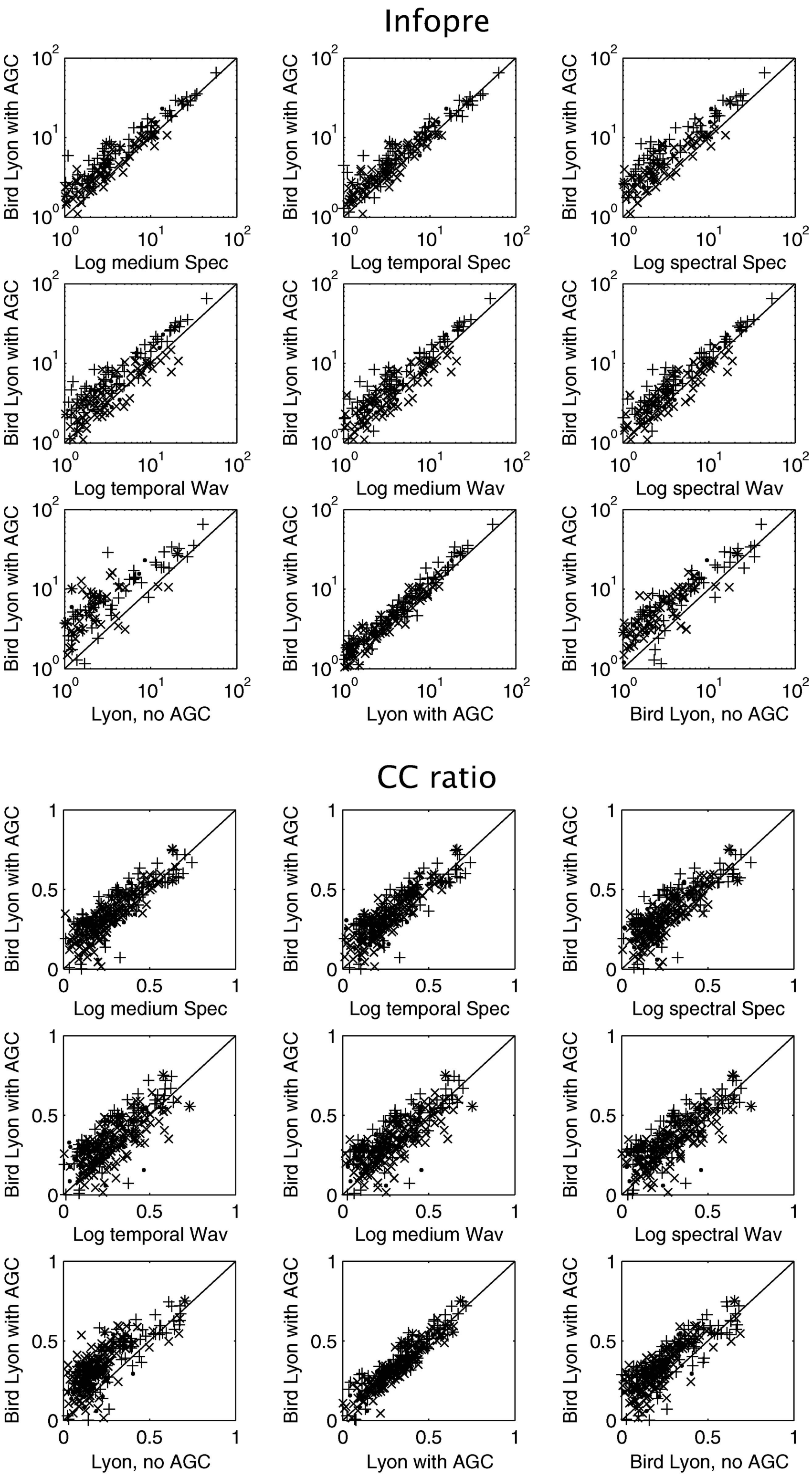

Fig. 5 Comparison of all other methods to the performance of the Bird Lyon representation with adaptive gain control, ml-noise used as the stimulus. The eighteen panels of this figure show scatter plots of the performance of the Bird Lyon model with AGC (*y* axis) against the performance of the spectrographic representation (top row), the wavelet representation (second row) and the original Lyon model or the Bird Lyon model without AGC (third row). MLd cells are denoted with a " + ", Field L cells with a "x" and CM cells with a "♦". In the top 9 panels the performance is quantified with the predicted information measure (Infopre) and units are in bits/s. In the lower 9 panels, the performance is quantified using the normalized correlation coefficient (CC Ratio). Points above the diagonal indicate the Bird Lyon with AGC representation yields a better-performing STRF. The only close contender is the original Lyon model using AGC, suggesting gain control is the major factor in this representation's success

strong boost in predictions. It seems therefore that AGC is not merely a mechanism for dynamic range management, but is a useful tool for higher brain areas too.

The optimal time-frequency scale will depend on the stimulus being used to drive the system. Moreover, this optimal point is not trivially just a factor of the stimulus but also of how the response property of the neurons change when different stimuli are being used (Theunissen et al., 2000). Here, we found that in responses to conspecific song, temporal information is relatively more important than spectral information

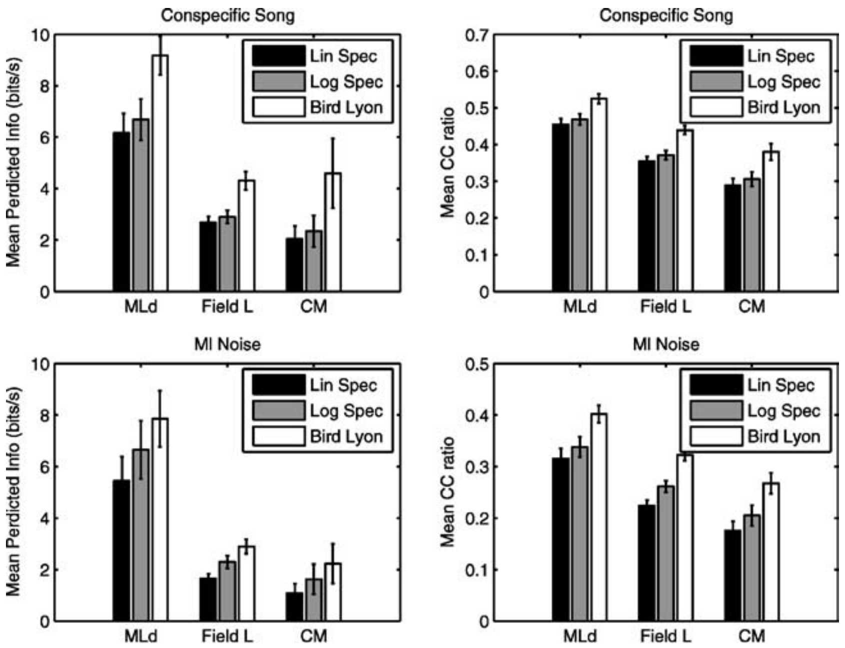

Fig. 6 Summary of the effects of amplitude scales. Shown are the performance of linear spectrograms, log spectrograms and the Bird Lyon model for all areas, both stimulus types and using both the information measure and the CC ratio for validation. All pair-wise differences are significant to at least $p < .0005$. The error bars show one standard error of the mean

as compared with modulation-limited noise. However since the modulation limited noise also included all the high temporal modulation frequencies that were also present in song, we must conclude that this change in optimal time-frequency scale occurred because the stimulus-response function of the neurons was different for each stimulus class. Our advice for auditory physiologists interested in estimating STRFs is to test multiple time-frequency scales. The choice of the range of time-frequency scale should of course be based on the nature of the stimulus and on the known properties of the auditory system but one should remain aware of the complexities described above.

As described in the methods, a time-frequency representation of the sound pressure waveform is a non-linear transformation. It is therefore natural to think about additional non-linear transformations that could be "tagged-on". The simplest non-linearities are static non-linearites of the spectro-temporal amplitude envelopes. We found that logarithmic compressions produce STRFs with more predictive power than linear or power law representations, regardless of time-frequency scale, stimulus type and brain area. In light of the overwhelming preference for logarithmic amplitude scales which we see in zebra finches, we recommend assessing logarithmic amplitude scales or other forms of compressive non-linearities for any auditory STRF estimation. Also, since the response-intensity curves of single auditory neurons vary significantly across neuron types, further improvements could be achieved by optimizing the degree of compression for each neuron.

Beyond simple static non-linearities, one can also evaluate dynamic non-linearities. Although the addition of more complex dynamic non-linearities might make the STRF results hard to interpret, simple forms such as adaptive gain control lead to interpretable results and have a clear physiological correspondence. Alternatively one could use a non-linear dynamic function which describes and is implemented as a model of a lower auditory processing stage than the one being studied. In that case, the STRF could be interpreted as connection weights between the putative lower level cells and the neurons being studied. In this paper, we tested a dynamic non-linear model that satisfied both of these features: first, it incorporated a relatively simple dynamic non-linearity, a form of local adaptive gain control. Second, the adaptive gain control was part of a model of the mammalian auditory periphery that included the filter bank of the cochlea and a compressive non-linearity of the hair cells. In our simulations, we found that this biologically inspired model led to better predictions and that adaptive gain control played a major role in this improvement. That we obtained significantly higher predictions with this simple non-linear function and with minimal fitting in the parameters of Lyon's model is very encouraging. It would not be too difficult to optimize the adaptive gain control feature by varying the parameters that affect, for example, the integration time, the spatial extent or the strength of the adaptive gain. This optimized adaptive gain component could also simply be added to the spectrographic or wavelet decomposition. We suspect that further improvements could be achieved in this manner.

Here we also addressed the methodological issue of validation. Although the CC ratio between the prediction and the response is often used, it should be noted that its value depends on a smoothing window used to describe the neural responses. A longer window reduces noise but could also filter out high frequency neural signal that is reliably predicted by the STRF. As a result, when analyzing parameters such as the optimal time-frequency scale, a long time window would clearly bias the results towards more spectral representations. To circumvent this unwanted effect, we propose the use of the coherence function between prediction and validation to quantify the goodness of fit. The coherence method has the advantage of separating both signal and noise into different frequency bands which can be analysed separately, thus avoiding the problem of filtering out high frequency signal through smoothing. The coherence can be integrated over all frequencies if one desires to report the goodness of the model with a single number as we have done here. Finally, it should also be noted that unbiased and noise-corrected values of the coherence and correlation coefficient can and should be obtained (Hsu et al., 2004a).

In summary, the careful choice of stimulus representation is important for the estimation of STRFs and can lead to better characterisation of the stimulus-response function of neurons given additional relatively simple nonlinear transformations. This is particularly true in the auditory system where STRFs for higher level auditory neurons can only be obtained from a time-frequency representation of the sound stimulus. The appropriate choice for this time-frequency representation with biologically inspired non-linearities can result in significant improvements. Such analysis could be performed for auditory cortical neurons that often respond in a predictive manner to the stimulus but where the simplest version of the STRF has resulted in poor predictions (Machens et al., 2004). The use of non-linear pre-processing steps followed by a linear spatio-temporal filter such as the STRF can also be applied to other modalities for the estimation of the general stimulus-response function of high-level sensory neurons.

Acknowledgments This work was funded by NIMH 015538 to fet.

References

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

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

Cohen L (1995) Time-Frequency Analysis.Prentice Hall, Englewood Cliffs, New Jersey.

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

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. 85: 1220–1234.

Dooling RJ (1982) Auditory perception in birds. In: DE Kroodsma, EH Miller, eds. Acoustic Communication in Birds, pp 95–130.

Eggermont JJ, Johannesma PM, Aertsen AM (1983a) Reverse-correlation methods in auditory research. Q. Rev. Biophys. 16: 341–414.

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.

Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA (2004) Dynamics of precise spike timing in primary auditory cortex. J. Neurosci. 24: 1159–1172.

Escabi MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J. Neurosci. 22: 4114–4131.

Escabi MA, Miller LM, Read HL, Schreiner CE (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J. Neurosci. 23: 11489–11504.

Gentner TQ, Margoliash D (2003) Neuronal populations and single cells representing learned auditory objects. Nature 424: 669–674.

Ghazanfar AA, Nicolelis MA (2001) Feature article: The structure and function of dynamic cortical and thalamic receptive fields. Cereb. Cortex. 11: 183–193.

Gleich O, Manley GA (2000) Hearing Organ of Birds and Crocodilia. In Comparative Hearing: Birds and Reptiles RJ Dooling, RR Fay, AN Popper, eds., Springer-Verlag, New-York: pp 70–138.

Hsu A, Borst A, Theunissen FE (2004a) Quantifying variability in neural responses and its application for the validation of model predictions. Network 15: 91–109.

Hsu A, Woolley SM, Fremouw TE, Theunissen FE (2004b) Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J. Neurosci. 24: 9201–9211.

Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design. J. Comp. Neurosci. 9: 85–111.

Lewicki MS (2002) Efficient coding of natural sounds. Nat. Neurosci. 5: 356–363.

Lyon RF (1982) A computational model of filtering, detection and compression in the cochlea. In IEEE Int. Conf. Acoust., Speech and Signal Processing. Paris, IEEE, France.

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. J. Neurosci. 24: 1089–1100.

Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Pattern Anal. and Machine Intell. 11: 674–693.

Marmarelis P, Marmarelis V (1978) Analysis of Physiological Systems. The White Noise Approach. Plenum, New York.

Okanoya K, Dooling RJ (1987) Hearing in passerine and psittacine birds: A comparative study of absolute and masked auditory thresholds. J. Comp. Psychol. 101: 7–15.

Painter T, Spanias A (2000) Perceptual Coding of Digital Audio. Proc. of IEEE 88: 451–513.

Palmer AR, Evans EF (1982) Intensity coding in the auditory periphery of the cat: Responses of cochlear nerve and cochlear nucleus neurons to signals in the presence of bandstop masking noise. Hear. Res. 7: 305–323.

Phillips DP (1990) Neural representation of sound amplitude in the auditory cortex: Effects of noise masking. Behav Brain Res. 37: 197–214.

Phillips DP, Hall SE (1987) Responses of single neurons in cat auditory cortex to time-varying stimuli: Linear amplitude modulations. Exp Brain Res. 67: 479–492.

Ruggero MA (1992) Physiology of the Auditory Nerve. In The Mammalian Auditory Pathway: Neurophysiology RR Fay, AN Popper, eds, pp. 34–93. Springer-Verlag, New-York.

Sachs MB, Abbas PJ (1974) Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli. J. Acoust. Soc. Am. 56: 1835–1847.

Schlauch RS, DiGiovanni JJ, Ries DT (1998) Basilar membrane nonlinearity and loudness. J. Acoust. Soc. Am. 103: 2010–2020.

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

Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. 114: 3394–3411.

Slaney M (1988) Lyon's Cochlear Model. In Apple Technical Report: 1–79.

Stevens SS (1956) The direct estimation of sensory magnitudes: loudness. Am. J. Psych. 69: 1–25.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J. Neurosci. 20: 2315–2331.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje W, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Comp. Neural. Syst. 12: 1–28.

Von Békésy G (1960) Experiments in Hearing. McGraw-Hill.

Willmore B, Smyth D (2003) Methods for first-order kernel estimation: Simple-cell receptive fields from responses to natural scenes. Network 14: 553–577.

Woolley SM, Casseday JH (2004) Response properties of single neurons in the zebra finch auditory midbrain: Response patterns, frequency coding, intensity coding, and spike latencies. J. Neurophysiol. 91: 136–151. Epub 2003 Oct. 2001.

Yao J, Zhang YT (2002) The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations. IEEE Trans. Biomed. Eng. 49: 1299–1309.

Zevin JD, Seidenberg MS, Bottjer SW (2004) Limits on reacquisition of song in adult zebra finches exposed to white noise. J. Neurosci. 24: 5849–5862.