Modulation spectra of natural sounds and ethological theories of auditory processing

Nandini C. Singh[a] and Frédéric E. Theunissen[b]
Department of Psychology and Neuroscience Institute, University of California, Berkeley, 3210 Tolman Hall, Berkeley, California 94720-1650

(Received 23 April 2003; revised 28 August 2003; accepted 15 September 2003)

The modulation statistics of natural sound ensembles were analyzed by calculating the probability distributions of the amplitude envelope of the sounds and their time-frequency correlations given by the modulation spectra. These modulation spectra were obtained by calculating the two-dimensional Fourier transform of the autocorrelation matrix of the sound stimulus in its spectrographic representation. Since temporal bandwidth and spectral bandwidth are conjugate variables, it is shown that the joint modulation spectrum of sound occupies a restricted space: sounds cannot have rapid temporal and spectral modulations simultaneously. Within this restricted space, it is shown that natural sounds have a characteristic signature. Natural sounds, in general, are low-passed, showing most of their modulation energy for low temporal and spectral modulations. Animal vocalizations and human speech are further characterized by the fact that most of the spectral modulation power is found only for low temporal modulation. Similarly, the distribution of the amplitude envelopes also exhibits characteristic shapes for natural sounds, reflecting the high probability of epochs with no sound, systematic differences across frequencies, and a relatively uniform distribution for the log of the amplitudes for vocalizations. It is postulated that the auditory system as well as engineering applications may exploit these statistical properties to obtain an efficient representation of behaviorally relevant sounds. To test such a hypothesis we show how to create synthetic sounds with first and second order envelope statistics identical to those found in natural sounds. [DOI: 10.1121/1.1624067]

PACS numbers: 43.80.Ka, 43.64.Bt, 43.64.Qh [WA] Pages: 3394–3411

I. INTRODUCTION

Natural sounds span a restricted range of all possible sounds just as natural scenes only represent a small subset of all possible images (Attneave, 1954; Field, 1987). This phenomenology can be quantified by calculating the degree of statistical redundancy found in natural sounds. The use of this redundancy is clearly demonstrated by the multiple forms of compression that are available for the digital storage of music and that result in relatively little perceptual degradation (Painter and Spanias, 2000). We will argue that the characterization of the statistics of natural sounds is also potentially important for understanding acoustical perception and its underlying neuro-physiological basis. A theory of neural representation and neural computation in sensory systems that takes into account the natural environment, as originally proposed by Attneave (1954) and Barlow (1961), has been fruitful in advancing our understanding of the visual system and we propose that a similar approach will lead to insights in auditory science. This theoretical framework leads to a series of predictions and experiments that have now demonstrated how neural computations and representations in the early stages of the visual system are adapted to the processing of natural scenes (reviewed in Simoncelli and Olshausen, 2001). For example, the spatio-temporal receptive fields of visual neurons have been shown to perform optimal filtering operations on natural images (van Hateren, 1992a; Dan *et al.*, 1996).

The use of natural sounds for understanding auditory processing has, for the most part, followed a different path. On one hand, auditory neuroethologists were pioneers in the use of behaviorally relevant stimuli to probe the physiology of the sensory systems. This approach led to the classic discoveries of pulse-echo tuned neurons in the bat (Suga *et al.*, 1978), song selective neurons in songbirds (Margoliash, 1983) and call selective neurons in the primate (Newman and Wollberg, 1978). In this respect, the auditory system appears to be at least as ''selective'' for specific natural sounds as the visual system is for specific natural images. On the other hand, a systematic study of the statistical structure that characterizes these natural vocalizations and then would yield theoretical predictions for the response properties of single or network of auditory neurons has not been pursued to the same degree as in the visual modality.

Two studies have taken this systematic approach by analyzing the statistics of the sound pressure waveform. In an initial study, Rieke *et al.* (1995) demonstrated that auditory nerve fibers in the frog transmitted information more efficiently when the power spectrum of broadband sounds matched the power spectrum of the natural frog call. More recently, by examining the higher-order statistics of natural sounds, Lewicki found that the basis set that best represented the independent components of vocalizations was obtained by a Fourier decomposition whereas the basis set that best

[a]Current affiliation: National Brain Research Center; Sector-15, Part II; Gurgaon, 122 001, Haryana, India.
[b]Author to whom correspondence should be addressed. Electronic mail: fet@socrates.berkeley.edu

represented the independent components of environmental sounds was obtained by a wavelet decomposition. The biological basis set generated by the filtering properties of the cochlea and the hair cells fell in the middle of these two solutions, suggesting that the initial stage of auditory processing could have evolved to be optimized to the different statistics of these two important groups of natural sounds (Lewicki, 2002).

Here we extend this theoretical approach to the auditory computations and representations not of the sound pressure waveform but of the spectro-temporal amplitude envelopes that are obtained by the decomposition of sound into frequency channels. This decomposition is performed in biological systems by the cochlea and in engineering applications, such as speech recognition, by the use of filter-banks. The importance of the spectro-temporal amplitude envelopes of sound in capturing the significant statistics of the natural sounds as well as in predicting the neural response of higher-level auditory neurons is very well documented. First, spectrograms are used extensively in the analysis of animal vocalizations, not only because they provide a clear pictorial representation of the different types of vocal gestures, but also because the spectrographic representation is a better pictorial match of our perception of the sound than any plot of the sound pressure waveform. For the same reasons, time-frequency representations are used extensively in preprocessing stages of speech recognition or sound compression algorithms (Painter and Spanias, 2000). Second, the importance of the statistical structure of these envelopes for speech perception has clearly been demonstrated. Degradation of this structure along either the spectral or temporal dimension results in loss of intelligibility (Drullman *et al.*, 1994; Drullman, 1995; Shannon *et al.*, 1995). Similarly, psychophysical studies have shown that humans are particularly sensitive to either temporal modulations alone (Viemeister, 1979), spectral modulations alone (Green, 1986) or the joint spectro-temporal modulations (Chi *et al.*, 1999) of these amplitude envelopes, and that this sensitivity is restricted to relatively low modulations rates. Finally, whereas auditory neurons in the auditory midbrain and forebrain are not sensitive to the phase of the sound pressure waveform, they do acquire novel temporal and spectral amplitude modulation tuning that is not observed at the lower levels of auditory processing stream (Popper and Fay, 1992). For these reasons, the characterization of the response properties of higher level auditory neurons has included their response to amplitude modulated tones (Phillips and Hall, 1987; Eggermont, 2002), spectrally modulated sounds (Schreiner and Calhoun, 1994; Calhoun and Schreiner, 1998) and more recently to complex spectro-temporal stimuli which are used to extract the joint spectral-temporal receptive fields (STRFs) of the neurons (Eggermont *et al.*, 1983; deCharms *et al.*, 1998; Theunissen *et al.*, 2000; Depireux *et al.*, 2001; Sen *et al.*, 2001; Escabi and Schreiner, 2002; Miller *et al.*, 2002).

Since both auditory perception and the responses of auditory neurons seem to be particularly sensitive to the structure in sound amplitude envelopes it becomes crucial to describe the statistical nature of this structure in natural sounds. Attias and Schreiner (1997) have begun to study the second order statistics of amplitude envelopes along the temporal dimension but very little is known about the joint statistics of the spectro-temporal modulations of natural sounds. For this reason, we investigated the lower order joint statistics of three different ensembles of natural sounds: human speech, zebra finch song and environmental sounds. As was done in Attias and Schreiner (1997), we calculated and fitted the probability distributions of amplitudes for such envelopes. We then calculated the joint second order statistics of the amplitude envelopes, which we call the modulation spectrum. We found that natural sounds have a characteristic modulation spectrum and discuss the implications of our results for an ethologically based theory of auditory processing.

II. METHODS

A. Estimating modulation spectra

Figure 1 illustrates how the modulation spectrum of a sound ensemble is defined. First, a specific time-frequency representation of the sound is calculated. For example, in Fig. 1(a), a spectrographic representation is chosen to display the spectral and temporal structure present in a zebra finch song. This time-frequency representation can be expressed in its Fourier domain. On the right panel, the 2-D image made by the spectrogram is shown as a weighted sum of sinusoidal gratings of variable period, orientation and phase. Each spectrographic ''grating'' corresponds to a particular broadband sound called a ripple sound. The ripple sounds are characterized by their sinusoidal amplitude modulations in time and in frequency. The function describing the amplitude envelope for each frequency band f of a particular ripple sound is written as

$$S_i(t,f)=A_i\cos(2\pi\omega_{t,i}t+2\pi\omega_{f,i}f+\phi_i). \tag{1}$$

The spectrogram for the sound of interest can then be written as a sum of such ripple components (or the equivalent integral in a continuous formulation):

$$S(t,f)=A_0+\sum_i S_i(t,f).$$

A_i determines the relative strength of the modulation depth (relative to the dc term A_0) for that particular ripple sound component. The parameter ω_t describes the modulation frequency of the amplitude envelope along the temporal dimension has units of Hz. In this report, it is referred to as the temporal modulation frequency or simply the modulation frequency. In other reports it has also been named ripple velocity or drifting velocity. Since ripple velocity has been used to describe the number of frequency units spanned per second by the ripple (ω_t/ω_f), we will avoid the use of that term. The parameter ω_f describes the modulation frequency of the amplitude envelope along the spectral dimension and has units of 1/Hz or for wavelet time-frequency representations 1/oct. In this report, it is referred to as the spectral modulation frequency. It has also been called ripple density or ripple peak density (Chi *et al.*, 1999; Klein *et al.*, 2000; Depireux *et al.*, 2001; Escabi and Schreiner, 2002). φ is the initial phase of the ripple. Note that although we use the

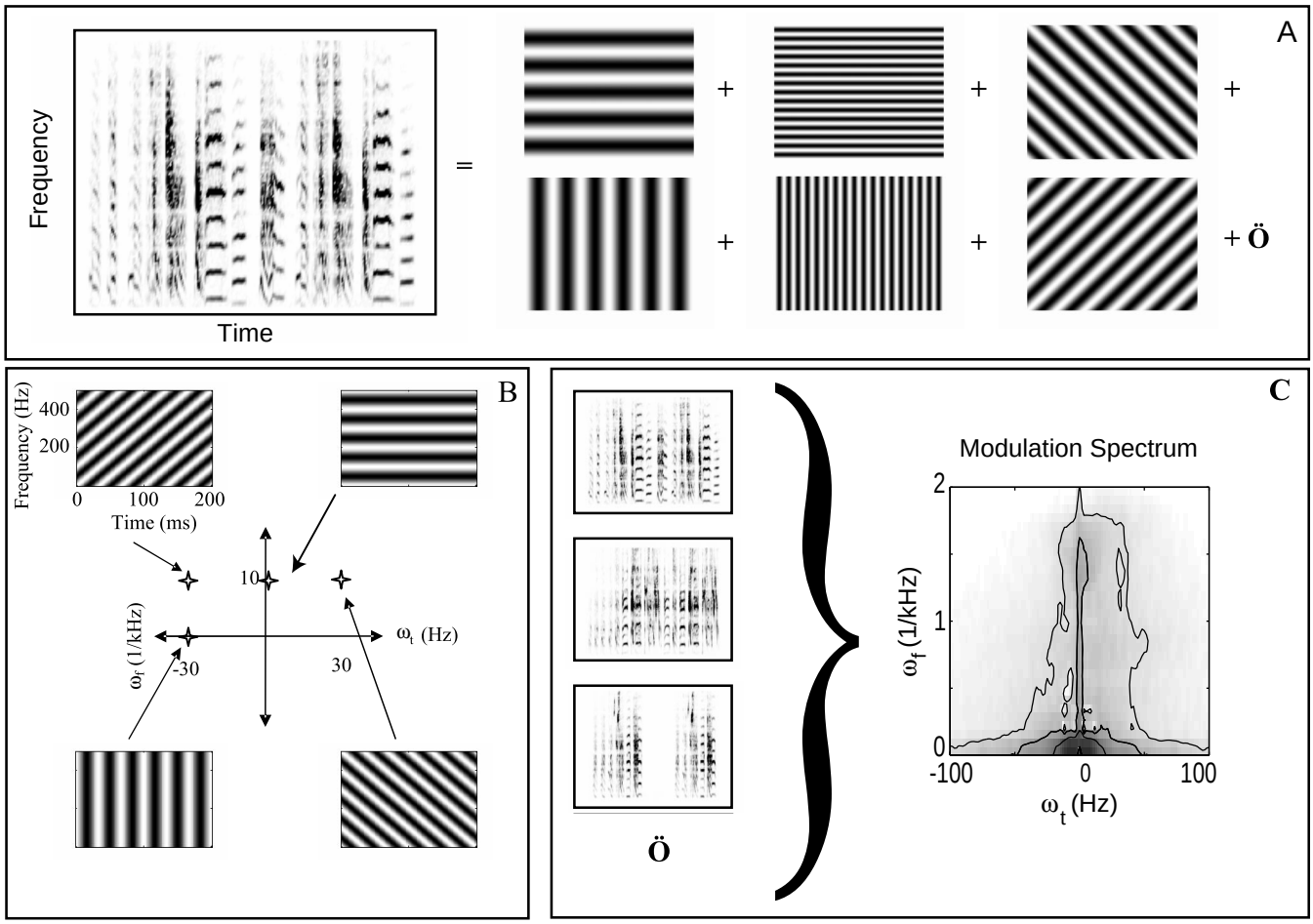

FIG. 1. *Definition of the modulation spectrum.* Panel (a) illustrates the decomposition of a sound represented by its spectrogram into its Fourier components: a sum of ripple sounds, which can be thought of as the acoustic analog of visual gratings. Each subfigure on the right side of the equation is the spectrogram of a single ripple sound component. The sound shown is a song from a zebra finch. As illustrated in (b), ripples are characterized by their temporal modulation, ω_t (Hz), their spectral modulations, ω_f (1/Hz or 1/oct), and their phase. A single point in a two-dimensional Cartesian plot can be used to represent the ripple sound components of a given spectral and temporal modulation, irrespective of phase. (c) To calculate the modulation spectrum, a representative group of sounds is decomposed into its ripple components and the power density of each ripple is estimated and plotted with gray-scale on the two-dimensional Cartesian plot. For the modulation spectrum shown in (c), we used 20 zebra finch songs of approximately 2 s each. The details of the calculations are illustrated in Fig. 2.

symbol ω, the modulation frequencies in Eq. (1) and in the rest of the paper are specified in units of oscillation frequencies and not in angular frequencies.

The modulation spectrum then shows the density distribution of amplitudes A_i of the component ripple sounds for an ensemble of sounds as a function of ω_t and ω_f. Figure 1(c) shows the modulation spectrum for an ensemble of zebra finch song. For time-frequency representations that yield a real valued amplitude envelope, the modulation spectrum is symmetric along the origin and therefore can be shown in two quadrants. Ripple sounds where $\omega_f = 0$ are broadband noise that are sinusoidally modulated in amplitude at a frequency given by ω_t. Ripple sounds where $\omega_t = 0$ are constant sounds with a sinusoidal frequency spectrum where the distance between peaks in the spectrum is given by $1/\omega_f$. Ripple sounds where $\omega_t \cdot \omega_f \geqslant 0$ are down-sweeps and are shown in the upper right quadrant. Ripple sounds where $\omega_t \cdot \omega_f \leqslant 0$ are up-sweeps and are shown in the upper left quadrant [see Fig. 1(b)]. As we will explain below, the range of possible values for ω_t and ω_f is restricted because of the mathematical nature of time-frequency representations.

The calculation of the modulation spectrum is similar to that of the standard frequency spectrum except that it requires the additional preprocessing step of calculating the spectrogram of the sound (or any other time-frequency representation) before calculating the modulus square of the 2-D Fourier transform of the ensemble of spectrograms [Fig. 1(c)]. As is the case for the frequency spectrum, the same result can be obtained by first estimating the auto-correlation function and then calculating the real valued 2-D Fourier transform. Figure 2 illustrates the entire calculation process using this second approach. A spectrogram is first obtained by decomposing the sound into an ensemble of narrow-band signals obtained from the output of a filter bank. The amplitude envelope of each narrow-band signal is obtained from the analytical signal (Flanagan, 1980; Cohen, 1995; Theunissen and Doupe, 1998). The value of the amplitude envelope calculated in that fashion is identical to the amplitude obtained in a short-time Fourier transform of a segment of sound centered at t and windowed with a function given by the Fourier transform of the gain function of the particular filter in the filter bank (Flanagan, 1980). As described in

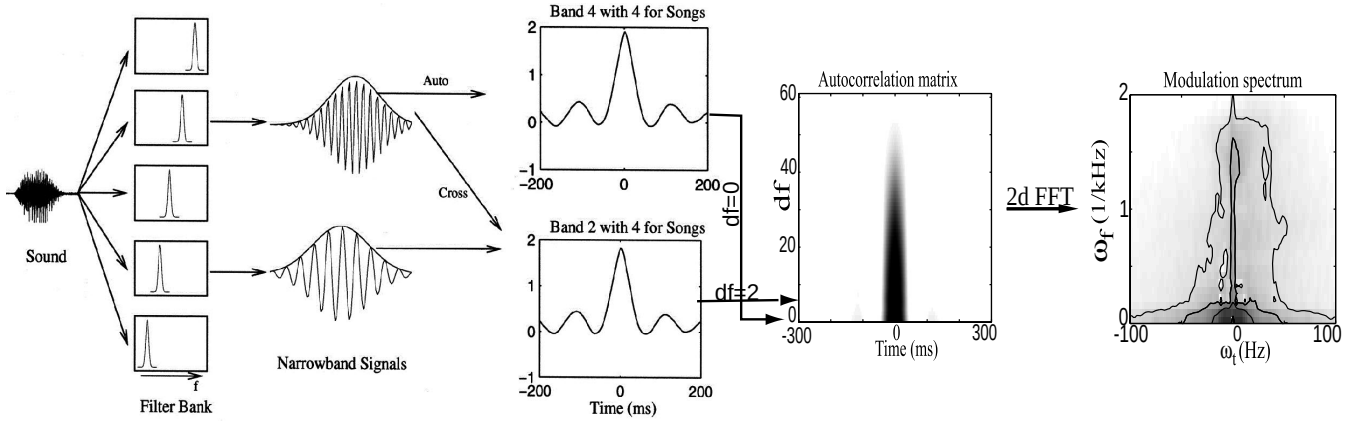

FIG. 2. *Calculation of the modulation spectrum*. A spectrographic representation of the sounds in the ensemble is first obtained by a decomposition into frequency bands using a bank of Gaussian filters. This decomposition results in a set of narrow-band signals with time-varying amplitude envelopes. The spectrogram is a pictorial representation of this time-varying envelope. The stimulus auto-correlation matrix is then obtained by cross-correlating the amplitude envelope of a particular band with the amplitude envelope of all the other bands, including itself. These cross-correlation functions are then averaged for all functions with equal frequency band offsets (df) and collapsed into an auto-correlation matrix, which shows the correlations as a function of time delay (X-axis) and frequency band offset (Y-axis). The two-dimensional Fourier transformation of this auto-correlation matrix is calculated to obtain the modulation spectrum of the sound ensemble.

more detail below, further transformations can then be applied to the amplitude of the envelopes before calculating the modulation spectrum. The amplitude envelopes (or its transformed value) in each band are then used to estimate an autocorrelation matrix, which shows the average product of the amplitude at frequency f and time t with the amplitude at frequency $f+df$ and time $t+dt$. The average is taken over all times t and frequencies f. The 2-D Fourier transform of this auto-correlation matrix yields the modulation spectrum: $P_{\mathrm{MS}}(\omega_t,\omega_f)$, where ω_t are the temporal frequencies corresponding to dt and ω_f are the spectral frequencies corresponding to df. For a modulation spectrum based on a wavelet decomposition, a similar calculation is done but with a filter bank of logarithmically spaced filters and fixed octave widths.

In our studies, we used three separate filter banks. In all three cases, the filters had Gaussian shapes and each filter bank was characterized by the fixed bandwidth of the filters. We used widths of 62.5, 125 and 250 Hz measured as the standard deviation parameter of the Gaussian function describing the gain of each filter. The filters were equally spaced on the frequency axis and separated from each other by one standard deviation. The auto-correlation matrix was calculated for time delays of ± 300 ms. Before taking the 2-D Fourier transform, the auto-correlation matrix was multiplied by a Hanning window. For most of our analyses, we calculated the modulation spectra using a log transformation on the amplitude values and we subtracted the mean log amplitude before windowing. The log transformation was used because, as shown here and previously (Attias and Schreiner, 1997), the distribution of envelope amplitude of natural sounds has a strong exponential component. The last plot in Fig. 2 shows a graphical representation of the modulation spectrum of a zebra finch song obtained in our calculations with the 125-Hz bandwidth filter bank and the log transform. As seen in the figure, the modulation spectrum has a low-pass characteristic both in temporal and spectral modulations and is slightly asymmetric with more energy for down-sweeps than up-sweeps.

B. Time-frequency scale and the estimation of modulation spectra

The bandwidth of the filters in the filter bank has a direct effect on the band occupancy of the amplitude envelope. In each frequency band, the temporal modulation spectrum of the amplitude square of the envelope is restricted to frequencies below the bandwidth of the filter (Flanagan, 1980). Therefore, high frequency temporal modulations can only be observed with wide bandwidth filters. Similarly, the spectral amplitude modulations for a given temporal window along frequency space are restricted to the modulation frequencies below the bandwidth given by the temporal window. Therefore, high frequency spectral modulations can only be measured with wide temporal windows. Since the temporal window is given by the Fourier transform of the gain function of the filter in the filter bank (Flanagan, 1980), high frequency spectral modulations can only be measured with narrow band-pass filters. These properties are another form of the well-known compromise between time and frequency resolution in time-frequency representations (Cohen, 1995).

Because of the time-frequency trade-off in resolution, one cannot generate a spectrographic representation that exhibits both high spectral and high temporal frequency modulations. Since spectrographic representations can be designed to be invertible (up to a single absolute phase) physical sounds that have simultaneously high spectral and high temporal amplitude modulations in the spectrographic time-frequency representation do not exist. More specifically, the uncertainty principle tells us that the product of the bandwidth, σ_f, and duration, σ_t, of the sound sample (the windowed signal used in the spectrographic decomposition) must satisfy the following inequality (Cohen, 1995):

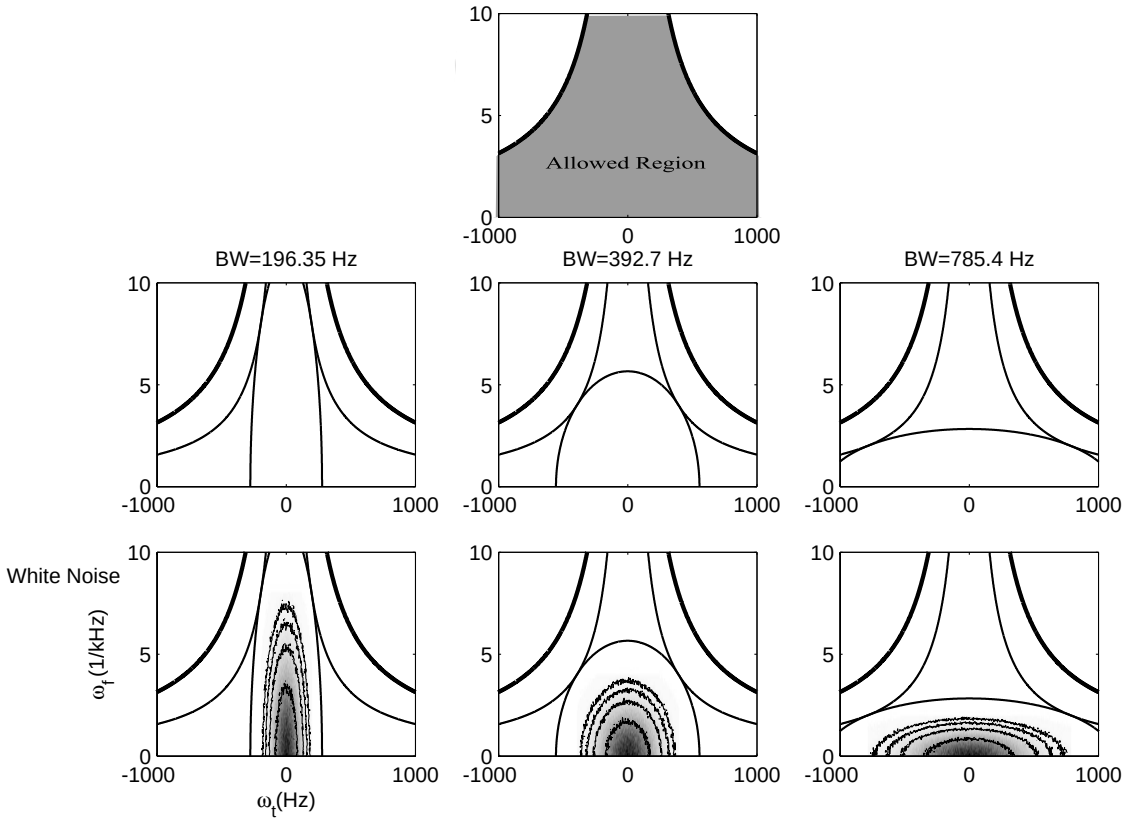

FIG. 3. *Schematics showing the physically plausible region of power for modulation spectrum and the region sampled by choosing a particular time-frequency scale*. The figure in the top row shows the region of physically plausible power for modulation spectra obtained from a spectrographic representation of sound. Since time and frequency are conjugate variables, the uncertainty principle imposes restrictions on the simultaneous sampling of fine temporal and spectral modulation frequencies (see text). The hyperbolic curves given by $\omega_t\ \omega_f = \pi$ enclose the allowed region shaded in gray. The *middle row* shows the actual region sampled by choosing a particular bandwidth for Gaussian filters in the filter bank: only the region within an ellipse can be sampled at a time. For spectrograms, the bounds of the ellipse are given by a second set of hyperbolic curves defined by $\omega_t\ \omega_f = \pi/2$. Choosing narrow bands in the spectrogram (narrow filters in the filter bank) results in a relatively low value for the highest temporal modulation that can be measured and a relatively high value for the highest spectral modulation that can be measured and vice versa. This principle is illustrated by estimating the modulation spectrum of white-noise shown in the *bottom row*. The modulation spectrum of white noise should fill the entire allowed region but, depending on the time-frequency scale chosen, we find power only within the region defined by the ellipse shown in bold. The contour lines show the areas that encompass 50%, 80%, 90% and 95% of the total power.

$$\sigma_f \sigma_t \geqslant \frac{1}{4\pi}.$$

Since we are measuring modulation frequencies (spectral and temporal) by estimating the correlations across multiple measurements of such samples, we can rewrite this inequality, by specifying the upper frequency limit of the temporal and spectral modulations which are given by $\max(\omega_t) = 1/2\sigma_t$, and $\max(\omega_f) = 1/2\sigma_f$. Sounds are therefore restricted to the range of modulation frequencies given by $|\omega_t \cdot \omega_f| \leqslant \pi$.

We provide an illustration of these properties by calculating the modulation spectrum of white noise for the three different bandwidths that we used in our filter bank (see Fig. 3). White noise sound includes all modulation frequencies and should therefore fill uniformly the entire area given by $|\omega_t \cdot \omega_f| \leqslant \pi$ as shown on the top panel (gray area labeled ''Allowed Region''). However, when we choose a particular time-frequency scale for our spectrographic representation, we are effectively only measuring modulation frequencies that are found in a subarea within this allowed region. As mentioned above, the maximum temporal and spectral modulation frequencies are given by the bandwidth of the filter in the filter bank. For Gaussian-shaped filters, the bandwidth of nonzero power is theoretically infinite but the power in the higher frequency modulations quickly decreases. The effective bandwidth of the filter, defined mathematically as the square root of the average deviations square from the center frequency, is simply the standard deviation parameter of the Gaussian filter in the filter bank, σ_{filt}. The bandwidth of the spectral modulation is then given by the maximum spectral modulation: $\text{BW}(\omega_f) = \max(\omega_f) = 1/2\,\sigma_{\text{filt}}$. The time window that corresponds to the Gaussian filter in the filter bank is also a Gaussian function with standard deviation parameter (also the effective duration of the window) given by $\sigma_t = 1/2\,\pi\sigma_f$. The bandwidth of the temporal modulations is then given by the maximum temporal modulation: $\text{BW}(\omega_t) = \max(\omega_t) = 1/2\,\sigma_t = \pi\sigma_f$. Thus when a spectrogram is obtained with Gaussian filters the product of the temporal modulation bandwidth and spectral modulation bandwidth is $\text{BW}(\omega_f) \cdot \text{BW}(\omega_t) = \pi/2 \leqslant \pi$, as required by the uncertainty principle. The hyperbola describing this function within the allowed region is shown in the middle plots of Fig. 3. In addition, for a given time-frequency scale set by the parameter describing the width of the filters in the filter bank (or

equivalently the width of the time window), the modulation spectrum will be restricted to frequencies given by the specific BW(ω_t) and BW(ω_f) corresponding to σ_f. For a Gaussian filter, the regions sampled are not rectangular but, as shown in Fig. 3, ellipsoid with major and minor axes given by $\sqrt{2}$BW(ω_t) and $\sqrt{2}$BW(ω_f).

Figure 3 shows the modulation spectrum for white noise calculated for the three values of the time-frequency scale parameter used in the spectrograms in our analyses, which was determined by the width of the filter in the filter bank (σ_{filt}=62.5,125,250 Hz). The corresponding spectral-temporal modulation ellipse was then determined by BW(ω_t) (196.35, 392.7, 785.4 Hz) and BW(ω_f) (8,4,2 kHz^{-1}). All of the significant estimated energy in the modulation spectrum fell within this ellipse. For white noise, a large area at the center of the ellipse was uniformly sampled, illustrating the fact that white noise also has white modulation spectra. Note, however, that the power in the modulation does decrease significantly before reaching the edge of the ellipsoid area. The modulation spectra for white noise obtained at different time-frequency scales has the same geometric shape but occupies a different modulation frequency area. Clearly, for white noise, there is no appropriate time-frequency scale since the entire range would be required to properly describe the spectral and temporal modulations actually present in the sound. In this case, the modulation spectra obtained from a spectrographic representation are uniquely a reflection of the shape and bandwidth of the filters in the filter bank. For other sound ensembles, however, there might be time-frequency scales that include most of the energy in the spectral and temporal modulations present in the sounds. For those sound ensembles, one could therefore estimate the modulation spectrum from a single spectrographic representation with the realization that modulation frequencies that would fall outside the sampled ellipse would be filtered out.

We attempted to estimate a measure of the optimal time-frequency scale of the spectrogram by measuring the entropy of the power density function given by the modulation spectrum. We reasoned that the modulation spectrum with the highest entropy would be the one that included most of the temporal and spectral modulation structure found in the song. To calculate the entropy, we transformed the modulation spectra into a discrete probability function that specified the probability of the occupancy of a discrete subdivision of the modulation spectrum defined into small rectangles defined by $(\omega_t - d\omega_t, \omega_x - d\omega_x)$ and $(\omega_t + d\omega_t, \omega_x + d\omega_x)$, $p(\omega_t, \omega_f)$. The limits for ω_t and ω_f were chosen to cover the space given by the corresponding sample ellipse for each bandwidth and $d\omega_t$ and $d\omega_f$ were set for all three bandwidths at $d\omega_t$=1.66 Hz and $d\omega_f$=0.065 cycles/kHz. The entropy of the probability distribution is then obtained with

$$H(P) = \sum_{\omega_t} \sum_{\omega_f} -p(\omega_t, \omega_f) \log_2(p(\omega_t, \omega_f)).$$

This entropy measure has units of bits but its absolute value is not interpretable since it is dependent on the size of $d\omega_t$ and $d\omega_f$ and on the choice of units. We are using the measure solely in a relative manner to compare the variability in the modulation spectrum obtained at different time-frequency scales.

C. Descriptive quantifiers of modulation spectra

To quantitatively describe some of the structure observed in the modulation spectra of our sound ensembles, we used a small set of simple measures that estimated the separability, symmetry, low-pass quality and shape. For these four measures, we used the modulation spectra on the log amplitude with the mean level subtracted as explained in Sec. II.A. We also calculated separately the relative power of the dc component of the linear amplitudes to yield a measure of modulation depth. Similar measures have been also used to quantify the modulation spectrum of the spectral-temporal receptive field of auditory neurons (Depireux *et al.*, 2001).

1. Separability

A fully separable modulation spectrum is one that will factorize into a function of ω_t and ω_f over all quadrants. $P_{\text{MS}}(\omega_t, \omega_f) = G(\omega_t).H(\omega_f)$. A separable modulation spectrum signifies that the probability of occurrence of joint spectral-temporal modulations (down-sweeps of up-sweeps) is expected from the average probability of the spectral or temporal modulations measured separately. To quantify the separability, we calculated the singular value decomposition of the modulation spectrum,

$$P_{\text{MS}}(\omega_t, \omega_f) = \sum_{i=1}^{n} \lambda_i g_i(\omega_t) \cdot h_i(\omega_f), \lambda_1 > \lambda_2 > \cdots > \lambda_n,$$

and calculated the ratio of the first singular value relative to the overall power given by the sum of all singular values:

$$\alpha_{\text{sep}} = \frac{\lambda_1}{\Sigma_{i=1}^{n} \lambda_i}.$$

When the modulation spectrum is fully separable, α_{sep} will be close to 1.

2. Asymmetry

The modulation spectrum will be asymmetric if there are more down-sweeps than up-sweeps in the sound ensemble. We quantified the asymmetry by calculating the relative power in the first and second quadrants:

$$\alpha_{\text{asym}} = \frac{P_{\text{down}} - P_{\text{up}}}{P_{\text{down}} + P_{\text{up}}},$$

where P_{down} is the total power in upper right quadrant (of positive ω_t and ω_f) and P_{up} is the total power in the upper left quadrant (of negative ω_t and positive ω_f). If α_{asym} is close to 0, the modulation spectrum is symmetric. If α_{asym} is positive, then there are more down-sweeps than up-sweeps in the sound ensemble and vice-versa.

3. Low-pass coefficient and starriness

We observed that in natural sounds most of the energy is concentrated in the low temporal and spectral frequencies and that the energies in the higher temporal and spectral modulations are not distributed uniformly but instead are

concentrated along the axes. In particular, for animal vocalizations and human speech, most of the spectral modulations are found only for low temporal modulations. To quantify these effects, we calculated two parameters. The first parameter measures the energy in the low frequencies relative to the total energy:

$$\alpha_{\text{low}} = \frac{P_{\text{low}}}{P_{\text{total}}},$$

where

$$P_{\text{low}} = \int_{-\Delta\omega_t}^{+\Delta\omega_t} \int_{-\Delta\omega_f}^{+\Delta\omega_f} P(\omega_t, \omega_f) d\omega_t d\omega_f .$$

We chose $\Delta\omega_t = 10$ Hz and $\Delta\omega_f = 0.195$ kHz^{-1}.

The second parameter measures the relative energy of the modulation spectrum that excludes the regions of joint high temporal and spectral frequency as well as the region of very low joint temporal and spectral frequencies calculated with P_{low}. This area is found next to the x axis and y axis and includes the high temporal modulations but only at low spectral modulations and vice-versa. It is calculated with

$$\alpha_{\text{star}} = \frac{P_{\Delta\omega_t} + P_{\Delta\omega_f} - 2P_{\text{low}}}{(P_{\text{total}} - P_{\text{low}})},$$

where

$$P_{\Delta\omega_t} = \int_{-\infty}^{+\infty} \int_{-\Delta\omega_f}^{+\Delta\omega_f} P(\omega_t, \omega_f) d\omega_t d\omega_f$$

is the total modulation power in a band of spectral frequencies limited by $\delta\omega_f$ and similarly for $P_{\Delta\omega_f}$. P_{total} is the total power in the modulation spectrum.

4. Shape separability

The measure of separability defined above is critically dependent on the power distribution. Since natural sounds have a high concentration of power in the low frequencies, we found that α_{sep} was relatively high for all the natural sound ensembles. However, we could also observe and further quantify with the starriness parameter that the energy outside the low spectral and temporal frequencies was not uniformly distributed. To examine the shape of the distributions, we calculated the separability for an occupancy matrix: given a contour line defined by the percent of the total power within the contour, we set all the values within the contour to have a value of 1 and all values outside to have a value of 0. We then calculated a separability index for this occupancy matrix.

5. Modulation depth

A measure of modulation depth can be estimated by looking at the ratio between the dc power and the power in the rest of the frequencies:

$$\alpha_{\text{mod}} = \sqrt{\frac{P_{\text{total}} - P_{\text{dc}}}{P_{\text{dc}}}}$$

where P_{total} is the total power and P_{dc} is the power at dc, i.e., power at $\omega_t = 0$ and $\omega_f = 0$. We used the square root because modulation depth is usually defined from the amplitude of the envelopes. This measure was applied to the modulation spectrum obtained without the logarithmic transformation.

D. Generating synthetic sounds from a modulation spectrum

We describe a straightforward methodology to generate complex sounds that can match both the frequency and modulation spectrum of a sound ensemble. Our interest is in designing synthetic sounds that match the first and second order modulation statistics of natural sounds. In particular, these sounds can be used to estimate the spectro-temporal tuning of auditory neurons as well as their potential sensitivity to the phase of the modulations that are present in natural vocalizations. Similarly, one could use these synthetic sounds to study perceptual sensitivity to specific spectral-temporal modulations or phase in human or animals.

The method is similar to the method used by Klein *et al.* (2000) and Escabi and Schreiner (Escabi and Schreiner, 2002) to generate synthetic sounds with a band-limited flat modulation spectrum, which has been called noise ripple. For noise ripple the space of ω_t and ω_f is sampled uniformly within some frequency bounds. In our case, we wanted to match our sampling to the modulation spectrum obtained from a particular sound ensemble. For this purpose, we sampled the desired modulation spectrum by normalizing the power spectral density to obtain a probability density function and randomly choosing N distinct pair of values for ω_t and ω_f from that distribution.

To generate the function that describes the amplitude envelope for our synthetic sound, we then obtained the envelope function for a sum of ripple sounds (see Sec. II.A):

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

where φ_i is a random phase for each ripple component. In our implementation we generated synthetic sound ensembles made of 20 synthetic sounds, each 2 s in duration. To synthesize the envelope of each noise ripple sound, we used $N = 100$ ripple components.

An ensemble of sounds with an amplitude envelope given by $S(t,f)$ will have the same modulation spectrum as the original sound ensemble but will also have, on average, a flat frequency spectrum and, on average, a flat temporal envelope. The flat average temporal envelope will also be found in the original sound (if the sound ensemble is stationary in time) but the average flat frequency spectrum is unlikely to be found in natural sounds. To match the overall frequency spectrum and the dc value of the modulation spectra (the modulation depth), we normalized $S(t,f)$ by the average standard deviation of the amplitude modulation in each frequency band f and added the mean amplitude envelope. Calling $A(f)$ the average amplitude in each frequency band measured in the original ensemble, $\sigma(f)$ the standard deviation in each frequency band measure in the original ensemble and $\sigma_S(f)$ the standard deviation obtained from

the ensemble of $S(t,f)$ functions for the synthetic ensemble, we generate a new function for the amplitude envelopes given by

$$S_{\text{Norm}}(t,f)=A(f)+\frac{\sigma(f)}{\sigma_S(f)}S(t,f).$$

Finally, to synthesize the sound a direct method or a more precise iterative method can be used. For the direct method, one simply creates an ensemble of carrier frequencies that will be modulated by S_{Norm}. To prevent any artifacts in periodicity, the frequency of the carrier frequencies should be chosen randomly with a uniform distribution between the lower and upper bounds of the desired frequency range. In our case, we set the lower frequency bound $f_{\min}=250$ Hz and the upper bound $f_{\max}=8$ kHz. Each carrier sound has the form

$$s_i(t)=\cos(2\pi f_i t+\theta_i),$$

where f_i is a random frequency between $f_{\min}$ and $f_{\max}$ and θ_i is a random phase and $i=1$ to Nc. We found that $Nc=1000$ carrier frequencies were more than sufficient to sample our range of frequencies. The synthetic sound is finally given by

$$s_{\text{syn}}(t)=\sum_{i=1}^{Nc} S_{\text{Norm}}(t,f_i)s_i(t).$$

The more precise iterative method is called spectrographic inversion and effectively involves iteratively adjusting the phase of the carrier sounds, θ_i, in order to minimize the difference between the desired spectrogram S_{Norm} and the spectrogram obtained from the synthesized sound (Griffin and Lim, 1984). We used the implementation of the Griffin and Lim algorithm provided by Malcolm Slaney as a Matlab program (1994).

When we used the simple direct method, we found that the ensemble of synthesized sounds had very similar frequency spectrum, modulation depth and modulation spectrum as the original sounds. However, the spectrogram obtained from specific sample sounds from the synthetic ensemble could be quite different from the desired spectrogram due to random phase interferences. The iterative method yielded a much better one-to-one match of the spectrogram and a slight improvement on the match between the ensemble modulation spectra of the natural and synthesized sounds.

E. Natural sound ensembles

We analyzed the statistics of three natural sound ensembles: two types of animal vocalizations (speech and zebra finch song) and an ensemble of environmental sounds.

The speech ensemble was made of 20 sentences chosen randomly from the audio-visual speech test library recorded by the Otolaryngology Department at the University of Iowa (Tyler *et al.*, 1990). The sentences *corpus* consists of 100 short complete sentences read by six different adult male and female speakers. Examples of these sentences are ''It rained all day yesterday,'' ''The book tells a story,'' ''The mother reads a paper'' and ''They have only one son.'' The sentences are read out of context, in an acoustically controlled environment. The total length of sound sampled was approximately 40 s. These speech signals have been used previously in speech perception research (Shannon *et al.*, 1995; Dorman *et al.*, 1997).

The zebra finch song ensemble consisted of the songs of 20 different adult males (age$>$100 days) that were raised by their parents in a large zebra finch colony in our laboratory. The recordings were obtained in a noise-free environment by isolating individual male birds in a sound proof recording chamber. Multiple samples of each song were obtained and a particularly clean exemplar was chosen. Each song lasted approximately 2 s and the 20-song ensemble was approximately 40 s in duration.

The ensemble of environmental sounds was 45 s in duration and consisted of a rustling brush, crunching leaves and twigs, rain, fire and forest and stream sounds. These were recorded and provided to us by Michael Lewicki. Lewicki used these sounds to study the higher order statistics of the sound pressure waveform (Lewicki, 2002).

III. RESULTS

We examined the statistics of the temporal-spectral envelope obtained from spectrographic representations of speech, zebra finch song and environmental sounds.

A. Probability distributions of the modulation amplitude

We first examined the probability distribution of the amplitude of the modulation envelopes (Fig. 4). In this analysis, we used a spectrographic representation based on our intermediate value for the time-frequency scale [$\text{BW}(\omega_t)=392.7$ Hz]. The modulation envelopes were obtained as described in Sec. II.A. The middle row in Fig. 4 shows the distribution obtained for $p(A)$ in each frequency band. The distribution of amplitudes for all three natural sounds is strikingly different from that of white noise. The distribution of amplitudes for Gaussian white noise is given by the Rayleigh distribution: $p(A)\alpha Ae^{-A^2}$. The fit of our data with the theoretical distribution is shown in the bottom panel and the two distributions are indistinguishable (K-S test). On the other hand, the distributions for the natural sounds examined here have a strong exponential component and are best fitted with an exponential distribution or a gamma distribution. The exponential distribution gave good fits for song and speech and for the higher frequencies of environmental sounds. The gamma distribution gave good fits for the lower frequencies of the environmental sounds. The exponential shape of these distributions reflects the fact that, for vocalizations and for the higher frequencies of environmental sounds, there is a finite probability of finding sounds that are arbitrarily soft as in, for example, the silent pauses between speech syllables. We also found that there were systematic trends as functions of the center frequency of the band. Some of the differences can be explained simply by changes in amplitude and not by changes the shape of the probability distribution and these differences would not appear if we had normalized our probability distributions by their variance. For example, the coefficient of the exponential fit decreased from low to high

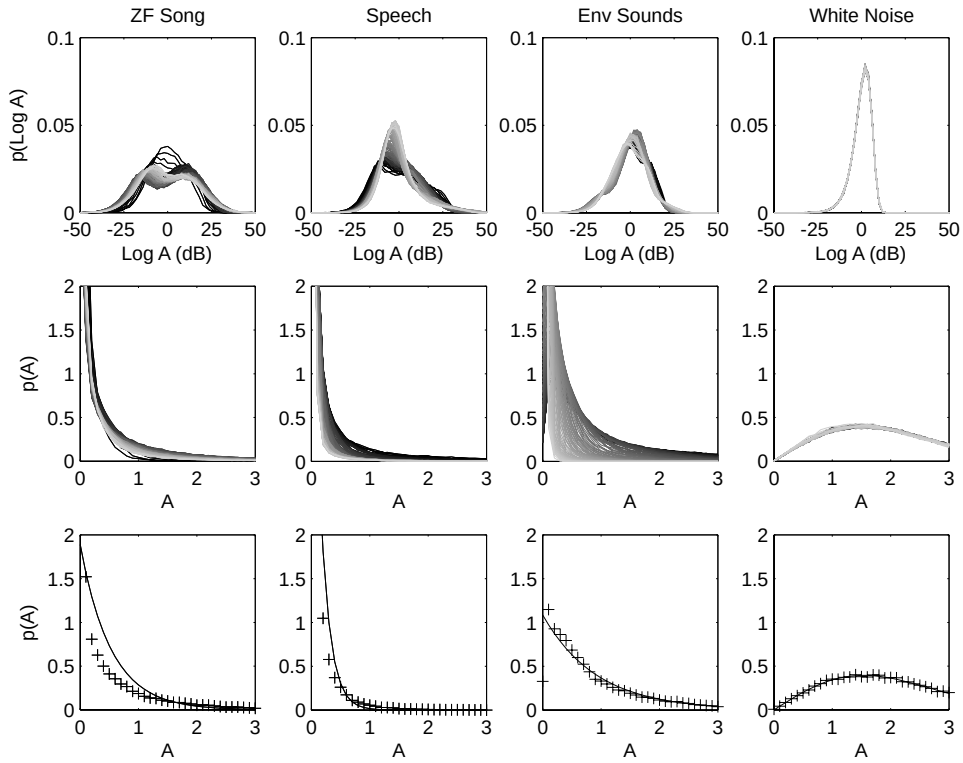

FIG. 4. *Amplitude distributions in different sound ensembles*. Probability density functions were estimated for the amplitude (bottom two rows) and the log amplitude (top row) of the envelope of the four sound ensembles examined in this study. The amplitudes of the envelopes were obtained by calculating the Hilbert transform of the decomposition of the sound into narrow bands as shown in Fig. 2 and explained in Sec. II.A. The filter bandwidth of $\sigma_{\text{filt}} = 125$ Hz was used [$\text{BW}(\omega_t) = 392.7$ Hz]. The gray scale lines show the results for the different frequency bands with the dark black line corresponding to $f = 875$ Hz and the lightest gray line corresponding to $f = 7312.5$ Hz. In the top row the probability distribution of the log amplitude is shown with the x-axis in dB units and with 0 dB corresponding to the mean of the distribution in each band. The middle row shows the probability distribution of the amplitude. The bottom row shows the probability distribution for the frequency band centered at 2500 Hz (crosses) and the best fit (solid) given by an exponential distribution for the natural sounds and given by the Rayleigh distribution for white noise.

frequencies, reflecting the higher probability of soft sounds in the upper frequency range. However, for speech and birdsong, we also observed systematic changes in the shape of the probability distribution as described in more detail in the next paragraph. Finally, although the fits were good in a mean square sense, the data showed systematic deviations from either the exponential form or the gamma form and these two theoretical models were rejected by the K-S test. The theoretical distribution of these natural sounds is therefore complex, potentially composed of multiple components.

The probability distributions were also examined in the logarithmic scale as shown in the top row of Fig. 4 where we plotted $p(\log(A))$ as a function of $\log(A)$. Here, we can also distinguish features that distinguish the distributions of the natural sounds from those of Gaussian white-noise. First, for zebra finch song and the lower frequency bands in speech, the distributions are approximately rectangular (low kurtosis) with similar probability distributions in a 40-dB range (-20 to 20). The distribution for zebra finch song has two peaks (bi-modal) corresponding to syllables and intersyllable silences. The relative peak probability of each peak changes as a function of frequency since there are more song syllables with energy only in the lower frequency range. The probability for speech sounds in the lower frequency range also has low kurtosis but it is not bimodal. Instead the distribution is asymmetric with an almost linear decrease in probability as a function of sound intensity. As frequency increases, the kurtosis also increases reflecting the much steeper slope in the linear region: the kurtosis is below 3 for frequencies below 3500 Hz and above 3 from all frequencies above. The mean kurtosis above 6000 Hz is 4.6. In comparison, the distribution of the log amplitude for the environmental sounds is more symmetric and has a kurtosis close to the normal distribution (mean kurtosis across all frequency bands is 3.26 relative to 3 for the normal distribution).

B. Modulation spectra of natural sounds and time-frequency scale

We calculated the modulation spectrum of the three natural sound ensembles using the methodology described in Sec. II.A and Sec. II.B. Figure 5 shows the modulation spectra for zebra finch song, speech and environmental sounds calculated for the three values of time-frequency scale that we investigated [$\text{BW}(\omega_t) = 196.35$ Hz, $\text{BW}(\omega_t) = 392.7$ Hz and $\text{BW}(\omega_t) = 785.4$ Hz]. The three time-frequency scales capture the principal features in the spectra since most of the energy is found at low spectral and temporal modulations. We can also visually verify the validity of the chosen range of time-frequency scale by noting that, for all three ensembles, the energy for the fastest temporal modulations decay to zero for the wide temporal bandwidth filter

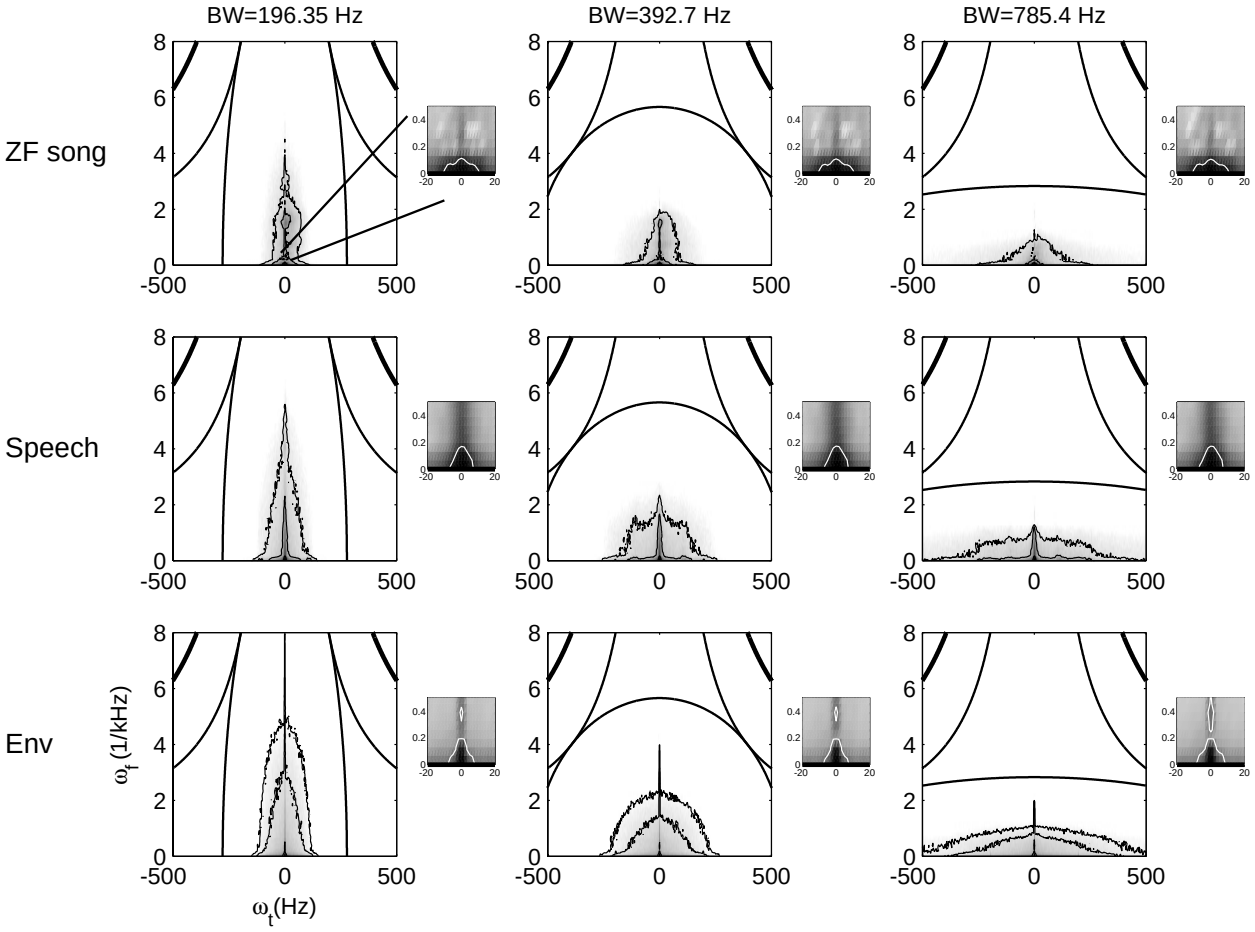

FIG. 5. *Modulation spectra for three different sound ensembles—zebra finch song, speech and environmental sounds obtained at three different time-frequency scales*. The modulation spectra of these three sound ensembles are shown for an effective temporal bandwidth [BW(ω_t)] of 196.35, 321 and 642 Hz. The contours are drawn to circle 50%, 80% and 90% of the total power. The small inset zooms in on the modulation spectra for the lower frequency range. The 50% contour is shown in white on these insets. Since a large fraction of the energy in modulation spectra of these natural sounds is concentrated in the low frequencies, the 50% contour is identical at all three time-frequency scales. Small differences are observed for the 80% contour and large differences for the 90% contour illustrating the fact that the time-frequency compromise affects approximately 20% of the modulation power found at the higher spectral and temporal modulations.

[BW(ω_t) = 785.4 Hz]. Similarly, the energy for the fastest spectral modulation decays to zero for the narrow temporal bandwidth filter [BW(ω_t) = 196.35 Hz]. Nonetheless, the smallest temporal bandwidth misses some of the fast temporal modulations and the widest frequency bandwidth misses some of the fast spectral modulations observed in these natural sounds.

We wanted to find a single time-frequency scale that yielded the best compromise for the representation of the fastest spectral-temporal modulations so that we could quantitatively describe and compare the spectra of these three natural sound ensembles. For this purpose, we calculated an information theoretic entropy measure as described in Sec. II.B. The results of that analysis are shown in Fig. 6. The entropy measures were similar for all three time-frequency scales, reflecting the fact that the probability distributions are well captured at any one of the three scales or that similar compromises are achieved. The highest entropy for the speech ensemble was obtained at the widest temporal bandwidth BW(ω_t) = 785.4 Hz whereas for zebra finch song ensemble the entropy is the highest at the narrowest bandwidth BW(ω_t) = 196.35. The differences, however, were not statistically significant. In the remainder of the paper, we show the results of our analysis of the modulation spectra at the intermediate time frequency scale [BW(ω_t) = 392.7 Hz] but very similar results were obtained at all three scales. To further estimate the effect of the potential compromise, we also calculated the power in the modulation spectra found outside the sampled ellipse. The power outside the area sampled by the intermediate time-frequency scale and found in the area sampled by the BW(ω_t) = 196.35 Hz scale was 1.6% of the total power for zebra finch song, 2.3% for human speech, and 3.7% for environmental sounds.

C. Modulation spectra of natural sounds

Within the allowed space imposed by the uncertainty principle and the time-frequency scale used in the measurements, we observed that the modulations in natural sounds have a characteristic distribution. As mentioned above, for all three sound ensembles, most of the energy is found for low spectral and temporal modulations. These natural sounds and in particular the vocalizations (Zebra finch song and human speech) are further characterized by nonoval distribu-

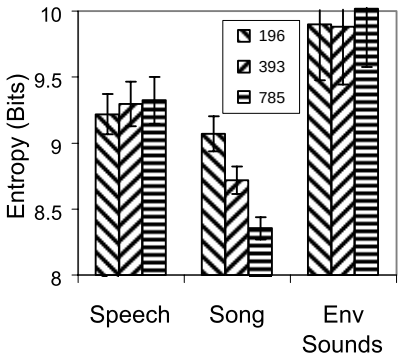

FIG. 6. *Entropy of the modulation spectrum*. The entropy of the modulation spectrum was calculated by treating it as a discrete probability distribution. The entropy was calculated to investigate the optimal time-frequency scale, which would be defined as the scale with the highest entropy value. The error bars show the standard error of the measure obtained by the jack-knife resampling method on the sound ensemble. The legend shows the temporal bandwidth BW(σ_t) in Hz.

tions, reflecting the fact that most of the high frequency spectral modulation power is found at the very lowest temporal modulation and vice versa. In other words, there is a scarcity of sounds with both high spectral and high temporal modulations. This property is best seen in Fig. 7 where we display a contour plot of the modulation spectra of the three natural sound ensembles and white noise for comparison. In this figure, we show all four quadrants of the modulation spectrum to visually emphasize the shape difference. The white noise contours are oval, reflecting the Gaussian-shaped filters, which are symmetric in time-frequency and the relative length of the temporal and spectral axes is determined by the time-frequency scale set by the bandwidth of the filters. For the natural sound ensembles, the contours that bound 50% of the energy are very close to the origin reflecting the low-passed property. The contours that bound 70% and 80% of the total energy draw a star-shape pattern reflecting the low probabilities of finding sound components with jointly high spectral and high temporal modulations. We quantified these observations by calculating various parameters describing the shape of these spectra as described in Sec. II.C. The results of these analyses are shown in Figs. 8 and 9.

First the separability index shows that the three natural sounds ensembles are quite separable. Only the speech ensemble, with an index of 0.84, is significantly different from the index found for the white noise ensemble, which is completely separable [Fig. 8(a)]. On the other hand, our starriness index, which calculates the relative energy in a the low temporal modulation band (the band of ± 10 Hz along the y axis) added to the energy in a low spectral modulation band ($\pm 0.195\ \mathrm{kHz}^{-1}$ along the x axis), is much larger in natural sounds than it is in white noise, reflecting the star-shaped pattern observed in Fig. 7 [Fig. 8(d)]. Although these two results seem contradictory, they have a simple explanation. As quantified by the low pass coefficient, a large fraction of the modulation energy spectrum (64% for zebra finch song, 61% for speech and 51% for environmental sounds) is found at the very low spectral and temporal modulations and the

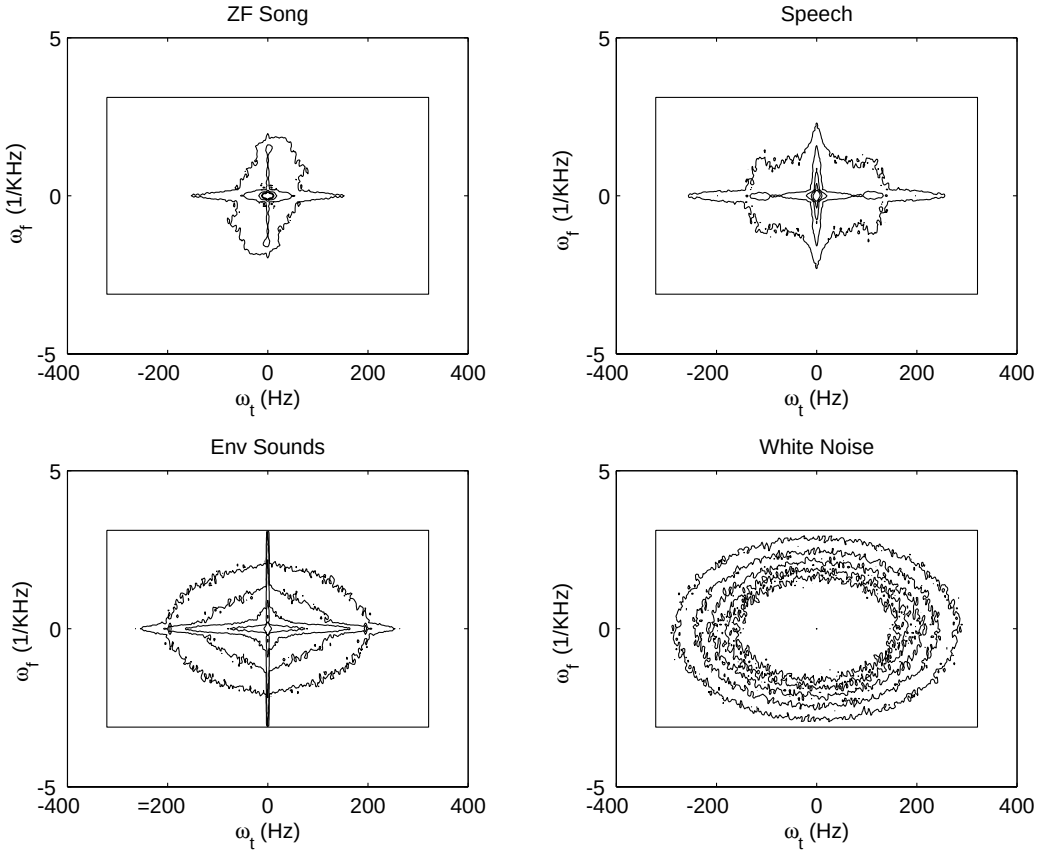

FIG. 7. *Modulation spectra displayed as contour plots*. The modulation spectrum for zebra finch song, speech, environmental sounds and white-noise are displayed with contour plots. The contours plots are drawn at the fixed power values that surround 50%, 60%, 70%, 80% and 90% of the total power.

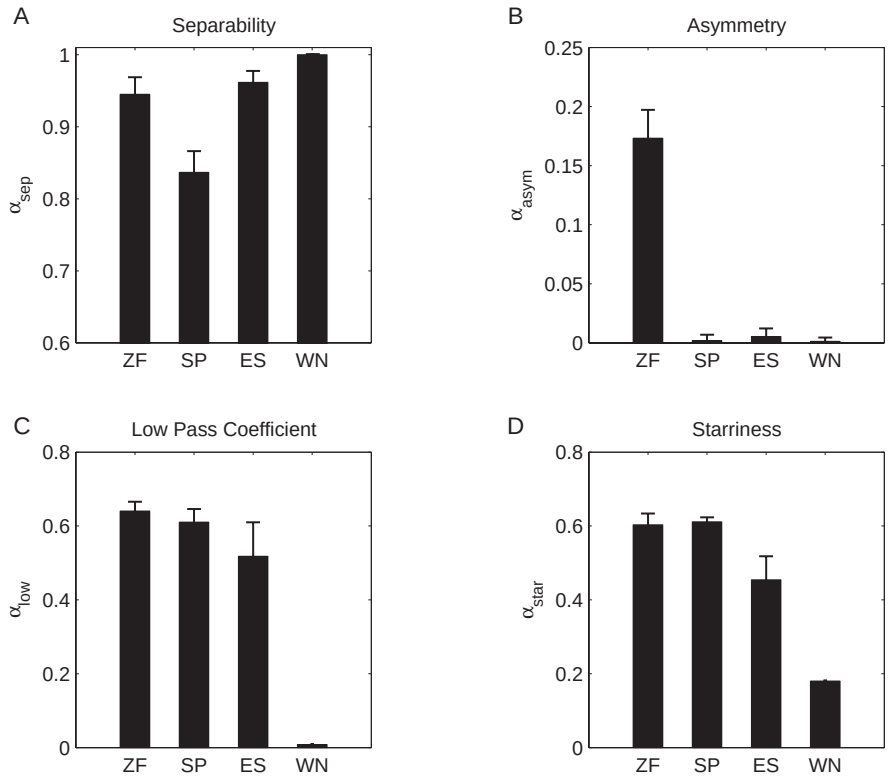

FIG. 8. *Separability, asymmetry, low-pass coefficient and starriness coefficients*. Four quantifiers that measure different aspects of the shape of the modulation spectrum were calculated for each ensemble (ZF: zebra finch song, SP: speech, ES: environment sounds and WN: white noise). The error bars show one standard error obtained with the jack-knife resampling technique. See Sec. II.C for the exact definition of the coefficients.

modulation distribution in this area is highly separable as illustrated by the circular 50% contour in Fig. 7 [Fig. 8(c)]. For those reasons, the separability index remains high. However, when one looks at the tail of the distributions found at the high spectral and high temporal modulations, the natural sounds lack the joint high spectral and temporal power found in the random signal. To measure this effect, we calculated the separability of an occupancy matrix defined by the contours that bound 50% to 90% of the total energy. This analysis showed that speech and environmental sounds are inseparable relative to noise for energy found between the 60% and 80% contours and zebra finch song for energies found between the 80% and 90% contours: the power in the higher modulation frequencies of the spectrum making approximately 30% of the energy for speech and environmental sounds and 20% of the energy for song are particularly concentrated along the spectral and temporal modulation axes.

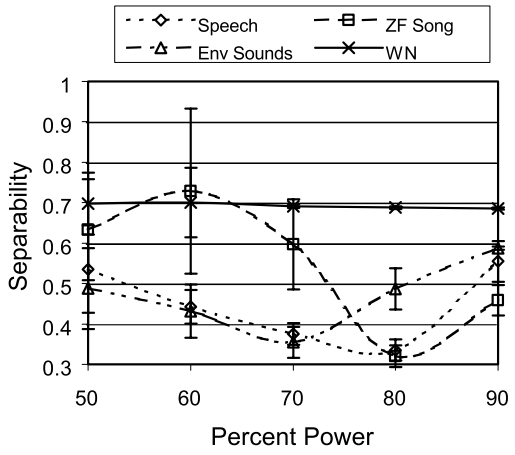

FIG. 9. *Occupancy separability at different thresholds*. An occupancy separability index was defined by calculating the separability coefficient for the space covered by the modulation spectrum. The space covered was defined by the contour line that bounded a given percent of the total power (shown on the *X* axis). The error bars show one standard error obtained with the jack-knife resampling technique.

Besides the low-pass filter characteristics and the lack of jointly high spectral and temporal modulations, the modulation spectra of speech and environmental sounds are remarkably symmetric. These sounds have equal representations of up-sweep and down-sweep ripple sound components (see Figs. 5 and 7). Zebra finch song, on the other hand, exhibits some asymmetry, with slightly more energy in down-sweeps (see Fig. 5). These observations are reflected in the asymmetry index, which measures the relative difference in the two quadrants [Fig. 8(b)]: only the zebra finch song ensemble shows a value of asymmetry that is different from zero.

Finally, we measured the modulation depth of the amplitude envelopes for the four sound ensembles. The modulation depth is traditionally defined as one minus the relative value of the amplitude minimum relative to the maximum: a signal with a modulation depth of 1 is intermittently silent. We used an alternative measure in which we quantified the modulation depth of our signals from their modulation spectra. To estimate the ''size'' of the joint temporal and spectral modulations, we calculated the square root of the ratio of the non-dc power relative to the dc power (see Sec. II.C). Although the actual amplitude modulation observed in the time domain will also depend on the phase of the ripple components of the sound, our measure will be large for signals that are dominated by non-dc ripple components. For example, one would expect isolated animal vocalizations to show larger amplitude modulations than white noise or environmental sounds. Indeed we found that speech had the largest

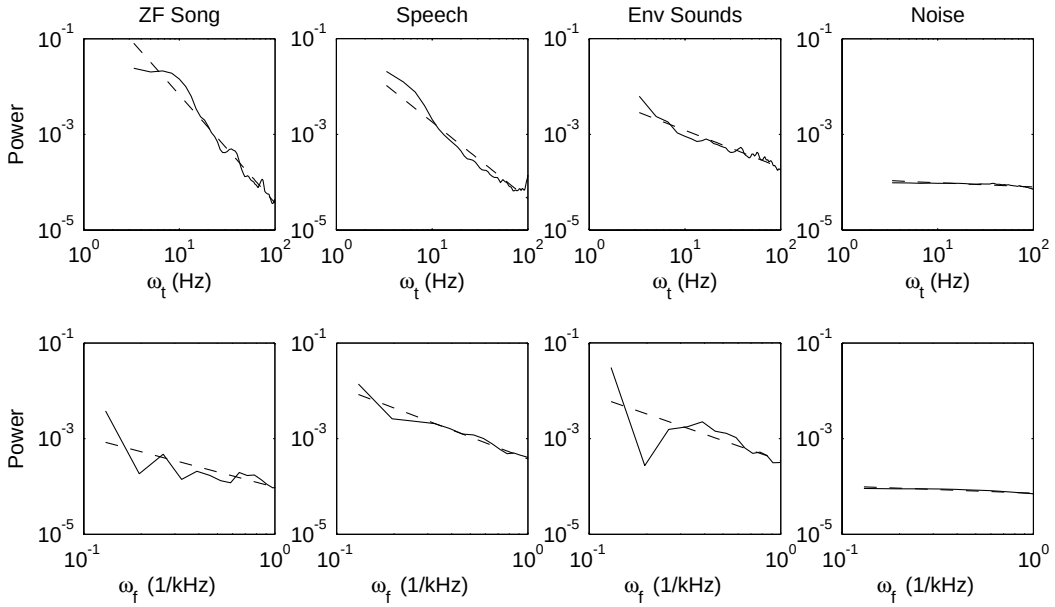

FIG. 10. *Average temporal modulation power and spectral modulation power*. The average temporal modulation power (upper row) and spectral modulation power (bottom row) is plotted as a function of modulation frequency on a log-log plot. The average spectral and average temporal modulations were obtained with a singular value decomposition of the joint modulation spectrum. The data is plotted with a solid line and the power function fit is plotted with a dashed line.

modulation depth (13.8 ± 2.0), followed by Zebra finch song (5.9 ± 0.2), environmental sounds (4.5 ± 0.2) and white-noise (3.92 ± 0.006).

Ultimately, one might want to fit the modulation spectra with a function or a theoretical model. We began this process by fitting the average temporal $[P(\omega_t)]$ and average spectral $[P(\omega_f)]$ components of the modulation spectrum with a power-law function: $P(\omega) \alpha \omega^{-\alpha}$. The average spectral and temporal components were obtained from the first $g(\omega_t)$ and $h(\omega_f)$ functions calculated in the singular value decomposition of the modulation spectrum that was used for the separability analysis (see Sec. II.C). Figure 10 shows these functions (solid line) and the corresponding fits (dashed lines) for all four sound ensembles. The fits were performed for temporal modulations between 3 and 100 Hz and for spectral modulations between 0.1 and 1 kHz^{-1}. This range corresponded to the area of the modulation spectrum where white-noise had a flat distribution, as shown in Fig. 10. Note that, although our effective bandwidth set by the standard deviation parameter of the Gaussian-shaped filters is given by $\mathrm{BW}(\omega_t) = 397.2$ Hz and $\mathrm{BW}(\omega_f) = 4\ \mathrm{kHz}^{-1}$, the boundaries for the areas that showed flat modulation power for white noise are approximately $\frac{1}{4}$ of the total effective bandwidth. Our estimation of the shape of the power distribution must be restricted to that central area or it will be affected by the shape and bandwidth of the filters.

The average temporal components of the modulation spectrum for all three natural sounds were well fitted by the power law (with $R^2 > 0.9$ and $P < 10^{-4}$ in all cases). The vocalizations had steeper slope coefficients with a value for α close to 2 (ZF song: $\alpha = 2.26$, lower 95% = 2.15, upper 95% = 2.36; speech $\alpha = 1.6$, lower 95% = 1.48, upper 95% = 1.72). The slope for environmental sounds was between that of vocalizations and white noise (env sounds: $\alpha = 0.78$, lower 95% = 0.72, upper 95% = 0.84). The approximate $1/\omega_t^2$ relationship is reminiscent of the $1/f^2$ relationship found for spatial frequencies in natural images and its significance is discussed below.

The average spectral components of the modulation spectrum could also be fitted reasonably well with the power law function although additional structure can clearly be observed in the zebra finch song (ZF) and the environmental sounds ($R^2 = 0.95$ $P < 10^{-4}$ for speech; $R^2 = 0.57$ $P < 10^{-3}$ for ZF song, $R^2 = 0.54$ $P = 0.001$ for env sounds). The slope of the power law was shallower than for the temporal component closer to 1 for zebra finch song (ZF song $\alpha = 1$, lower 95% = 0.5 upper 95% = 1.6) and close to 1.5 for speech and environmental sounds ($\alpha = 1.52$, lower 95% = 1.32 upper 95% = 1.72; env sounds $\alpha = 1.4$, lower 95% = 0.6 upper 95% = 2.1).

D. Synthetic sounds with matched modulation spectrum

A final goal of our analysis was to demonstrate how one could synthesize sounds that had similar modulation spectra as arbitrary sound ensembles but different phases in their ripple components. Using the methodology described in Sec. II.D, we generated synthetic zebra finch song, which we called song ripples, and synthetic speech, which we called speech ripples. The top row of Fig. 11 shows the modulation spectrum of zebra finch song and speech and the bottom row shows the modulation spectrum obtained from 400 s of synthetic song ripples and speech ripples. The contour lines surround the area that encloses 50%, 80% and 90% power and the gray scale showing the power is logarithmic. By visual inspection, one can see that the match is relatively good. In particular, the areas of high power are practically identical. The contours that enclose 80% and 90% of the total power are different in the synthetic and natural ensembles but one

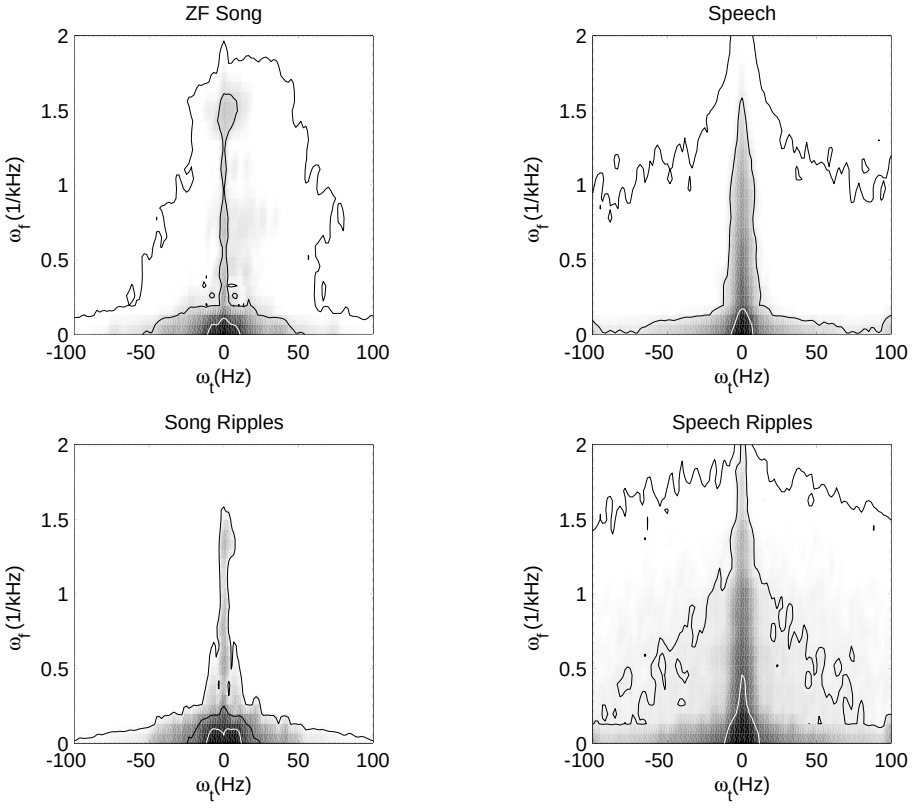

FIG. 11. *Modulation spectrum of natural sounds and their synthetic models*. The top panels show the modulation spectra for zebra finch song and speech. The bottom panels show the corresponding modulation spectra of synthetic sounds that we generated by sampling the spectra of the original sound. The contour lines surround the areas that enclose 50%, 80% and 90% of the total power. The 50% contour lines are drawn in white.

should realize that they correspond to contours drawn at 0.2% and 0.03% of max power for speech and at 0.3% and 0.04% of max power for zebra finch song. These areas of modulation space are therefore infrequently sampled in our synthesis. To better evaluate the quality of the fit, we calculated the cross-correlation coefficient between the modulation spectra of natural sound and that of the synthetic sound: for zebra finch song and song ripples it is 0.9649 and for human speech and speech ripple it is 0.9545.

Spectrograms of exemplars of song ripples and speech ripples are shown in Fig. 12. These sounds have a distinct quality that can be described as zebra-finch-song like and

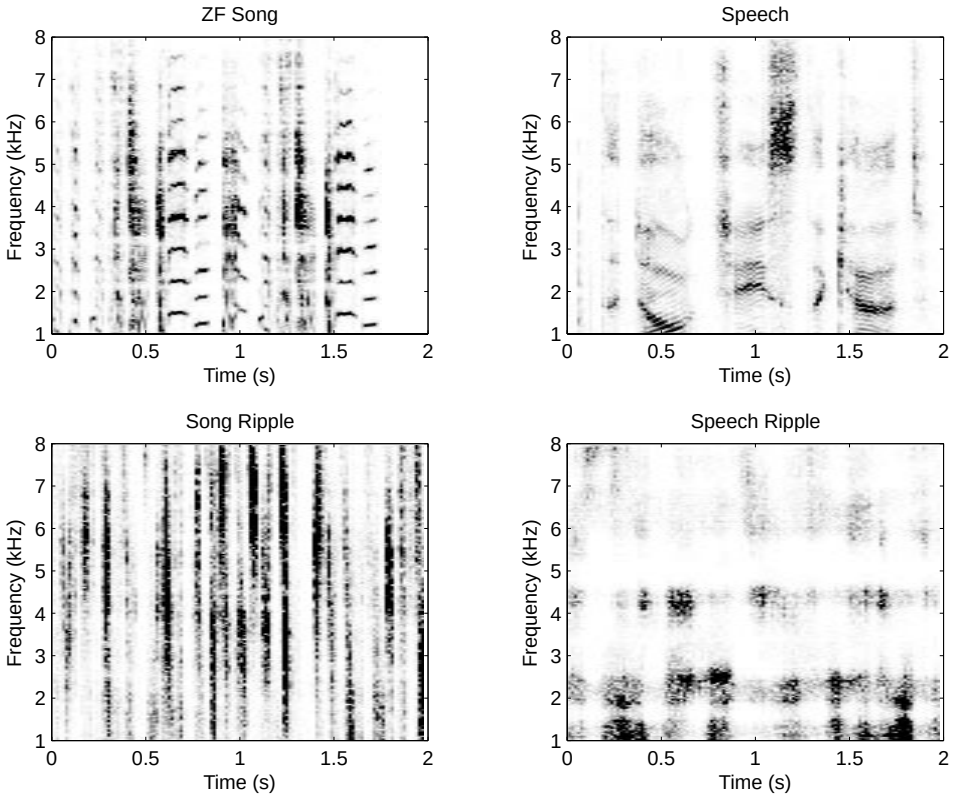

FIG. 12. *Representative spectrograms of a natural zebra finch song and a speech sample (top row) and a synthetic song ripple and speech ripple (bottom row).*

speechlike. Similarly, one can observe, in the spectrograms, similar temporal and spectral modulations in the natural and synthetic ensemble. On the other hand, it is clear that the phase of the ripple components plays an important role in the natural sounds. For example, in the temporal domain, the presence of complete silence is more common in the natural vocalizations than in the ripple sounds. The phase of the ripple sound components also plays a crucial role in the spectral domain, both in generating natural harmonic stacks (all cosine components) and in determining the exact frequency location of the formants in speech. For these reasons and also because the correlations are calculated in 300-ms windows (and 300-ms bits of speech sound parsed together randomly would be unintelligible), the speech ripple sounds are completely unintelligible.

IV. DISCUSSION AND THEORETICAL PREDICTIONS

We have shown that the statistical redundancies that are observed in natural sounds can be, in part, described by the lower-order joint temporal and spectral statistics of the envelopes of the sounds obtained from a time-frequency decomposition of the sound. The distribution of the amplitude of the envelopes of natural sounds has a strong exponential component, which distinguishes it from that of white noise. The modulation spectrum, given by the 2-D Fourier transform of the auto-correlation matrix of the amplitude envelopes, shows that the spectro-temporal envelope modulations in natural sounds are concentrated in the low frequencies, that the average temporal and spectral modulation power can be fitted with a power law and that vocalizations have most of their power in the higher spectral modulation frequencies concentrated at low temporal modulation frequencies.

A. Probability distribution and second order statistics of amplitude envelopes

Our results complement and support previous work that analyzed the statistics of natural sounds. The exponential form of the distribution for amplitude envelopes in speech has been known and exploited in speech processing applications for a long time (e.g., Paez and Glisson, 1972). More recently, Attias and Schreiner (1997) analyzed the amplitude distributions and the temporal modulation power in a larger ensemble of natural sounds which included symphonic music, speech, cat vocalizations and environmental sounds. They also noted the exponential form of the distribution of amplitudes in the natural sounds, which reflects the high probability of finding arbitrarily soft sounds. In their analysis, they found that the distribution in different frequency bands were very similar whereas we found systematic differences in the shape of the probability distribution as a function of the center frequency. However, simple methodological differences between our analyses and theirs could explain the discrepancy: first, we used smaller and more homogenous ensemble of sounds that they did and, second, we used a filter-bank with filters of linearly spaced center frequencies and of fixed bandwidth, whereas they used logarithmically space filters of $\frac{1}{8}$-oct bandwidth.

Attias and Schreiner also calculated the power spectrum of the temporal modulations and fitted their data with a modified power law. There results are similar to those we found by fitting the temporal component of the joint modulation spectrum. In their analysis, they found, as we did, that the power coefficient, α, was between 1 and 2.5. Similar results were also found in the analysis of the distribution of the overall amplitude envelope of speech and music (Voss and Clarke, 1975; Brillinger and Irizarry, 1998). In addition, we found that environmental sounds and animal vocalizations can be segregated into different groups based on these statistics as well as those obtained from the joint time-frequency modulation spectrum (see below). First, the distributions of the log amplitude of the envelopes in those animal vocalizations have low kurtosis, exhibiting a relatively uniform distribution for a 40-dB range of sound intensity, whereas environmental sounds have a log distribution of amplitudes that is approximately normal. Second, animal vocalizations exhibit temporal modulation power relationships that are approximately $1/\omega^2$ whereas environmental sounds are characterized by a significantly flatter power curve. A similar separation of natural sounds into these two broad classes was also suggested by Lewicki (2002) who analyzed the statistically independent components of the sound pressure waveform in these two classes of sounds. He found that vocalizations were best decomposed by Fourier filters and environmental sounds by wavelet filters (Lewicki, 2002).

B. Joint spectro-temporal statistics

In addition, we calculated the joint spectro-temporal modulation spectrum of the natural sounds. It is the natural extension of the purely temporal power-spectra analysis performed on amplitude envelopes or on the overall loudness as described above and the purely spectral power-spectra analysis of the cepstrum performed for speech analysis. We showed how to calculate the joint modulation spectrum and how its calculation was dependent on the choice of the time-frequency representation and on the time-frequency scale of the chosen filters. We chose to perform our analysis using linearly spaced filters with Gaussian shape and fixed bandwidth. The advantages to that representation are that (1) it is symmetric in time and frequency, (2) the limits of the sampled space are well defined and (3) the harmonic sounds, which are very common in animal vocalizations, have well localized occupancy on the spectral axis of modulation spectrum at the value corresponding to the inverse of their fundamental frequency. Since harmonic sounds are critical in acoustical behaviors (e.g., Lohr and Dooling, 1998) and contribute significantly to the statistical redundancy of the sound, the Fourier filter decomposition is in our opinion superior to the wavelet transformation for the characterization of the statistical properties of natural vocalizations. On the other hand, the wavelet transform has the advantage of spanning multiple time-frequency scales. The wavelet transform could therefore be more efficient to estimate the modulation spectrum of sounds that have a broad range of modulation components, such as environmental sounds. These facts and hypotheses are well supported by the analysis performed on the higher order statistics of the sound pressure waveform of vocalizations versus environmental sounds mentioned above (Lewicki, 2002). Moreover, a form of wavelet decomposition

is performed by the mammalian cochlea and the frequency sampling throughout most of the auditory system is approximately uniform on a logarithmic scale. For this reason, the spectro-temporal receptive fields of mammalian auditory neurons have also, until now, been exclusively estimated with a wavelet decomposition (e.g., Depireux *et al.*, 2001; Escabi and Schreiner, 2002). It would therefore be of great value to repeat the statistical analysis performed here using wavelet transforms or other biologically inspired time-frequency representations as those obtained from models of the auditory system (Chi *et al.*, 1999). Note, however, that, because of the physical limits on the frequency range of sounds (and of hearing), a wavelet transform will still result in a modulation spectrum that is bounded to a particular region of spectral and temporal modulations. Also, within the sampled region different frequencies would be analyzed at different scales. Therefore, the results obtained in such analyses would have to be carefully compared to those obtained for white noise and for colored noise, with identical overall frequency power as the ensemble of interest.

Our data suggest that the modulation spectrum for vocalizations can also be distinguished from that of environmental sounds. The vocalizations studied here show significant power in spectral modulations in a narrow band of temporal modulation frequencies (<5 Hz). This power corresponds to the voiced sections of speech and to the harmonic sounds found in zebra finch song. In both vocalizations, there is a scarcity of sound components with both high spectral and high temporal modulations, giving the modulation spectra for vocalizations a characteristic star shape. Environmental sounds are principally characterized by their low-passed quality both spectrally and temporally with a power law function describing both the average temporal and spectral modulation power. The power coefficient describing temporal modulations is smaller (flatter curve) and the power coefficient describing spectral modulations is greater (steeper curve) for environmental sounds than that for vocalizations. We also found that the zebra finch vocalizations are asymmetric with more energy in the down-sweeps than the up-sweeps. We found similar results for other animal vocalizations (bengalese finch song and bat calls; data not shown) and expect this asymmetry to be a common feature of the vocalizations of some animal species. It was strikingly absent from human speech, however.

C. Implications for audio processing

The statistical structure of the spectro-temporal envelopes of natural sounds could have direct implications for various forms of sound processing such as sound compression algorithms for storing music on digital media or speech pre-processing for auditory prosthetics. For example, most current sound compression algorithms (such as MP-3) use standard entropy compression methods on a time-frequency representation of the sound obtained with a filter bank (Painter and Spanias, 2000). The fact that these methods can obtain relatively high compression factors demonstrates that the redundancy in sounds like human music is well captured in the amplitude envelopes of the sound. Since the entropy compression is performed for short segments of time (and therefore optimized for the statistical redundancy at that point in time), it is not clear how the compression could be improved by prior knowledge of the average modulation spectra. On the other hand, information from the modulation spectra could be used to choose appropriate time-frequency decompositions for different classes of sound. Similarly, the knowledge of the differences in modulation spectra for speech versus environmental sounds could be of use for designing preprocessing modules in auditory prosthetics or hearing aids that maximize signal-to-noise ratios for particular signals and noises (see also Chi *et al.*, 1999).

D. Implications for neural coding

Our principal interest is to generate a theoretical framework with which to study the processing of complex sounds in the auditory system of animals. Following the school of thought started by Barlow (1961), we argue that the auditory system has evolved to process behaviorally relevant sounds. For that reason, we expect that the neural representations and computations in the auditory system will be affected by the statistics of behaviorally relevant sounds. In particular, we postulate that the statistics of the spectro-temporal amplitude envelopes of the sound are important for auditory brain areas that are responsible for sound identity. Our results generate various testable hypotheses, which are in some cases, at least qualitatively, supported by psychoacoustical or physiological data.

1. Amplitude coding

The nature of the distributions of the amplitude envelopes leads to a first set of hypotheses. The relatively flat distribution obtained for the log of the amplitude of the envelopes for vocalizations suggests that, in order to discriminate among natural sounds by their amplitude level, an approximate logarithmic amplitude-response curve should be used, as described by Weber's law. 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 nonlinearity is a common property of the auditory system found both at the level of the basilar membrane (Schlauch *et al.*, 1998; Ruggero and Rich, 1991) and in many auditory neurons (Sachs and Abbas, 1974; Palmer and Evans, 1982; Phillips, 1990). In addition, synthetic stimuli that have the natural amplitude distribution were shown to increase the coding efficiency of auditory neurons in the cat inferior colliculus (Attias and Schreiner, 1998) and in auditory neurons of the grasshopper (Machens *et al.*, 2001).

2. Spectro-temporal modulation coding

The characteristic modulation spectrum of natural sounds leads to a second set of hypotheses on neural coding in which the spectro-temporal receptive fields (STRFs) of auditory neurons would be matched to the modulation spectrum of behaviorally relevant sounds. This ''matching'' could take various forms. In the simplest form, a matched-filter

hypothesis, we would expect an approximate one-to-one match between the modulation spectrum of behaviorally relevant sounds and the ensemble modulation transfer function of an auditory processing stage. The modulation transfer function (MTF; also called ripple transfer function or RTF) of a neuron is given by the amplitude of the 2-D Fourier transform of its STRF. The MTF is the equivalent of the gain response (or Bode plot) of one-dimensional filters and shows the spectro-temporal modulations in the sound that will drive the cells (see also Chi *et al.*, 1999; Miller *et al.*, 2002). Given that the modulation spectra of all natural sounds is concentrated in the low frequencies, we would expect to find a concentration of best temporal modulation and best spectral modulation tuning in the low frequencies. This concentration has been observed experimentally in the mammalian thalamus and cortex (Miller *et al.*, 2002) although the quantification of the match has not yet been performed. A second line of evidence for neural tuning to the modulation spectrum of natural sounds was described in the songbirds. In the auditory forebrain, neurons that are selective for conspecific song show the greatest responses to synthetic sounds that have similar modulation spectra as the natural vocalizations. Moreover, to obtain responses similar in strength to those of the natural sounds matching the modulation spectrum was more critical than matching the frequency spectrum (Grace *et al.*, 2003).

An additional prediction can be made from the power law relationship between power and temporal modulation frequency that was observed in this analysis (Fig. 10): if one desires equal driving power for each neuron, then the approximate $1/\omega_t^2$ relationship found for zebra finch song and speech requires the bandwidth of the temporal modulation tuning to be fixed in octave units. Recent data on the temporal MTF is also consistent with this hypothesis (Miller *et al.*, 2002). The $1/\omega_t^2$ relationship is reminiscent of the $1/f^2$ power relationship found in natural images as a function of the spatial frequency, f (Field, 1987). For natural images, the $1/f^2$ relationship implies that the second order statistics of natural images are scale invariant. The equivalent statement for natural vocalizations is that the second order temporal statistics of the amplitude envelopes are invariant to time compression or dilation. This mathematical relationship might explain the perceptual effect whereby sped up vocalizations of a particular animal species often sound like the vocalizations of a different animal species.

A more complex form of matching between the modulation spectrum and the MTF of neurons is predicted if one theorizes a spectral-whitening of the input space. In this framework, neurons effectively amplify stimulus regions of low stimulus power to generate a white response output for each neuron or for an ensemble of neurons. The result is that entropy of the output is maximized and the transmission capacity of the neuronal channel is optimized from an information theoretic perspective (Atick, 1992; van Hateren, 1992b). Experimental data in support for this theoretical framework has been found in the visual system of insects (van Hateren, 1992a) and mammals (Dan *et al.*, 1996). With this perspective, one would then expect asymmetric MTF of neurons, with an amplification of the higher temporal modulation frequencies relative to the lower temporal modulations. In these models, the degree of amplification of the higher frequencies depends critically on the signal-to-noise ratio of the input to the neuron. For flat noise power, the theory predicts that at low signal-to-noise the amplification of higher frequencies is reduced or non existent and the neurons are low-pass filters whereas, at higher signal-to-noise ratio, the amplification comes into play and the neurons become band-pass filters (van Hateren, 1992b). Our data on the modulation spectra of natural sounds could be used to extend such noise analysis. If, for example, the role of an auditory area would be to encode speech in a noisy background of environmental sounds, we would expect to find MTF that amplify the areas of high signal-to-noise ratio and filter out the areas of low signal-to-noise: more explicitly, MTF would be tuned to the intermediate modulation frequencies that are present in speech and not so dominant in the environmental sounds. Interestingly, both psychophysically determined thresholds for ripple stimuli (Chi *et al.*, 1999) and ensemble MTF of neurons in the auditory thalamus and cortex of the cat (Miller *et al.*, 2002) exhibit this type of band-pass filtering.

E. Concluding remarks

In summary, the statistics of the envelopes of natural sounds are characteristically different from those of white noise stimuli. This statistical structure allows us to make predictions on theories of auditory processing based on natural sound statistics. Current psychological and physiological data is, in a qualitative fashion, consistent with these predictions. But these theories remain hypothetical until experimental data are generated to prove or disprove them directly. Also, it is almost certain that different coding hypotheses will apply to different stages of the auditory system and that other biological and physical constraints will be of importance. Given the recent work in the analysis of natural sounds (Attias and Schreiner, 1997; Brillinger and Irizarry, 1998; Lewicki, 2002) and in the development of analytical tools to generate complex synthetic sounds and to extract auditory receptive fields from responses to such sounds or to natural sounds (Theunissen and Doupe, 1998; Klein *et al.*, 2000; Theunissen *et al.*, 2001; Escabi and Schreiner, 2002), we are now in a position where we can directly test this theoretical framework and further advance our understanding of the computations occurring in the auditory system for sound identification.

ACKNOWLEDGMENTS

The authors would like to thank Sarah Woolley, Noopur Amin, and Lee Miller for critical comments on a previous version of the manuscript. The paper was greatly improved by following the suggestions and addressing the criticisms of two anonymous reviewers. The work was funded by NIMH Grant Nos. MH-58189 and MH-66990 to FET.

Atick, J. (**1992**). ''Could information theory provide an ecological theory of sensory processing?'' Network **3**, 213–251.

Attias, H., and Schreiner, C. E. (**1997**). ''Temporal low-order statistics of natural sounds,'' Adv. Neural Info. Process. Syst. **9**, 27–33.

Attias, H., and Schreiner, C. E. (**1998**). ''Coding of naturalistic stimuli by auditory midbrain neurons,'' in *Advances in Neural Information Processing Systems* (MIT, Cambridge, MA).
Attneave, F. (**1954**). ''Some informational aspects of visual perception,'' Psychol. Rev. **61**, 183–193.
Barlow, H. B. (**1961**). ''Possible principles underlying the transformation of sensory messages,'' in *Sensory Communication*, edited by W. A. Rosenbluth (MIT, Cambridge, MA), pp. 217–234.
Brillinger, D. R., and Irizarry, R. A. (**1998**). ''An investigation of the second- and higher-order spectra of music,'' Signal Process. **65**, 161–179.
Calhoun, B., and 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.
Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). ''Spectro-temporal modulation transfer functions and speech intelligibility,'' J. Acoust. Soc. Am. **106**, 2719–2732.
Cohen, L. (**1995**). *Time-Frequency Analysis* (Prentice Hall, Englewood Cliffs, NJ).
Dan, Y., Atick, J. J., and Reid, R. C. (**1996**). ''Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory,'' J. Neurosci. **16**, 3351–3362.
deCharms, R. C., Blake, D. T., and Merzenich, M. M. (**1998**). ''Optimizing sound features for cortical neurons,'' Science **280**, 1439–1443.
Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). ''Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex,'' J. Neurophysiol. **85**, 1220–1234.
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). ''Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,'' J. Acoust. Soc. Am. **102**, 2403–2411.
Drullman, R. (**1995**). ''Temporal envelope and fine structure cues for speech intelligibility,'' J. Acoust. Soc. Am. **97**, 585–592.
Drullman, R., Festen, J. M., and Plomp, R. (**1994**). ''Effect of temporal envelope smearing on speech reception,'' J. Acoust. Soc. Am. **95**, 1053–1064.
Eggermont, J. J. (**2002**). ''Temporal modulation transfer functions in cat primary auditory cortex: separating stimulus effects from neural mechanisms,'' J. Neurophysiol. **87**, 305–321.
Eggermont, J. J., Aertsen, A. M., and Johannesma, P. I. (**1983**). ''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, M. A., and Schreiner, C. E. (**2002**). ''Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain,'' J. Neurosci. **22**, 4114–4131.
Field, D. J. (**1987**). ''Relations between the statistics of natural images and the response properties of cortical cells,'' J. Opt. Soc. Am. A **4**, 2379–2394.
Flanagan, J. L. (**1980**). ''Parametric coding of speech spectra,'' J. Acoust. Soc. Am. **68**, 412–419.
Grace, J. A., Amin, N., Singh, N. C., and Theunissen, F. E. (**2003**). ''Selectivity for conspecific song in the zebra finch auditory forebrain,'' J. Neurophysiol. **89**, 472–487.
Green, D. (**1986**). ''Frequency and the detection of spectral shape change,'' in *Auditory Frequency Selectivity*, edited by B. C. Moore and R. Patterson (Plenum, Cambridge), pp. 351–359.
Griffin, D., and Lim, J. (**1984**). ''Signal estimation from modified short-time Fourier transform,'' IEEE Trans. Acoust., Speech, Signal Process. **32**, 236–242.
Klein, D. J., Depireux, D. A., Simon, J. Z., and Shamma, S. A. (**2000**). ''Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design,'' J. Comput. Neurosci. **9**, 85–111.
Lewicki, M. S. (**2002**). ''Efficient coding of natural sounds,'' Nat. Neurosci. **5**, 356–363.
Lohr, B., and Dooling, R. J. (**1998**). ''Detection of changes in timbre and harmonicity in complex sounds by zebra finches (*Taeniopygia guttata*) and budgerigars (*Melopsittacus undulatus*),'' J. Comp. Psychol. **112**, 36–47.
Machens, C. K., Stemmler, M. B., Prinz, P., Krahe, R., Ronacher, B., and Herz, A. V. (**2001**). ''Representation of acoustic communication signals by insect auditory receptor neurons,'' J. Neurosci. **21**(9), 3215–3227.
Margoliash, D. (**1983**). ''Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow,'' J. Neurosci. **3**, 1039–1057.
Miller, L. M., Escabi, M. A., Read, H. L., and Schreiner, C. E. (**2002**). ''Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex,'' J. Neurophysiol. **87**, 516–527.
Newman, J., and Wollberg, Z. (**1978**). ''Multiple coding of species-specific vocalizations in the auditory cortex of squirrel monkeys,'' Brain Res. **54**, 287–304.
Paez, M. D., and Glisson, T. H. (**1972**). ''Minimum Mean-Squared-Error Quantization in speech PCM and DPCM Systems,'' IEEE Trans. Commun. **COM-20**(2), 225–230.
Painter, T., and Spanias, A. (**2000**). ''Perceptual Coding of Digital Audio,'' Proc. IEEE **88**, 451–513.
Palmer, A. R., and Evans, E. F. (**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, D. P. (**1990**). ''Neural representation of sound amplitude in the auditory cortex: effects of noise masking,'' Behav. Brain Res. **37**, 197–214.
Phillips, D. P., and Hall, S. E. (**1987**). ''Responses of single neurons in cat auditory cortex to time-varying stimuli: linear amplitude modulations,'' Exp. Brain Res. **67**, 479–492.
Popper, A. N., and Fay, R. R. (**1992**). *The Mammalian Auditory Pathway: Neurophysiology*. (Springer-Verlag, New York).
Rieke, F., Bodnar, D. A., and Bialek, W. (**1995**). ''Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents,'' Proc. R. Soc. London, Ser. B **262**, 259–265.
Ruggero, M. A., and Rich, N. C. (**1991**). ''Application of a commercially-manufactured Doppler-shift laser velocimeter to the measurement of basilar-membrane vibration,'' Hear. Res. **51**, 215–230.
Sachs, M. B., and Abbas, P. J. (**1974**). ''Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli,'' J. Acoust. Soc. Am. **56**, 1835–1847.
Schlauch, R. S., DiGiovanni, J. J., and Ries, D. T. (**1998**). ''Basilar membrane nonlinearity and loudness,'' J. Acoust. Soc. Am. **103**, 2010–2020.
Schreiner, C. E., and Calhoun, B. M. (**1994**). ''Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions,'' Aud. Neurosci. **1**, 39–61.
Sen, K., Theunissen, F. E., and Doupe, A. J. (**2001**). ''Feature analysis of natural sounds in the songbird auditory forebrain,'' J. Neurophysiol. **86**, 1445–1458.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). ''Speech recognition with primarily temporal cues,'' Science **270**, 303–304.
Simoncelli, E. P., and Olshausen, B. A. (**2001**). ''Natural image statistics and neural representation,'' Annu. Rev. Neurosci. **24**, 1193–1216.
Slaney M. (**1994**). ''An introduction to auditory model inversion,'' Interval Technical Report IRC1994-014.
Stevens, S. S. (**1956**). ''The direct estimation of sensory magnitudes: loudness,'' Am. J. Psychol. **69**, 1–25.
Suga, N., O'Neill, W. E., and Manabe, T. (**1978**). ''Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the moustache bat,'' Science **200**, 778–781.
Theunissen, F. E., and Doupe, A. J. (**1998**). ''Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVc of male zebra finches,'' J. Neurosci. **18**, 3786–3802.
Theunissen, F. E., Sen, K., and Doupe, A. J. (**2000**). ''Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds,'' J. Neurosci. **20**, 2315–2331.
Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W., and Gallant, J. L. (**2001**). ''Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli,'' Network Comput. Neural Syst. **12**, 1–28.
Tyler, R. S., Preece, J. P. and Tye-Murray, K. (**1990**). ''Iowa Audiovisual Speech Perception Tests,'' Department of Otolaryngology. The University of Iowa, Iowa City, IA 52242.
van Hateren, J. H. (**1992a**). ''Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation,'' J. Comp. Physiol. [A] **171**, 157–170.
van Hateren, J. H. (**1992b**). ''A theory of maximizing sensory information,'' Biol. Cybern. **68**, 23–29.
Viemeister, N. F. (**1979**). ''Temporal modulation transfer functions based upon modulation thresholds,'' J. Acoust. Soc. Am. **66**, 1364–1380.
Voss, R. F., and Clarke, J. (**1975**). ''1/*f* noise in music and speech,'' Nature (London) **258**, 317–318.