nature neuroscience

Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds

Sarah M N Woolley, Thane E Fremouw, Anne Hsu & Frédéric E Theunissen

Vocal communicators discriminate conspecific vocalizations from other sounds and recognize the vocalizations of individuals. To identify neural mechanisms for the discrimination of such natural sounds, we compared the linear spectro-temporal tuning properties of auditory midbrain and forebrain neurons in zebra finches with the statistics of natural sounds, including song. Here, we demonstrate that ensembles of auditory neurons are tuned to auditory features that enhance the acoustic differences between classes of natural sounds, and among the songs of individual birds. Tuning specifically avoids the spectro-temporal modulations that are redundant across natural sounds and therefore provide little information; rather, it overlaps with the temporal modulations that differ most across sounds. By comparing the real tuning and a less selective model of spectro-temporal tuning, we found that the real modulation tuning increases the neural discrimination of different sounds. Additionally, auditory neurons discriminate among zebra finch song segments better than among synthetic sound segments.

For vocally communicating animals, auditory perception is crucial. Juvenile birds and human infants selectively attend to, discriminate among and learn to produce conspecific vocalizations[1–4]. Neural mechanisms underlying the selective and precise perception of natural sounds are not well understood. It is possible that selective auditory tuning acts to match neural responsivity to the acoustics of biologically relevant natural sounds. Such 'matched-filter' tuning may facilitate the detection of stereotyped sounds in background noise[5–7]. Auditory neurons may also be tuned such that the acoustic differences between sounds are maximized in the brain, thereby facilitating auditory discrimination. Here, we provide evidence that selective tuning for spectro-temporal modulations serves as a neural mechanism for the discrimination of natural sounds in the auditory system.

Vocalizations and other natural sounds are characterized by spectro-temporal modulations, oscillations in power across frequency and time[8]. Spectral modulations (cycles (cyc)/kHz) are oscillations in power across a frequency spectrum (kHz) at particular times, such as harmonic stacks. Temporal modulations (Hz) are oscillations in power (amplitude) over time. This characterization of complex sounds has been used to understand the response properties of high-level auditory neurons that cannot be captured by frequency tuning curves[9,10]. It has been proposed that tuning to spectro-temporal modulations, in addition to frequency spectrum, is used to code communication sounds[7,11–14], including speech[15,16]. We used two quantitative analyses to test the hypothesis that, in songbirds, tuning for spectro-temporal modulations in auditory neurons spans the spectro-temporal composition of natural sounds in such a way as to facilitate differences in the neural responses to different natural sound classes and individual sounds. First, we calculated the linear spectro-temporal receptive fields (STRFs) of single neurons and evaluated the relationship between the ensemble tuning derived from these receptive fields and the spectro-temporal modulations characterizing various natural sounds. Second, we measured the neural discrimination of different natural sound segments to test whether modulation tuning facilitates discrimination.

RESULTS

Modulation power spectra and spectro-temporal receptive fields

The spectro-temporal modulations contained in zebra finch song were analyzed by decomposing the songs of 20 birds into their Fourier components (**Fig. 1a**) and calculating a modulation power spectrum (MPS; **Fig. 1b**)[8], which shows signal power as a function of temporal modulations and spectral modulations. Negative temporal modulations distinguish upward (**Fig. 1b**, left half) versus downward (**Fig. 1b**, right half) frequency modulations (FM sweeps). The song MPS shows that zebra finch song contains a limited range of modulations; high frequency spectral modulations occur at low temporal modulation frequencies, and high temporal modulations occur at low spectral modulation frequencies. For example, the harmonic stacks in zebra finch songs are composed of relatively high spectral modulations but are temporally continuous for up to 200 ms and therefore occur at low temporal modulation frequencies. The power centered at temporal modulations below 5 Hz and between 0.5 and 1.5 cyc/kHz spectral modulation shows the power of the harmonic stacks often present in zebra finch songs (**Fig. 1a**,**b**). A synthetic noise stimulus that was limited

Helen Wills Neuroscience Institute and Department of Psychology, University of California at Berkeley, Berkeley, California 94720, USA. Correspondence should be addressed to S.M.N.W. (sw2277@columbia.edu).

Received 11 May; accepted 15 August; published online 4 September 2005; doi:10.1038/nn1536

in spectral and temporal modulations (modulation-limited noise; see Methods) was generated so that the maximum spectral and temporal modulations matched those of song (**Fig. 1b**,**c**). The MPSs show that modulation-limited noise was designed to contain the spectral and temporal modulations in zebra finch song and other modulations by more uniformly sampling the acoustic space bounded by the maximum temporal and spectral modulations in song (**Fig. 1c**,**d**).

We examined the responses of 273 single neurons in the adult male zebra finch auditory midbrain region mesencephalicus lateralis dorsalis (MLd), primary auditory forebrain (field L) and a secondary auditory forebrain region (caudal mesopallium; CM) to modulation-limited noise (**Fig. 2a**). CM was studied because it has been implicated in the perception of familiar sounds[17]. Normalized reverse correlations between stimulus spectrograms and neural responses (spike trains) were used to obtain STRFs for individual neurons (**Fig. 2b**)[18,19]. An STRF describes the linear relationship between a sound and the neural responses to that sound. In the STRF, a spike occurs at time 0, and the *x*-axis shows time preceding the spike. The *y*-axis shows spectral frequency (in kHz). Red indicates the sound that is reliably associated with excitation and blue indicates an absence of sound that is reliably associated with spiking. The pattern shows the presence and absence of specific frequencies that maximally drive a neuron. The mirror image of the STRF can be thought of as the spectrogram of the sounds that best drive a neuron. In this way, a neuron's characteristic frequency, temporal response pattern, excitatory and inhibitory spectral bandwidths, spike latency and temporal precision can be measured. The STRF in **Figure 2b** shows a neuron that has onset characteristics and is driven by lower spectral frequencies. The linear portions of responses, those that are captured by the STRF, ranged between 30 and 89% of the total response, depending on the cell and brain region.

To calculate the modulation tuning of a neuron, a two-dimensional Fourier transform was applied to each STRF to yield the modulation transform function (MTF). Each MTF (**Fig. 2c**, center) plots the modulation tuning of a neuron in terms of spectral and temporal modulations. The axes are the same as the MPS for sounds (**Fig. 1b**,**c**). To obtain the modulation tuning of the neuronal ensembles in each brain region, MTFs for single neurons were averaged for all cells in a brain region to yield the ensemble modulation transfer function (eMTF; **Fig. 2c**, right). Using this approach, the modulations that characterize classes of sounds and the modulation tuning of auditory neurons can be directly compared.

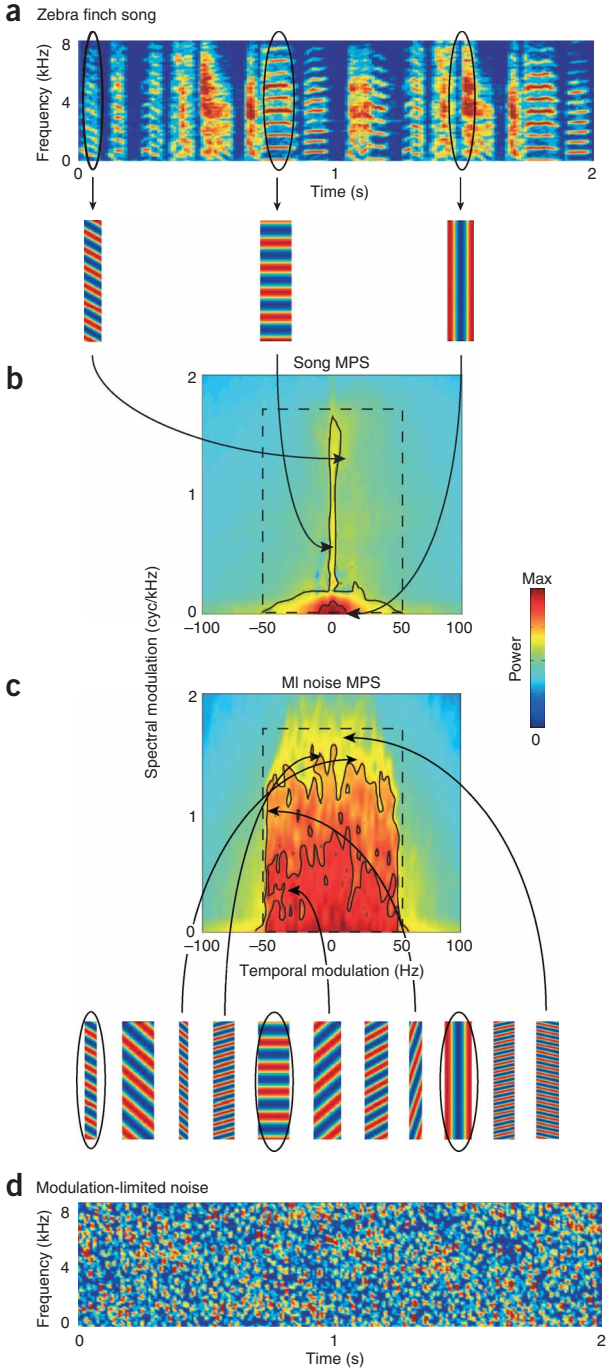


Figure 1 Spectro-temporal modulations in song and modulation-limited (ml) noise. (**a**) A spectrogram of zebra finch song with example spectro-temporal modulation patterns below. Red indicates high intensity and blue indicates low intensity. (**b**) A modulation power spectrum (MPS) shows the spectral and temporal modulations in song. The outer and inner black contour lines delineate the modulations contained in 80% and 50% of the modulation power, respectively. The dashed box indicates the borders of the highest spectral and temporal modulations in song (defined by the 80% contour line) and the defined borders of the modulations included in modulation-limited noise. (**c**) The MPS of modulation-limited noise, with example modulations below. The modulations circled in black (below) are those that also occur in song. The other modulations are not contained in song (see arrows). (**d**) A spectrogram of the modulation-limited noise. Arrows in **b**,**c** indicate the specific points on the MPSs that correspond to the example modulations.

Selective modulation tuning

In response to modulation-limited noise, eMTFs for MLd, field L and CM showed that the ascending auditory system acts as a low-pass filter by selectively tuning to low spectral modulations. **Figure 3a** shows the eMTFs obtained from responses to modulation-limited noise with the 80% power contour line from the modulation-limited noise MPS superimposed in black. This contour outlines the modulations that are prevalent in modulation-limited noise (**Fig. 1c**). Field L shows broader modulation tuning than do MLd and CM because the tuning properties of cells in field L are more complex and varied than in MLd or CM; field L contains multiple cell types and at least five subregions. MLd STRFs showed that most midbrain neurons had strong onset characteristics and simple frequency tuning (single-frequency peaks), with a wide variety of best frequencies across neurons. Individual MTFs were correspondingly simple, showing little tuning for upward versus downward frequency sweeps or spectral modulations. Field L STRFs showed more multiple frequency peaks, more complex excitatory and

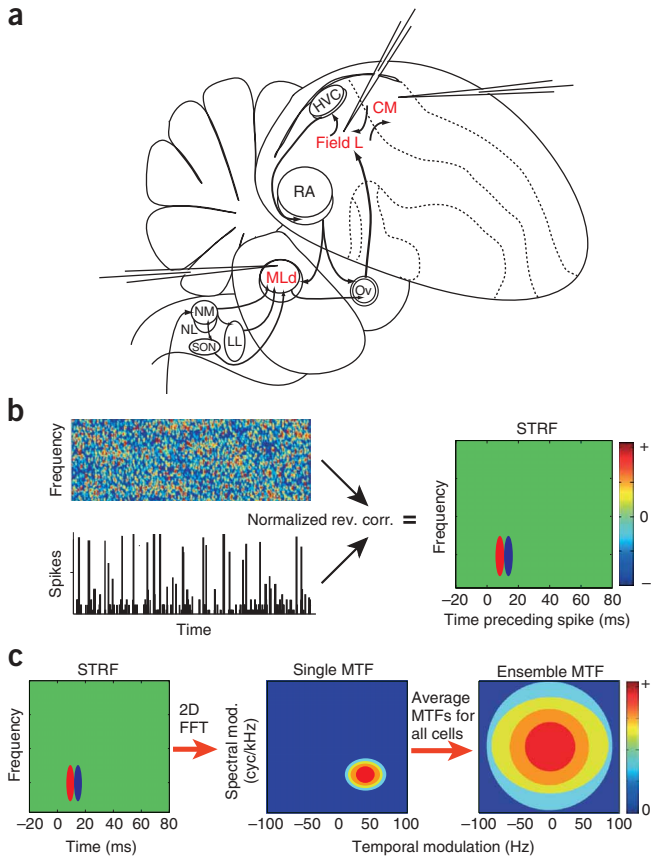


Figure 2 Recording sites and analysis of neural tuning. (**a**) Single-unit recordings were made in the auditory regions MLd, field L and CM. (**b**) A model spectro-temporal receptive field (STRF), showing the linear time-frequency tuning of a single neuron. STRFs were obtained by reverse correlations between stimulus spectrograms and responses (PSTHs). (**c**) Modulation tuning of single neurons was measured by calculating modulation transfer functions (MTFs) from STRFs. Ensemble modulation tuning of all neurons within one brain region was calculated by averaging the MTFs for single neurons to obtain the ensemble MTF (eMTF). For illustrative purposes, hypothetical STRFs and MTFs are shown.

inhibitory tuning and a wider variety of tuning patterns than did MLd STRFs. CM neurons that were linear enough to characterize using STRFs showed tuning properties that were similar to those of MLd cells. Despite the tuning differences across MLd, field L and CM, spectral modulation tuning in all three auditory regions was limited to low frequencies, even though a much wider range of spectral modulations was present in the stimulus. The spectral modulation filter (**Fig. 3b**, left) illustrates the progressive decrease in tuning strength from low to high spectral modulation frequencies. In contrast to the low-pass spectral modulation tuning, temporal modulation tuning was band-pass; neurons were selectively responsive to modulation frequencies between 5 and 55 Hz, with peak gain occurring at frequencies between 25–35 Hz. The temporal modulation filter (**Fig. 3b**, right) shows tuning gain across temporal modulation frequencies. The gains for positive and negative modulations were very similar and were therefore averaged. In all three brain regions, a steep decrease in tuning gain occurs between ~25 and 0 Hz and a decrease on the high frequency side occurs between 35 and 55 Hz, resulting in band-pass tuning.

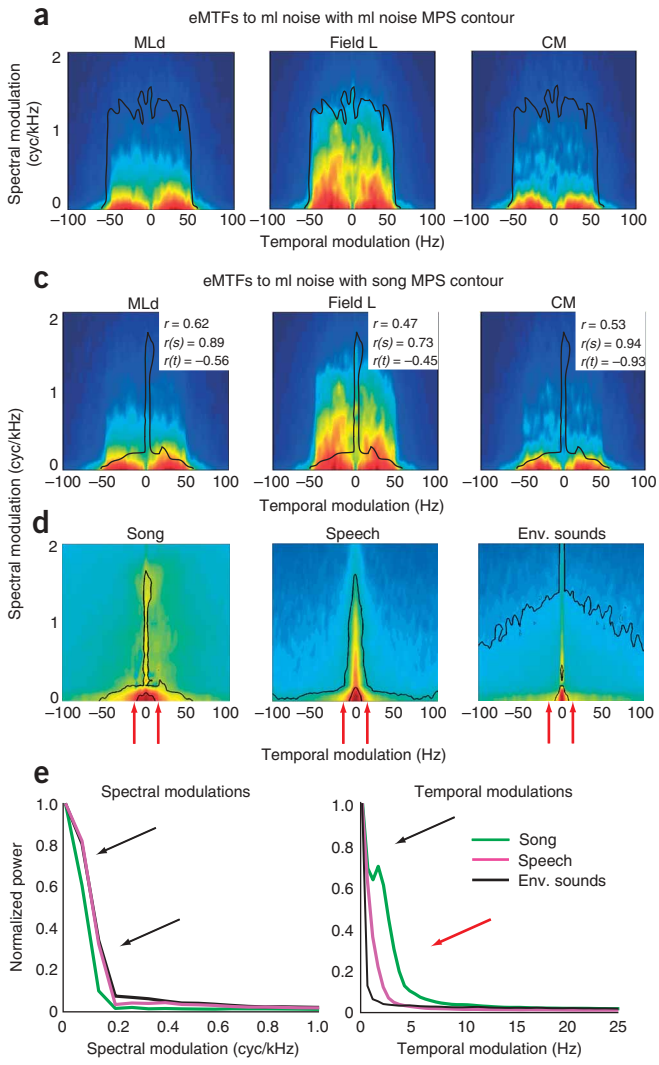
To investigate the relationship between the selective modulation tuning and the modulations that characterize zebra finch song, we compared the ensemble tuning observed in response to modulation-limited noise, shown by the eMTF, with the MPS for zebra finch song, for all three brain regions (**Fig. 3c**). The tuning pattern (color) generally fits within the contour lines showing the modulations that are present in song, although more precisely in MLd and CM than in field L, where neural tuning is more complex. The low-pass spectral modulation tuning matches zebra finch song in that most of the power in song occurs at low spectral modulations (**Figs. 1b** and **3d**,**e**). The correlation coefficients between the eMTFs and the MPS for zebra finch songs were 0.62, 0.47 and 0.53 for MLd, field L and CM, respectively (**Fig. 3c**). The correlation coefficients between the spectral modulation tuning and the spectral modulations in zebra finch song were 0.89, 0.73 and 0.94 for MLd, L and CM, respectively. This general match between selective tuning for low spectral modulations and the prevalence of low spectral modulation power in song supports a matched-filter coding hypothesis, in the spectral domain.

Temporal modulation tuning showed a different pattern. The modulations to which the most tuning gain was devoted were not the most strongly represented temporal modulations in song. While the tuning power peaked at 25–35 Hz and decreased at lower frequencies (**Fig. 3b**, right), the power in the song MPS peaked at the low frequencies and decreased with increasing frequency (**Fig. 3d**,**e**). The correlations between temporal modulation tuning and the temporal modulations in song were negative. Correlation coefficients for MLd, L and CM were −0.56, −0.45 and −0.93, respectively (**Fig. 3c**). For sensory coding, one optimal coding strategy is to whiten the response with a gain function that is the inverse of the stimulus power[20]. Whitening leads to maximal entropy in the response and the highest degree of neural discrimination for a stimulus. The observed tuning is consistent with this strategy; tuning increases the gain of frequencies that are represented with lower power in natural sounds. Perfect whitening would result in a correlation coefficient of −1. The temporal tuning may therefore facilitate the discrimination among similar sounds that contain those frequencies, such as segments of song.

Coding efficiency and discrimination among sounds

To examine the functional significance of the observed temporal modulation tuning for discrimination among different types of sounds such as song and other natural sounds, we compared the ensemble tuning to the spectro-temporal modulations in three classes of natural sounds: zebra finch song, speech and environmental sounds (see Methods, **Fig. 3d**). Although these sounds differ in power at temporal modulation frequencies between ~4 and 50 Hz (red arrows), they all contain high power at low spectro-temporal modulations. The power distributions across spectral modulation frequencies are highly similar across classes (**Fig. 3e**, left), showing high power at low frequencies and rapid decreases in power as frequency increases. The three sound classes differ in the distribution of power across temporal modulation frequencies (**Fig. 3e**, right). At frequencies between 4 and 20 Hz, each sound class shows a different power distribution (red arrow).

Because the lowest-frequency spectro-temporal modulations are present at high power across the three natural sound classes, a neural coding strategy that facilitates sound discrimination might attenuate the representation of these common sounds and focus on sounds with modulations that distinguish one class of sounds from another. To test this, we calculated the modulation frequencies that were present in high power (common) across all three natural sound classes and compared them to the neural modulation tuning in each brain region. We extracted the modulations that contribute 50% of the total power in



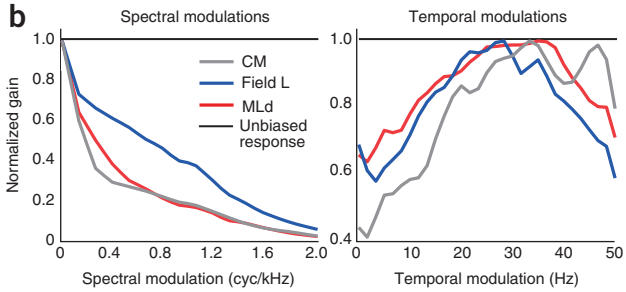


Figure 3 Selective ensemble modulation tuning. (**a**) eMTFs for MLd, field L and CM with the 80% contour line from the stimulus (ml noise) MPS superimposed in black. (**b**) Spectral and temporal modulation filters showing the gain distribution of tuning across spectral and temporal frequencies. In the temporal modulation filter, the gains of negative and positive temporal modulations were averaged. For clarity, only gain from 0 to 50 Hz is shown. Each line shows the gain distribution for one brain region. The black line indicates the theoretical gain function for a response showing no selective tuning (unbiased response). (**c**) eMTFs from responses to modulation-limited noise, with the 80% contour line from the song MPS superimposed. *r* is the correlation coefficient between the eMTF and the song MPS. *r*(*s*) is the correlation coefficient between the spectral modulation filter and the spectral modulations in song. *r*(*t*) is the correlation coefficient between the temporal modulation filter and the temporal modulations in song. (**d**) MPSs for zebra finch song, speech and environmental sounds, showing the modulations in each sound class. Red arrows represent ranges in which temporal modulations differ most. (**e**) Power distributions of spectral modulations in the three natural sound classes show high similarity at all frequencies (black arrows). Power distributions of temporal modulations in the three classes show that they all have high power at low frequencies (black arrow) but differ at mid-frequency modulations (red arrow). This difference corresponds to the regions with the red arrows on the MPSs in **d**.

each sound class (outlined by the inner black contour lines on the MPSs in **Fig. 3d**), and then further extracted only the frequencies that are shared across all three sounds classes (see Methods). **Figure 4a** shows these high-power, common modulations as a red (maximum power) region on an MPS with a blue background (zero power). The sounds represented by these common modulations can be thought of as slowly oscillating, broadband segments. **Figure 4d** shows high-magnification eMTFs for tuning in MLd, field L and CM. The white contour line surrounding the modulations that are common across natural sounds is superimposed over the eMTFs. The band-pass nature of the temporal modulation tuning results in little tuning for the modulations that are most highly represented (common) across sound classes. The degree of filtering of these common sounds by selective tuning was quantified by calculating the maximum possible overlap between the tuning and the range of common natural sounds and comparing that to the actual overlap. Given that no overlap (that is, complete neural attenuation) would be 100% filtering and maximally effective at avoiding the common sounds, the actual neural tuning achieved 38%, 34% and 45% of the maximum possible attenuation in MLd, field L and CM, respectively. These plots and analyses demonstrate that the tuning avoids these common modulations.

Is the avoidance of these common modulations a good coding strategy for the discrimination of natural sounds? Although the low spectro-temporal modulation frequencies had high power in all natural sounds, it is still possible that these modulations would be informative if they varied greatly across the different natural sounds. To quantify which modulations would be most informative for discrimination, we calculated the coefficient of variation (c.v.) across the MPSs of song, speech and environmental sounds (natural sounds, **Fig. 4b**). This c.v. shows the spectro-temporal modulations that vary the most in terms of relative power across sound classes. The high-power regions (red) show the modulations that are most variable, and the low-power regions (blue) show which modulations vary the least. The high-power, common modulations described above are the least variable across the sound classes and can therefore be considered redundant. High variability is found for the high spectral modulations (**Fig. 4b**) and for an intermediate range of temporal modulations (**Fig. 4b,c**). The high spectral modulation region of the c.v. corresponds to the harmonic stacks in song and formants in speech and the absence of these modulations in environmental sounds. The intermediate range of temporal modulations corresponds to a range of frequencies (between 5 and ~25 Hz) that are both characteristic and variable in the natural sounds (**Fig. 3e**, right). The relationship between the modulation tuning of the neurons and the variability in the modulations composing natural sounds is illustrated in **Figures 4d,e**. The eMTFs and the c.v. are compared for the spectral modulation range in which the neural

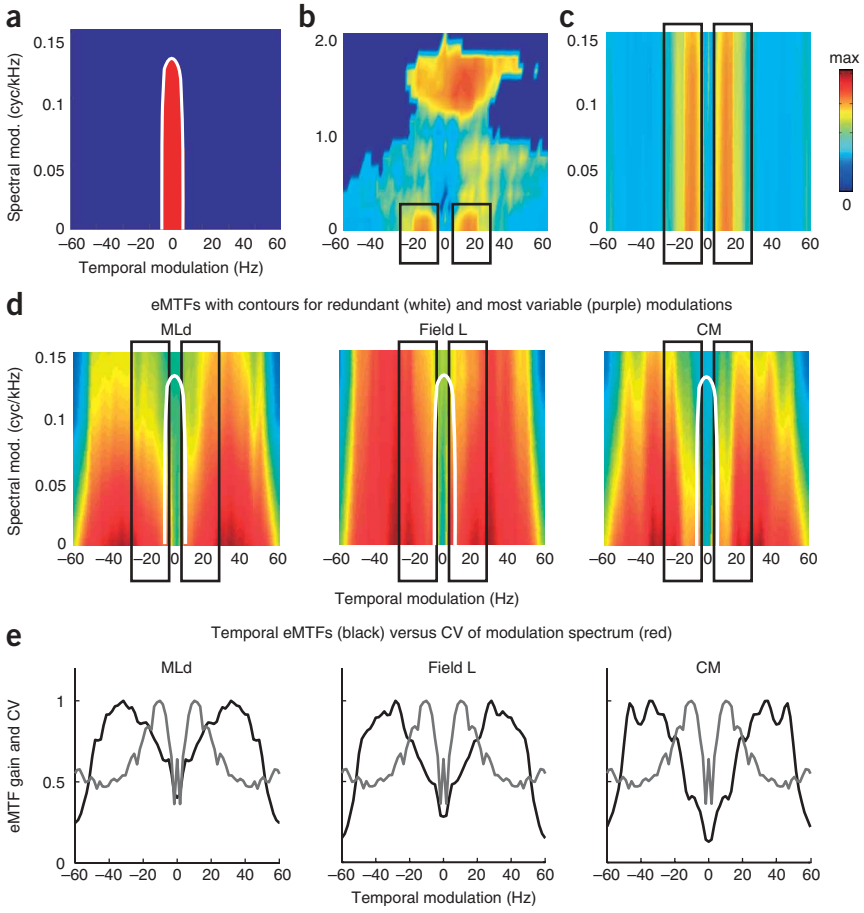


Figure 4 Tuning efficiency in response to modulation-limited noise. (**a**) An MPS showing the spectral and temporal modulations that are redundant across all three classes of natural sounds. (**b**) The coefficient of variation (c.v.) for natural sounds (song, speech, environmental sounds) showing the modulation frequencies with the highest variability across sounds. These sounds vary most at high spectral modulations and at mid-range temporal modulations (black boxes). (**c**) A higher-magnification view of the c.v. showing the temporal modulations that differ most across natural sounds. (**d**) A high-magnification view of the modulation tuning in three classes of natural sounds in response to modulation-limited noise shown in **Figure 3**, with the contour line for the redundant modulations (white) and the boxes showing the temporal modulations that vary most across sound classes (black) superimposed for comparison. The redundant modulations fall where tuning power is attenuated and the most variable temporal modulations overlap with the tuning. (**e**) Line plots of the summed temporal eMTF gain and c.v. for three classes of natural sounds for spectral frequencies below 0.15 cyc/kHz. The plot shows that the modulations that vary the most across natural sounds are found where the temporal modulation gain function is steepest.

tuning is observed (that is, a range that is limited to between 0 and 0.15 cycles/kHz). In **Figure 4d**, the black boxes from the c.v. plots are superimposed on the eMTFs to compare regions of tuning with the temporal modulations that most distinguish one class of natural sounds from another (that is, the most variable modulations). The temporal frequency range of higher variability is in the area of intermediate gain and the highest slope in the temporal modulation tuning (**Fig. 4e**). In this intermediate frequency range, the largest change in ensemble tuning gain is obtained for a given change in temporal modulation frequency in the stimulus. By contrast, the low variability for the redundant modulations can be seen as a blue band running down the center of the c.v. plot (**Fig. 4c**) and a trough in the line plot of the c.v. of the temporal modulations (**Fig. 4e**, gray line). The lack of variability in power at these frequencies matches the lack of tuning for those frequencies (**Fig. 4e**, black line). In this way, neurons avoid tuning for the very low frequencies that vary least among natural sounds and have high gain slope for modulations that vary the most. Large neural gain is also observed for higher temporal frequencies that, nonetheless, vary little across different natural sounds. These higher temporal frequencies could, however, be informative for discrimination within a sound class and, because they have little power, would need to be amplified as predicted by the whitening hypothesis.

Effects of modulation tuning on neural discrimination

Midbrain and forebrain neurons process sounds efficiently by selectively filtering out redundant modulations and showing high gain sensitivity for variable modulations, a mechanism that may promote the neural discrimination of natural sound classes such as songs, speech and environmental sounds. The neural tuning gain also seems to whiten the modulation spectra of specific classes of natural sounds, which may also facilitate discrimination. To further investigate the relationship between modulation tuning and the neural discrimination of sounds, we generated a model of neural tuning that did not show band-pass temporal modulation tuning (see Methods). The observed (band-pass) and model (non–band-pass) eMTFs are shown in **Figure 5a**. Responses to example natural sounds were predicted using both the observed (real) and the model modulation tuning patterns and analyzed for neural discriminability. The responses yielded by the real tuning pattern showed larger dynamic ranges and higher frequencies than responses yielded by the model tuning pattern, suggesting that the selective modulation tuning observed in responses to modulation-limited noise facilitates the discrimination of different sounds (**Fig. 5b**).

To quantify the potential advantage of band-pass temporal modulation tuning, we calculated a measure of neural discrimination based on the normalized distance between the neural responses obtained for two different 100-ms segments of sound (see Methods). In this case, the neural response was the ensemble linear response obtained by convolving the stimulus with an ensemble STRF obtained from the eMTF, either the band-pass (real) eMTF or the Gaussian (model) eMTF. The real band-pass tuning results in significantly greater neural discrimination than does the non–band-pass model (**Fig. 5c**). This effect was observed both for discrimination of sounds across natural sound classes (top) and for individual sounds within a class (bottom). This analysis does not show that the observed band-pass tuning is the optimal tuning. But it demonstrates that the observed tuning pattern is advantageous for discriminating among sounds across and within classes because higher frequency modulations are emphasized, resulting in responses with larger bandwidth (**Fig. 5b**). Across the natural sounds that we examined, the intermediate temporal frequencies vary

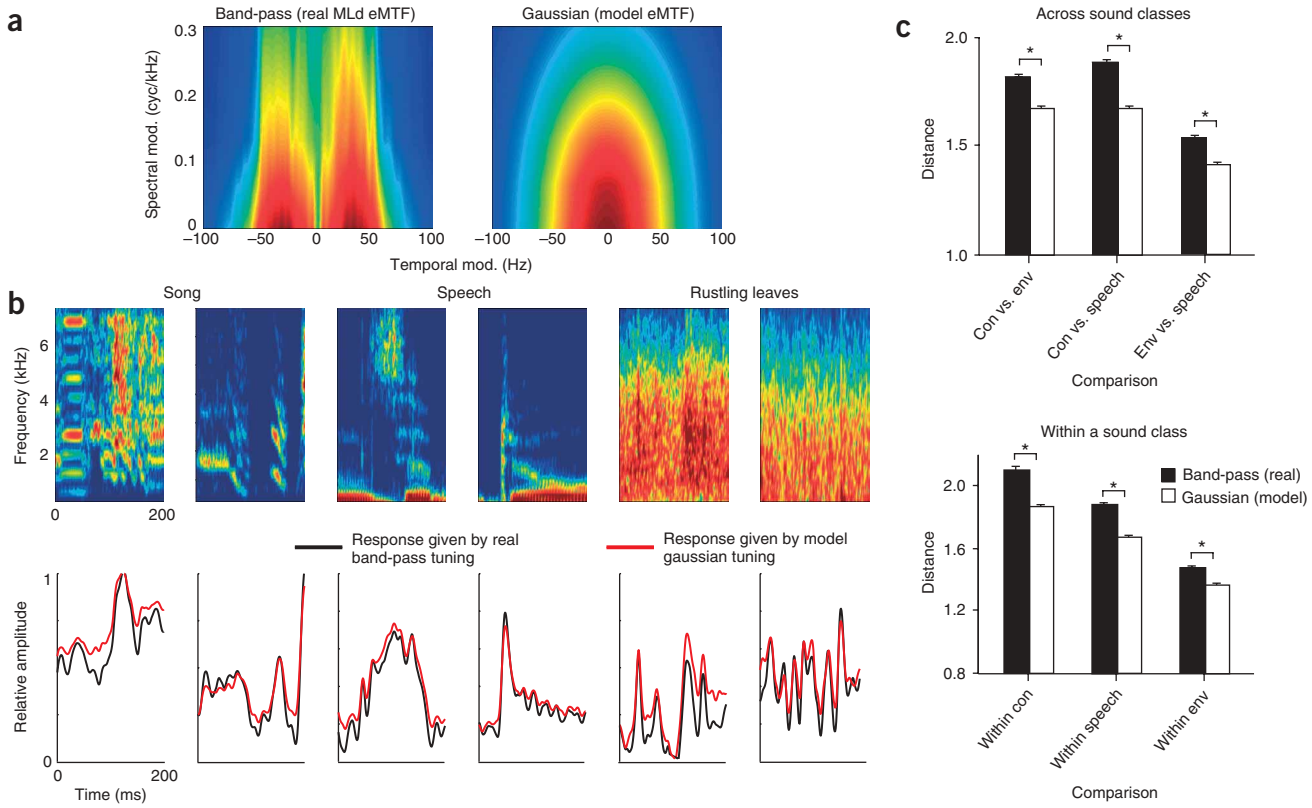


Figure 5 Natural versus model modulation tuning and the neural discrimination of natural sounds. (**a**) High-magnification eMTFs showing the real, band-pass neural tuning in MLd and an alternative model tuning pattern with Gaussian neural tuning. (**b**) Spectrograms of song, speech and environmental sound segments with the predicted ensemble responses to those sounds below. The ensemble responses were obtained by convolving the spectrograms of the sounds with an ensemble STRF obtained from the eMTF. The responses differ depending on the tuning pattern used to calculate them. The real, band-pass tuning yields a higher dynamic range of responses than does the model tuning. (**c**) Average normalized distances between segments of song, speech and environmental sounds in MLd, field L and CM. The distances are calculated for the neural response to random pairs of sound segments, either across sound classes or from the same class. Distance measures show that responses calculated using the real, band-pass temporal tuning pattern increase neural discrimination when compared with responses obtained using the model eMTF. Error bars: s.e.m. *, $P < 0.001$.

the most. These frequencies are amplified relative to lower frequencies by the tuning gain function. The advantage of high gain sensitivity (high-slope) for these intermediate, informative frequencies has not been explored in this model of the ensemble response. It is possible that high gain sensitivity could result in differential recruitment of the number of neurons for different classes of sounds.

Neural discrimination of song versus synthetic sounds

The match between the observed modulation tuning and the informative modulations in natural sounds suggests that zebra finch auditory neurons should be better at discriminating among natural sounds relative to synthetic sounds, which lack natural modulation and phase spectra. We tested this prediction by comparing the neural discrimination of modulation-limited noise and of zebra finch song segments (see Methods). Modulation-limited noise is 'unnatural' in that the modulation power spectrum is approximately flat within the band limits (**Fig. 1c**), and the modulation phase spectrum is random, resulting in uniform variability across modulation frequencies (measured with c.v., data not shown). For each neuron, we recorded responses to both modulation-limited noise and song. The average spike rates in responses to these stimuli did not differ (data not shown). We estimated the discriminability of the neuronal ensembles in each brain region for zebra finch song segments and modulation-limited noise segments using the normalized distance measure. We also separated the linear fraction of the response from the entire response to show that the linear tuning described by the STRFs facilitates the discrimination of natural sounds (**Fig. 6**).

The responses of all neurons to an individual song are shown in a neurogram (**Fig. 6a**). Each row shows the response of one neuron. The STRF for each neuron was convolved with the stimulus to obtain the linear prediction of the response, shown as a linear neurogram. To measure neural discrimination, we calculated the average noise-normalized distance between two neurograms of randomly chosen pairs of song segments (**Fig. 6b**; see Methods). This normalized distance was calculated for the linear neurograms and the neurograms of the entire responses. The process was repeated for modulation-limited noise.

In all three brain regions, the neurogram distance was significantly higher for song than for modulation-limited noise (**Fig. 6c**, left). This result agrees with an information theoretic analysis of these data[21]. Moreover, we show here that the increased discrimination for natural sounds can, at least in part, be explained by the linear tuning properties of the neurons; the linear fraction of the response shows the same trend as the entire response and contributes significantly to the difference in discriminability between the stimulus types.

DISCUSSION

Efficient neural coding strategies that extract meaningful signal from noise and reduce stimulus redundancies are predicted by behavioral

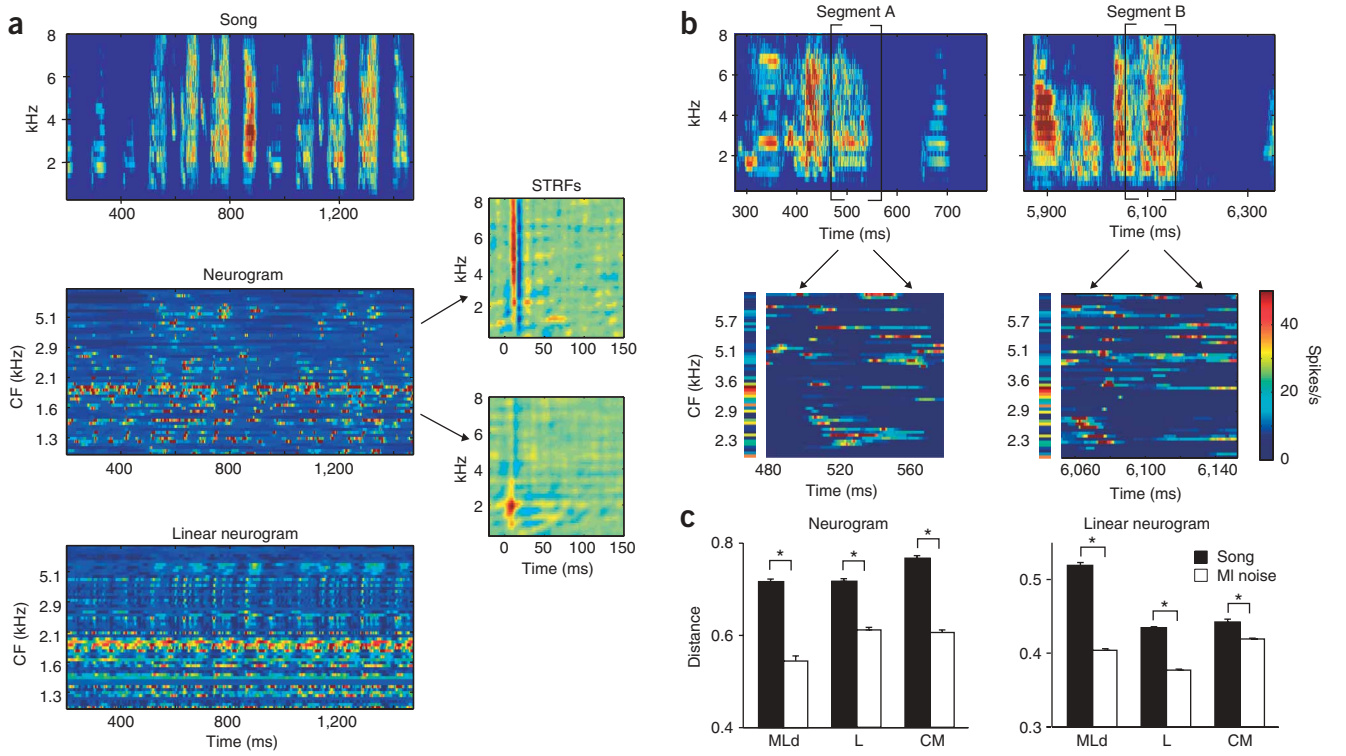


Figure 6 Neural discrimination of natural versus synthetic sounds. (**a**) A song spectrogram is shown above a neurogram, showing the response of each neuron in MLd to that song. The neurogram is a color-coded matrix of instantaneous neural responses (color indicates spike rate) across time (*x*-axis) and cell number (*y*-axis). The instantaneous response is obtained by averaging the PSTH with a time-varying Gaussian window. The cells are organized according to their characteristic frequencies (CFs), as shown on the *y*-axis. The STRF for each cell was used to obtain a predicted response shown as the linear neurogram. (**b**) The neural discrimination for two randomly chosen sound segments A and B (top) was calculated by measuring the normalized Euclidian distance between two rows in the neurogram. This distance is normalized by the neural noise, which is the variance in the response of each neuron (see Methods). The color strip to the left of each neurogram shows the relative variance for each neuron in this neuronal ensemble. An average normalized distance is obtained by averaging across all rows (cells). This neural discrimination measure is called the neurogram distance. (**c**) Average neurogram distances between segments of song and modulation-limited noise for MLd, field L and CM. The neurogram distances are calculated for the overall response and the linear fraction of the response obtained using the STRFs. Error bars: s.e.m. *, $P < 0.001$.

studies; vocalizations are perceived and understood in widely varying environments and with significant information removed[13,15,22]. Our findings indicate that the discrimination of natural sounds through linear spectro-temporal modulation tuning promotes efficient coding in four ways: (i) tuning in midbrain and forebrain neurons selectively filter spectral modulations such that only the lower spectral modulation frequencies that typify natural sounds are encoded, (ii) modulation tuning attenuates low spectro-temporal modulation frequencies that are redundant across natural sounds, (iii) temporal modulation tuning overlaps in its region of maximal gain sensitivity (slope) with the temporal modulation frequencies that vary the most among the classes of natural sounds that we examined and (iv) the ensemble temporal tuning whitens the temporal modulation power function, increasing the bandwidth of the neural response to signals from within a natural sound class. Although the data do not match the theoretical principles perfectly, the linear modulation tuning showed several efficient coding patterns that seem to facilitate the encoding of natural sounds. The two analyses that we used, comparing neural discrimination using the real tuning and a model of tuning that lacked the band-pass temporal tuning gain and using distance measures to quantify neural discrimination, support the suggestion that the observed tuning is beneficial for discriminating among natural sounds. The modulation tuning we found in zebra finches is similar to that found in the inferior colliculus (IC) and auditory cortex of cats[14,23]. In mammals and birds, spectral modulation tuning is low-pass and temporal modulation tuning is band-pass. This ensemble tuning property may therefore be a general property of vertebrate auditory systems and, as shown here, reflect a coding strategy for the efficient representation of spectro-temporal modulations found in natural sounds.

Discrepancies between the theoretical principles of efficient sensory coding and these data exist. Our analysis of natural sounds suggests that sounds defined by low temporal and high spectral modulation frequencies could be informative for the discrimination of different vocalizations. Such sounds are the tonal aspects of vocalizations (for example, harmonic stacks in zebra finch song and formants in speech). Although such sounds could be highly informative, we found that they were poorly sampled by the linear ensemble tuning. It is possible that the non-linear fraction of the response, which is not captured by the STRF, could be sensitive to these modulations. Non-linear tuning for higher spectral and temporal modulations has been observed in the cat inferior colliculus[14] and in the primate auditory cortex[24,25]. And our findings suggest that the non-linear fractions of responses are useful for the discrimination of natural sounds (**Supplementary Figs. 1** and **2**). Another possibility is that, in zebra finches, the discrimination of songs relies more heavily on temporal than on spectral structure. Behavioral studies suggest that temporal cues are highly informative in recognizing

vocalizations[13,15]. Most zebra finch auditory neurons show strong onset characteristics and therefore encode temporally modulated signals well[26–28].

The finding that field L neurons are more complexly tuned than are MLd neurons raises the question of how that tuning complexity develops. MLd cells show less tuning for spectral modulations and more precise tuning for temporal modulations than do field L cells. This corresponds well with what is known of these cells from other studies. MLd cells are highly responsive to temporally complex sounds, show strong onset responses and respond to white noise and tones with a high degree of temporal accuracy[26,27]. Cells in field L are more selective and complexly tuned than in MLd[28–30]. Receptive field properties typically increase in complexity as sensory information progresses away from the periphery. Most well-known is the transition from geniculate to simple and complex receptive fields in the visual cortex[31]. Here, also, it is likely that convergent inputs from more simply tuned thalamic cells and other regions of field L create complex auditory receptive fields. The auditory system contains strong descending projections, which may have a role in the development of tuning complexity[32].

We have previously shown that natural power and phase in the modulation spectrum of song is important for information coding in the auditory system[21]. Preserving the natural distribution of envelope amplitude[33,34] or frequency power spectrum[35] also leads to higher information rates. This research indicates that neural responses to multiple stimulus parameters yield efficient representations of natural sounds. Here, we have offered a mechanism for this increase in information coding by demonstrating how selective neural tuning for spectro-temporal modulations is advantageous for the encoding of natural sounds.

METHODS

Stimuli. The modulation-limited noise was synthesized by combining 100 ripples. Ripples are broadband sounds that are the auditory equivalent of sinusoidal gratings; their amplitude envelopes oscillate sinusoidally across the spectral and temporal domains (**Fig. 1a,c**)[8,35]. We sampled a uniform distribution of ripples in a spectro-temporal rectangle that covered the range of modulations found in song: temporal modulations <50 Hz; spectral modulations <2 cyc/kHz[21]. Modulation-limited noise stimuli were 2 s in duration. Two-second samples of song from forty zebra finches were recorded using standard procedures, band-pass frequency filtered at 250 and 8,000 Hz and stored as .wav files. Peak power was balanced between stimuli. Stimuli were presented at 70 dB sound pressure level (SPL; peak).

The modulation power spectra (MPS) for song and modulation-limited noise were calculated by taking the two-dimensional Fourier transform of the auto-correlation matrix of the sound spectrogram[8]. A jack-knifing procedure yielded the s.d. of the MPS. The coefficient of variation (c.v.) is the s.d. divided by the mean. A time window of $\pm$ 300 ms was used to calculate the stimulus auto-correlation and to estimate the cross-correlation between the stimulus and the responses in the estimation of the STRF. We used the c.v. to quantify the MPS variability because, given neural noise proportional to the MPS power, the c.v. shows the discriminating modulations that yield the higher signal to noise ratio in the modulation tuned neurons. Speech samples were 20 randomly selected sentences from a speech test library[36]. Environmental sounds included rustling leaves, fire, rain and rushing water recorded by M. Lewicki (Carnegie Mellon University)[37].

Electrophysiology. *In vivo* electrophysiological recordings were obtained using standard extracellular recording procedures from urethane-anesthetized adult male zebra finches. Stimuli were presented free-field in a sound-attenuation chamber. Only single units that responded to both stimuli were included. Ten trials of ten modulation-limited noise samples and 20 zebra finch song samples were presented. Spikes were collected from up to two brain regions simultaneously at a sampling resolution of 1 ms. Responses from 91 MLd cells, 147 field L cells and 35 CM cells were obtained. Recording locations were confirmed by identifying electrolytic lesions using standard histological procedures.

STRFs and MTFs. The linear modulation tuning of individual neurons was measured by first calculating a spectro-temporal receptive field (STRF), a linear model of the neuron's time-frequency response[16,18,19,35]. The linear fraction of a neuron's response is given by convolving the neuron's STRF with the spectrogram of the stimulus. The STRF was calculated using a generalized reverse correlation method, in which the averaged spectrogram of the sound preceding each spike is normalized by the auto-correlations in the stimulus[18,19]. Regularization and cross-validation techniques were also used to prevent over-fitting and to estimate the reliability of the STRF. The algorithms are available at http://strfpak.berkeley.edu. The modulation transfer function (MTF) for a neuron was calculated by taking the modulus of the two-dimensional Fourier transform of the STRF. The MTF shows the tuning gain of the neuron as a function of the spectral and temporal modulations present in the sound. Ensemble modulation tuning for all the neurons in one brain region was measured by averaging the MTFs for all cells to get a single ensemble modulation transfer function (eMTF).

Analysis of tuning gain and stimulus power. The spectral and temporal modulation filters were calculated by projecting the eMTFs into single vectors. For the spectral modulation filter, gain was summed across all temporal frequencies at each spectral frequency to measure the total gain at each spectral frequency. For the temporal modulation filter, gain was summed across all spectral frequencies to measure the total gain per temporal frequency. The procedure was also applied to the MPSs of the three natural sounds to compare the power distribution of the tuned response with the stimulus power distribution such that a match or mismatch between tuning and stimulus could be determined. To quantify the relationships between tuning and the modulations in zebra finch song, we calculated the correlation coefficients between the eMTFs for MLd, field L and CM and the MPS for song. Furthermore, we calculated the correlation coefficients between the spectral modulation filters and the spectral modulations in song. The process was repeated for the temporal modulation filters and the temporal modulations in song.

The common modulations across natural sounds were defined as the modulations present in high power across the three sound classes. To define these, we increased the percentage of power encompassed by contour lines until the contours for the three sound classes diverged at $\sim$50% of the total power in each MPS. We therefore defined modulation frequencies as high-power and shared across sound classes if they fell within the contours defining 50% of the total power of the sound (**Fig. 3d**, inner contour lines). The modulations falling within these contour lines were extracted from the MPSs and combined such that modulations contained within the 50% contour of all three sounds were retained, and those that were not common across the extracted portions of the MPSs were discarded. The result was an MPS showing the high power modulation overlap among all the natural sounds (**Fig. 4a**).

Analysis of ensemble STRFs and predictions. To quantify the effects of ensemble linear tuning on the discrimination of natural sounds, we generated an ensemble STRF and ensemble predictions. The ensemble STRF was obtained by taking the inverse two-dimensional fast Fourier transform of the eMTF after setting the phase to zero (cosine phase). This STRF was centered at $f = 4$ kHz, in the center of the hearing range of zebra finches. Predicted ensemble responses were then obtained by convolving this ensemble STRF with the spectrograms of the natural sounds. To compare the observed band-pass tuning with the null hypothesis, we generated a model eMTF that had similar bounds in its tuning but had low-pass temporal gain instead of band-pass temporal gain. This low-pass eMTF was modeled with a two-dimensional Gaussian gain function, with the peak centered at the origin (0 Hz, 0 cyc/kHz) and with standard deviation parameters of 50 Hz and 0.2 cyc/kHz. The outputs for these two predicted ensemble responses (from the real eMTF and model eMTF) were scaled to have equal mean (5 spikes/s) and power (s.d. 1 spike/s).

Normalized distance measures of discrimination. The discriminability of a response to two signals depends on the mean separation of the responses and the spread, or noise, in the response. Discriminability is improved by increasing the separation (stronger signal) and/or by decreasing the spread (noise). We

quantified discriminability using a normalized distance measure similar to the signal detection measure of d'. d' is equal to two times the difference in means between signals divided by the s.d. of the noise. The normalized distance, D', is similar but uses the Euclidean distance between the time-varying responses to two sound segments.

$$D' = 2 \cdot \sqrt{\sum_{t=1}^{N_{\mathrm{TBin}}} \frac{(R_t(A) - R_t(B))^2}{\sigma_c^2}}$$

$R_t(A)$ and $R_t(B)$ are the instantaneous ensemble neural responses for segments A and B, respectively, measured in spikes/s. Segment lengths were 100 ms and the time bins were 1 ms ($N_{\mathrm{TBin}} = 100$). The s.d. of the noise was taken to be 1 for the observed and modeled response.

To quantify the actual neural discriminability obtained from the neurograms (**Fig. 6**), we estimated the average D' for all single cells in one brain region. The s.d. of the noise was estimated directly from the data by comparing the difference between mean responses from post-stimulus time histograms (PSTHs) obtained from half of the total number of trials. In this case, the average measure of discriminability is

$$D' = 2 \cdot \frac{1}{N_{\mathrm{Cells}}} \sum_{c=1}^{N_{\mathrm{Cells}}} \sqrt{\sum_{t=1}^{N_{\mathrm{TBin}}} \frac{(R_t(A) - R_t(B))^2}{\sigma_c^2}}$$

The neural responses are vectors over time (t) and cell number (c). A time bin of 1 ms was used and the length of the segments was 100 ms. The time-varying neural response was obtained using a variable width Gaussian smoothing[21]. The variance in a neuron's response, σ_c^2, is given by the difference between a single response and the average time-varying firing rate. This neural noise was estimated by dividing the PSTH of ten trials into halves and obtaining, by extrapolation to large number of trials, an unbiased measure of noise[38]. The average discriminability for an ensemble of sounds and for a brain region was determined by repeating this calculation for ~100,000 random sound segment pairs, for each stimulus type. The discrimination measure correlated only weakly with spectrographic distances and normalizing by the spectrogram distance did not change the results significantly (**Supplementary Figure 2**).

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank P. Gill for discussion and assistance in the data analysis. We thank J. Mazer for insightful comments on an earlier version of this manuscript. This work was supported by US National Institute of Deafness and Communication Disorders grants to S.M.N.W. and F.E.T. and US National Institute of Mental Health grants to F.E.T. and T.E.F.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kuhl, P.K. & Meltzoff, A.N. The bimodal perception of speech in infancy. *Science* **218**, 1138–1141 (1982).
2. Marler, P. & Peters, S. Selective vocal learning in a sparrow. *Science* **198**, 519–521 (1977).
3. Braaten, R.F. & Reynolds, K. Auditory preference for conspecific song in isolation-reared zebra finches. *Anim. Behav.* **58**, 105–111 (1999).
4. Doupe, A.J. & Kuhl, P.K. Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* **22**, 567–631 (1999).
5. Suga, N. Functional properties of auditory neurones in the cortex of echo-locating bats. *J. Physiol. (Lond.)* **181**, 671–700 (1965).
6. Pollak, G.D. & Bodenhamer, B.D. Specialized characteristics of single units in inferior colliculus of mustache bat: frequency representation, tuning, and discharge patterns. *J. Neurophysiol* **46**, 605–620 (1981).
7. Rose, G. & Capranica, R.R. Temporal selectivity in the central auditory system of the leopard frog. *Science* **219**, 1087–1089 (1983).
8. Singh, N.C. & Theunissen, F.E. Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–3411 (2003).
9. Calhoun, B.M. & Schreiner, C.E. Spectral envelope coding in cat primary auditory cortex: linear and non-linear effects of stimulus characteristics. *Eur. J. Neurosci.* **10**, 926–940 (1998).
10. Eggermont, J.J. Temporal modulation transfer functions in cat primary auditory cortex: separating stimulus effects from neural mechanisms. *J. Neurophysiol.* **87**, 305–321 (2002).
11. Creutzfeldt, O., Hellweg, F.C. & Schreiner, C. Thalamocortical transformation of responses to complex auditory stimuli. *Exp. Brain Res.* **39**, 87–104 (1980).
12. Narins, P.M. & Capranica, R.R. Neural adaptations for processing the two-note call of the Puerto Rican treefrog, *Eleutherodactylus coqui*. *Brain Behav. Evol.* **17**, 48–66 (1980).
13. Woolley, S.M. & Rubel, E.W. High-frequency auditory feedback is not required for adult song maintenance in Bengalese finches. *J. Neurosci.* **19**, 358–371 (1999).
14. Escabi, M.A. & Schreiner, C.E. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *J. Neurosci.* **22**, 4114–4131 (2002).
15. Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. *Science* **270**, 303–304 (1995).
16. Chi, T., Gao, Y., Guyton, M.C., Ru, P. & Shamma, S. Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* **106**, 2719–2732 (1999).
17. Gentner, T.Q. & Margoliash, D. Neuronal populations and single cells representing learned auditory objects. *Nature* **424**, 669–674 (2003).
18. Theunissen, F.E., Sen, K. & Doupe, A.J. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural stimuli. *J. Neurosci.* **20**, 2315–2331 (2000).
19. Theunissen, F.E. *et al.* Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network* **12**, 289–316 (2001).
20. Simoncelli, E.P. & Olshausen, B.A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
21. Hsu, A., Woolley, S.M.N., Fremouw, T.E. & Theunissen, F.E. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *J. Neurosci.* **24**, 9201–9211 (2004).
22. Assmann, P.F. Tracking and glimpsing speech in noise: Role of fundamental frequency. *J. Acoust. Soc. Am.* **100**, 2680 (1996).
23. Miller, L.M., Escabi, M.A., Read, H.L. & Schreiner, C.E. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol.* **87**, 516–527 (2002).
24. Lu, T., Liang, L. & Wang, X. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci.* **4**, 1131–1138 (2001).
25. Lu, T. & Wang, X. Information content of auditory cortical responses to time-varying acoustic stimuli. *J. Neurophysiol.* **91**, 301–313 (2004).
26. Woolley, S.M. & Casseday, J.H. Response properties of single neurons in the zebra finch auditory midbrain: response patterns, frequency coding, intensity coding, and spike latencies. *J. Neurophysiol.* **91**, 136–151 (2004).
27. Woolley, S.M. & Casseday, J.H. Processing of modulated sounds in the zebra finch auditory midbrain: responses to noise, frequency sweeps and sinusoidal amplitude modulations. *J. Neurophysiol.* **94**, 1143–1157 (2005).
28. Grace, J.A., Amin, N., Singh, N.C. & Theunissen, F.E. Selectivity for conspecific song in the zebra finch auditory forebrain. *J. Neurophysiol.* **89**, 472–487 (2003).
29. Gehr, D.D., Capsius, B., Grabner, P., Gahr, M. & Leppelsack, H.J. Functional organisation of the field-L-complex of adult male zebra finches. *Neuroreport* **10**, 375–380 (1999).
30. Sen, K., Theunissen, F.E. & Doupe, A.J. Feature analysis of natural sounds in the songbird auditory forebrain. *J. Neurophysiol.* **86**, 1445–1458 (2001).
31. Ringach, D.L. Mapping receptive fields in primary visual cortex. *J. Physiol. (Lond.)* **558**, 717–728 (2004).
32. Suga, N., Xiao, Z., Ma, X. & Ji, W. Plasticity and corticofugal modulation for hearing in adult animals. *Neuron* **36**, 9–18 (2002).
33. Machens, C.K. *et al.* Representation of acoustic communication signals by insect auditory receptor neurons. *J. Neurosci.* **21**, 3215–3227 (2001).
34. Rieke, F., Bodnar, D.A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. Biol. Sci.* **262**, 259–265 (1995).
35. Klein, D.J., Depireux, D.A., Simon, J.Z. & Shamma, S.A. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *J. Comput. Neurosci.* **9**, 85–111 (2000).
36. Tyler, R.S., Preece, J.P. & Tye-Murray, K. in *Department of Otolaryngology* (University of Iowa, Iowa City, Iowa, 1990).
37. Lewicki, M.S. Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
38. Hsu, A., Borst, A. & Theunissen, F.E. Quantifying variability in neural responses and its application for the validation of model predictions. *Network* **15**, 91–109 (2004).