Behavioral/Systems/Cognitive

Probabilistic Encoding of Vocalizations in Macaque Ventral Lateral Prefrontal Cortex

Bruno B. Averbeck[1,2] **and Lizabeth M. Romanski**[3]

[1]Center for Visual Science, Department of Brain and Cognitive Sciences and [2]Sobell Department of Motor Neuroscience and Movement Disorders, Institute of Neurology, University College London, London WC1N 3BG, United Kingdom, and [3]Department of Neurobiology and Anatomy, and Center for Navigation and Communication Sciences, University of Rochester, Rochester, New York 14627

We examined strategies for classifying macaque vocalizations into their corresponding categories, as well as whether or not there was evidence that prefrontal auditory neurons were related to this process. We found that static estimates of the spectral and temporal contrasts of the calls were not effective features for discriminating among the call classes. A hidden Markov model (HMM), however, was more effective at discriminating among the call classes, reaching a performance of almost 75% correct. Finally, we found that the responses of prefrontal auditory neurons could be predicted more effectively as linear functions of the probabilistic output of the HMM than as linear functions of the spectral features of the calls. This provides evidence that, for call recognition, the macaque auditory system likely performs dynamic processing of vocalizations, and that prefrontal auditory neurons carry a signal related to the output of this processing.

Key words: **prefrontal cortex; vocalizations; macaque; hidden Markov model; encoding; primate**

Introduction

Understanding sensory processing in the brain is a complex problem. Although representations close to sensory receptors can be reasonably well characterized by models that assume the responses are linearly related to simple features of the stimuli (Kim and Young, 1994; Pillow et al., 2005), these models fail as early as the cochlear nucleus (Nelken et al., 1997) and the inferior colliculus (Escabi and Schreiner, 2002) in the auditory system. Many laboratories are investigating the special problems associated with natural stimuli and sensory systems (Theunissen et al., 2000; Vinje and Gallant, 2000; Hsu et al., 2004; Machens et al., 2004). An ecologically important class of natural stimuli for primates is their vocalizations (Ghazanfar and Hauser, 1999, 2001; Seyfarth et al., 2005), and an extensive body of behavioral research has shown that primates, including rhesus macaques, discriminate among different calls (Dittus, 1984; Gouzoules et al., 1984; Cheney and Seyfarth, 1988; Gouzoules et al., 1998). These studies have also provided evidence that the calls provide semantic meaning (i.e., they are mapped to symbolic representations or referents) (Seyfarth et al., 1980; Macedonia and Evans, 1993; Zuberbühler et al., 1997, 1999; Hauser, 1998) (but see Owren and Rendall, 2001), making primate vocalizations an important model system for studying human language (Seyfarth et al., 2005).

How does perceptual processing map the sound pressure waveform of calls from different classes into a percept? The first step toward answering this question is to understand which features of the calls are useful for distinguishing among categories. If a feature is not useful for distinguishing among the classes of calls, it would not be a useful feature for the auditory system to encode for the purpose of call discrimination. Behavioral research suggests that primates use multiple features to discriminate among calls, including the interpulse interval in noisy calls, the overall amplitude envelope and the location of inflections in the frequency contour (Hauser et al., 1998; Le Prell and Moody, 2000; Ghazanfar et al., 2001a, 2002). Previous research on the primate auditory cortex suggests that the frequency (Barbour and Wang, 2003) or temporal contrast of sounds is being represented. Another possibility, not mutually exclusive to the others, is that the auditory system is performing a dynamic analysis of the sound, taking into account the time-varying structure of its spectral features, which can be modeled by a hidden Markov model (HMM) (Rabiner, 1989).

In this study, we began by estimating how well macaque vocalizations could be discriminated using spectral and temporal contrast, and compared these results to the performance obtained by an HMM. These analyses showed that the HMM performed better than the spectral and temporal contrast for discriminating among calls from different classes. Although this doesn't necessarily imply that the auditory system is explicitly implementing an HMM, it does imply that the HMM models reasonably well, at a formal level, the computation being performed by the auditory system during the recognition of vocalizations. After this we compared encoding models which tried to predict the time-varying response of neurons in ventral-lateral prefrontal cortex (VLPFC) based on either linear functions of the time-frequency representation of the sounds, or the time-varying

Received June 30, 2006; revised Sept. 10, 2006; accepted Sept. 10, 2006.

This work was supported by National Institute on Deafness and Other Communication Disorders Grants DC-04845 and DC-05409 (L.M.R.) and by a Schmitt foundation postdoctoral grant (B.B.A.).

Correspondence should be addressed to Dr. Bruno B. Averbeck, University College London, Institute of Neurology, Sobell Department, Queen Square, London WC1N 3BG, UK. E-mail: b.averbeck@ion.ucl.ac.uk.

DOI:10.1523/JNEUROSCI.3466-06.2006

probabilistic output of the HMM. We found that linear functions of the probabilistic HMM output were more effective at predicting the firing rates of neurons than linear functions of the spectral representations. This suggests that the computations performed by the HMM might approximate the computations performed in the brain, between the time-frequency representation present in the cochlea, and the representation in prefrontal auditory neurons.

Materials and Methods

Electrophysiological recording methods. We recorded extracellular neuronal activity from the VLPFC of two awake, behaving macaque monkeys (*Macaca mulatta*) in response to a set of species-specific vocalizations. Single and multiunit activity was recorded from chronically implanted recording chambers centered over the VLPFC auditory region (Romanski et al., 1999; Romanski and Goldman-Rakic, 2002). All surgical, behavioral and electrophysiological procedures were in accordance with National Institutes of Health guidelines and with University of Rochester Committee on Animal Resources and have been described previously (Romanski et al., 2005).

Macaque vocalizations. Monkey vocalizations were provided by M. D. Hauser (Harvard University, Boston, MA) and included a large repertoire of rhesus macaque vocalizations recorded on the island of Cayo Santiago, Puerto Rico. The vocalization type, context and caller identity of all vocalizations have been characterized. The types of vocalizations presented in the current experiment included aggressive calls (i.e., barks and pant threats), coos, copulation screams, gekkers, grunts, girneys, harmonic arches, shrill barks, submissive screams, and warbles. These vocalization categories are based on the behavioral context in which the vocalizations were emitted, as well as spectral features of the calls.

All calls were presented to the animals at their original sampling frequencies, which were between 20 and 44.1 kHz, but were subsampled to 20 kHz for the analyses performed with the HMM. Subsampling was performed using the “resample” command of Matlab. This function first low-pass filters, and then subsamples the signal, when it has to downsample. There is little information in the calls above 10 kHz, so the subsampling helps reduce the dimensionality of the calls without throwing away information. Furthermore, it is necessary to have all calls represented at the same sampling rate for comparison of frequency contrast values across calls.

Task. Neuronal activity was acquired and digitized during a passive listening task in which monkeys fixated a central point on a monitor while vocalization and nonvocalization stimuli were presented from speakers (Audix, PH5-vs), located 30 inches in front of the monkeys. Sounds were presented at 60–75 dB sound pressure level measured at the level of the monkey’s ears. Eye position was continuously monitored using either an implanted scleral search coil (one animal) or an ISCAN (Burlington, MA) infrared pupil monitoring system. The animals were required to fixate a central point for the entire trial, which included a 500 ms pretrial fixation period, the stimulus presentation, and a 500 ms poststimulus fixation period. A juice reward was delivered at the termination of the poststimulus fixation period and the fixation requirement was then released. Losing fixation at any time during the task resulted in an aborted trial. There was a 2 s intertrial interval.

Each isolated unit ($n = 301$) was tested with one of several lists of 10 vocalizations, including one vocalization from each of the 10 categories. Each cell was tested with a single list, with 9–12 repetitions of each call in a randomized block design. For the next cell isolated, the next stimulus list was used and, thus, the population of VLPFC cells was tested with a large stimulus ensemble spread over the population of cells. Results reported here are based on only the units ($n = 122$) that were significant for call type in an ANOVA ($p < 0.05$) of the firing rate of the neuron. Here we were interested in whether or not we could predict the temporal structure of the response of the neurons which responded to the vocalizations.

Analysis of spectral and temporal contrast. The spectral and temporal contrasts of each call were characterized by first calculating the spectrogram of each call. Spectrograms were calculated using a 512 point Blackman windowed discrete Fourier transform, at an interval of 100 samples. They were subsequently filtered using a symmetric two-dimensional (2D) Gaussian window with an SD of 1.3 samples. Modulation spectra were then calculated by taking the 2D Fourier transform of the spectrogram, and computing its power (Singh et al., 2003; Hsu et al., 2004). The average frequency modulation was then calculated by averaging along the temporal axis of the modulation spectra, and the average temporal modulation was calculated by averaging along the frequency axis. To collapse these average vectors into a scalar estimate of the contrast, each average was normalized so that it summed to one, and then its entropy was calculated. The entropy associated with a particular call, *c*, which is an estimate of the amount of variability in a particular dimension, is given by the following:

$$H(c) = -\sum_{i=1}^{N} p_c(i) \ln p_c(i), \tag{1}$$

where N is the number of dimensions along either the spectral or the temporal axis of the modulation spectra and $p_c(i)$ is the value of the contrast in dimension i. Because the calls were of varying lengths we zero-padded each spectrogram, to make all spectrograms the same length. The zero padding makes the temporal modulation dimension the same for all modulation spectra. Because zero-padding the spectrograms introduces smoothing in the modulation spectra, we smoothed all the modulation spectra to normalize the amount of resolution obtained across calls of different lengths. This did not have a strong effect on our classification performance, and because the spectral and temporal contrast were not found to be good measures for discrimination, any residual discriminability attributable to call length does not impact any of our scientific claims. Specifically, we tried combinations of zero padding and not zero padding, as well as using the SD of the distribution instead of the entropy, and including the squares of the predictor variables in the decoding analysis, and achieved classification performances from 27 to 40% correct. SD did not tend to work as well, and without zero padding the classification performance also decreased slightly.

Classification on the spectral and temporal contrast measures was performed using both linear discriminant analysis (LDA), and *k*-nearest neighbor classification (Duda et al., 2001). *K*-nearest neighbor classification was used because LDA is limited to using linear separating hyperplanes to separate the categories. We found the performance of LDA to be superior to that of *k*-nearest neighbors, and so we only report analyses using LDA in the results. Classification was performed using twofold cross-validation. Therefore, the set of calls was divided in half and the discriminant functions were estimated on half the data, then the decoding performance was calculated on the other half of the data. The two datasets were then switched and the analysis repeated such that all calls were classified. The number of calls in each category is given in Table 1 by the row totals.

To extend the average spectral and temporal contrast measures, we used non-negative matrix factorization (NNMF) (Lee and Seung, 2001) to extract more information from the modulation spectra. NNMF is similar to singular value decomposition (SVD) in that it is an algorithm for factoring matrices into component dimensions. However, NNMF was used instead of SVD because the calculation of entropy (Eq. 1) requires all positive values. Thus, 20 factors were extracted from each modulation spectra matrix and the entropy of each of these factors was calculated. This provided a higher dimensional representation of the calls on which to do classification. As with the results on the average spectral and temporal contrast, classification was done using LDA.

HMM. The HMM is a statistical tool for characterizing time-varying stimuli (Rabiner, 1989). The hidden part of the HMM name comes from the fact that the observations are assumed to come from the hidden states of the system. The Markov part of the name comes from the fact that the transition from one hidden state to the next depends only on the previous hidden state, not hidden states further removed in time. The model has two sets of variables that have to be fit to the data, the probability distribution of observed values associated with each hidden state and a set of transition probabilities that describe the probability of transitioning from one hidden state to another (Fig. 1).

Table 1. Classification table for best temporal and frequency entropy

	Predicted category									
Call category	AG	CO	CS	GK	GY	GT	HA	SB	SC	WB
AG	22	4	1	3	11	9	0	1	5	2
CO	1	18	1	0	3	0	3	2	3	0
CS	3	0	29	15	1	10	5	9	11	0
GK	0	0	3	7	4	5	0	3	2	1
GY	5	2	2	3	6	3	0	2	2	3
GT	3	0	1	3	2	30	2	2	4	1
HA	0	0	3	0	1	0	7	2	5	2
SB	2	0	2	2	1	4	1	4	5	0
SC	3	0	8	5	7	3	7	1	12	2
WB	4	0	0	0	0	0	0	0	0	1

Correct, 0.42. WB, Warble; CO, coo; GK, gekker; GY, girney; SB, shrill bark; CS, copulation scream; SC, submissive scream; AG, aggressive call; GT, grunts; HA, harmonic arches.

In the model, the observed variables are generated probabilistically conditioned on the state of the system, which is hidden and therefore not directly observed. In the case of the simple example shown in Figure 1*a*, the observed variables ($O(t)$) are the power at each frequency in the spectrogram at a single point in time. To simplify the exposition, we can assume that each state corresponds approximately to a pattern of power in the spectrogram. For example, the initial harmonic segment of this call could correspond to one state, whereas the frequency sweep could be represented by a couple of states, depending on the position in the sweep. Because the probability of observing a particular pattern at a single point in time in the spectrogram is only dependent on the current state, the down sweep can be represented by the same states as the up sweep (compare the fourth and sixth examples in spectrogram). The transition probabilities represent the probability of transitioning from one state to the next (Fig. 1*b*). Thus, the transition probabilities for the harmonic section at the beginning of the call reflect the fact that the call tends to stay in this state for awhile with a high self-transition [i.e., $p(s(t+1) = 1 \mid s(t) = 1) = 0.6$]. They also reflect the fact that when it does transition, it tends to transition to state 2, which is the first state of the frequency sweep. State 2 also tends to transition to state 3, and state 3 to state 4. Thus, the frequency sweep is represented by a sequence of states, and the transition between the states. The hidden states represent probable patterns in the observable variables (the spectrogram at a single time point), and the transition probabilities represent the temporal sequence of the states. In the case of a simple call, like the coo (see Fig. 2), which is harmonic and has relatively little temporal modulation, a few states will be sufficient to represent the frequency patterns, and the transition probabilities would reflect the fact that these calls tend to remain in states for a relatively long time.

All HMM analyses were implemented using the HMM Matlab toolbox developed by K. P. Murphy (University of British Columbia, Vancouver, British Columbia, Canada). Model fitting in this toolbox is done by the expectation maximization algorithm. In our model, the probability distribution of observed variables is parameterized as a multidimensional Gaussian distribution conditioned on the hidden state. Specifically,

$$p(O(t)|s(t) = i) \sim N(m_i, Q_i), \tag{2}$$

where m_i is the vector mean of the distribution, and Q_i is the covariance matrix for state *i*. A separate mean vector and covariance matrix are estimated for each hidden state. The state transition matrix, *A*, describes the probability of transitioning from one state to another and, thus, contains entries which specify $p(s(t+1) = i \mid s(t) = j)$. Once the state transition matrix and the probability distribution of the observed variables have been specified for all of the hidden states, the probability, or likelihood of any sequence of observations, given the model, can be calculated. The log of this quantity, referred to as the log likelihood, will be used in several places in the manuscript. If we write the sequence of *T* observations as follows:

$$O_{1,T} = O(1), O(2), \ldots, O(T), \tag{3}$$

and collect all of the model parameters in $V = (A, m_1, Q_1, \ldots, m_H, Q_H)$, where *H* is the number of hidden states in the model, we can write the likelihood of the entire sequence of observations for a given model as, $p(O_{1,T} \mid V)$. Correspondingly, we can consider a subset of the observations up to time *t*, as $p(O_{1,t} \mid V)$. These probabilities are estimated in the toolbox using the forward-backward algorithm (Rabiner, 1989).

The HMM that was implemented followed closely HMMs used in speech recognition (Huang et al., 2001). The first step in the implementation of the HMM was preprocessing of the vocalizations to generate a series of time-varying observables for each call. We did not use the spectrogram, but rather a cepstral representation of the sound (Fig. 1*d*), because this has been found to be more efficient. Statistically, this is closely related to the log of the spectrogram. Specifically, the time sequence of observables, $O_{1,T}$, was generated by first calculating a spectrogram, as described above. Each time slice of the spectrogram was then passed through a bank of triangular filters whose width increased with frequency according to mel frequency scaling (Huang et al., 2001). These filters overlapped such that filters to the left and right of a central filter tapered to zero at the center point of the central filter. Furthermore, the left and right endpoints of each filter were evenly distributed in mel frequency space, which is related to untransformed frequency space by $B(f) = 1125\ln(1 + f/700)$. This generated a representation similar to that seen in the auditory nerve (Evans, 1972), as well as significantly reducing the dimensionality of the observation vector, as we used a rather small number of filters (see below). The discrete cosine transform (DCT) was then computed on the log transform of the filter outputs. The DCT of the log filter outputs results in a representation known as the cepstrum (Huang et al., 2001), and the specific implementation we calculated is known as the mel-frequency cepstrum (Huang et al., 2001), because of the increasing bandwidth of the filters with increasing center frequency.

Several parameters of the HMMs had to be optimized for classification. These parameters were the number of filters in the filter bank, *F*, the number of coefficients of the cepstrum that were retained, *C*, whether or not the first and second derivative of the observable variable were included as an observable, *D*, and the number of hidden states for each class, *H*. The number of hidden states was optimized by first splitting the dataset for each class of calls in half, with half of the calls forming an estimate set and half a test set. The HMM was then fit repeatedly to the estimate data with 2–30 hidden states. Because the model-fitting procedure is subject to local minima, every model was fit 10 times on the estimate data with random initial conditions, and the fit resulting in the maximum log likelihood on the estimate data was used. The HMMs with the largest log likelihood were then applied to the test data, which had not been used to fit the model. A plot of the log likelihood of the test data versus the number of hidden states was produced. Either the maximum of this curve or the point at which the change in likelihood was <0.1% was selected as the appropriate number of hidden states for the model. Thus, a different number of hidden states was used for each class.

The remaining parameters, *F*, *C*, and *D*, were optimized by selecting the parameter values that maximized classification performance, again using cross-validation. For classification, a separate HMM is fit to each class of calls. Thus, we can calculate the probability of the observation sequence for an individual call, for each separate HMM, as $p(O_{1,T} \mid V_j)$, where *j* refers to the model parameters for class *j*. Individual calls were classified by running them through the 10 HMMs, one for each class, and calculating which HMM most probably gave rise to the call. Formally, we can represent this as follows:

$$\hat{c} = \arg\max_{j} p(O_{1,T}|V_j), \tag{4}$$

where ĉ is the class estimate for the call with the sequence of observations $O_{1,T}$ and $j = (1, \ldots, 10)$ indexes the HMM fit to each call class. When classification was performed, the dataset was split in half, and the models were estimated on the first half of the data using the number of hidden

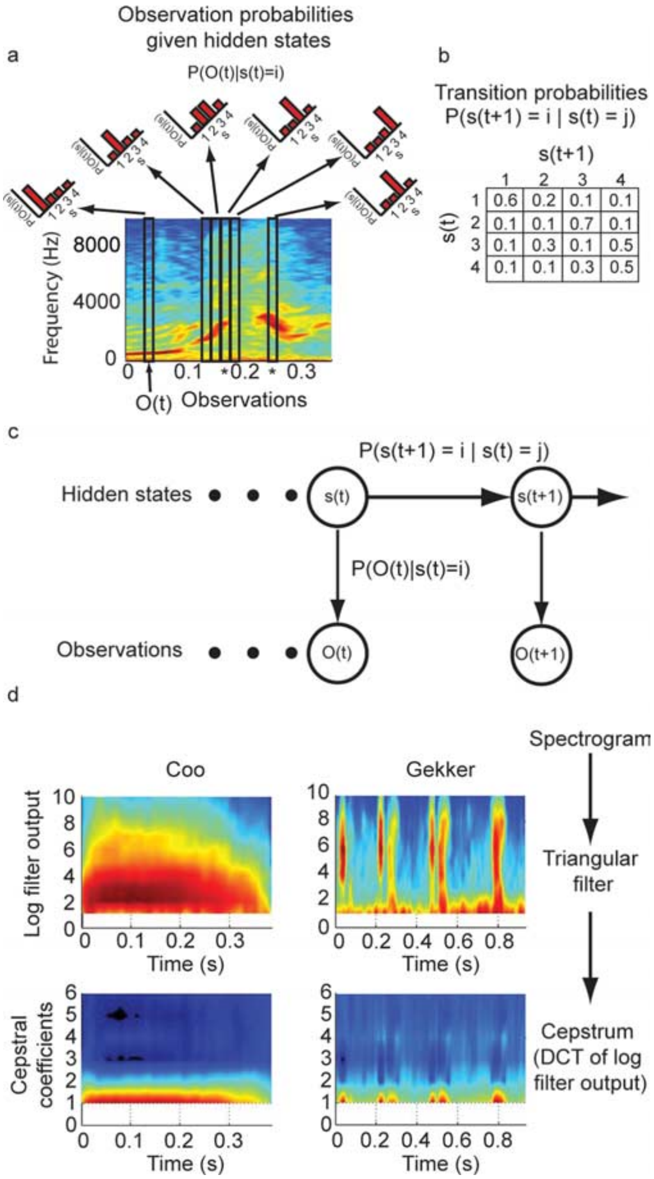

Figure 1. Hidden Markov model. ***a***, Probability of observation given hidden state. Bar graphs show the hidden state probabilities that correspond to each time slice in the spectrogram. Asterisks at the bottom of spectrogram indicate points in the up sweep and down sweep that are similar spectrally. ***b***, Transition probability matrix. This matrix shows the probability of transitioning from one state to another. ***c***, Schematic of model. Circles correspond to hidden (*s*) or observed (*O*) variables. Lines indicate statistical dependencies. Hidden states depend only on previous hidden states, and observed variables depend only on hidden states at the same point in time. ***d***, Log filter output and cepstral coefficients for two example calls. HMM was fit to the first half of the DCT coefficients.

states determined as described above. Classification was then done on the other half of the data. The test and estimate halves were then switched and the process repeated. In this way, the percent correct classification performance could be calculated.

The number of filters, F, used in the filter bank controlled the dimensionality of the observable vector. The number of filters was varied first between 8 and 32 in steps of eight. In each case, the number of hidden states was optimized as described above, and the classification performance was computed. In all cases, eight filters was optimal. To refine this estimate, we redid the analysis using 6–12 filters. In this case, 10 filters proved to be optimal, although similar performance was obtained with 6–10 filters. We also found, similar to the results for speech recognition (Huang et al., 2001), that if we retained only the first $C = F/2 + 1$ components of the cepstrum, classification was as good as when we retained all of the components. Thus, in the case of $F = 10$ filter outputs, the first $10/2 + 1 = 6$ cepstral coefficients were used, a fairly compact representation. Although the HMM analyses were performed on this reduced cepstral representation, predictions of the time-varying neural responses were performed using the full 10 cepstral components (see Fig. 6).

The final optimization that was performed was a determination of whether or not including the first and second derivatives of the observable variables would improve classification performance. These derivatives are a way to get around the Markov assumption of the model without building in higher-order transition probabilities explicitly. Because of the Markov assumption, the transition to the next hidden state is only dependent on the previous hidden state. By including the derivatives, temporal information beyond the previous time step was included in the observable variables. We found that this improved classification performance slightly. Thus, when we included the first and second derivatives and used 10 filters in our filter bank, our final observable vector was $(10/2 + 1) \times 3 = 18$ for each time slice.

Cluster and multidimensional scaling (MDS) analyses were performed on the confusion matrix (Table 2), using the cluster and mdscale functions in the Matlab statistics toolbox.

Decoding analysis using neural responses. LDA was also used to classify the single-trial responses of individual neurons with respect to the stimuli that generated them. We showed in previous work (Averbeck et al., 2003; Romanski et al., 2005) that LDA is an effective means of characterizing the information in neural responses. All classification was done on a 300 ms window of neural responses, starting 75 ms after stimulus onset. We divided this interval into a variable number of bins, and performed the classification analysis separately for each binwidth. As the bin width became smaller, the number of bins increased. If the information in the neural response is present at a coarse time scale, dividing the response into a smaller number of bins would not increase the amount of information extracted. Classification performance was estimated using two-fold cross-validation. The number of stimuli correctly classified divided by the total number of stimuli was used as our estimate of percent correct classification performance.

Predicting the firing rate in 60 ms bins. We used a linear model to predict the firing rate in 60 ms bins as a function of either the output of the HMM or the cepstrum coefficients. The following model was fit across all calls, for each neuron:

$$\hat{r}(t) = \sum_{k=0}^{K-1} \sum_{j=1}^{J} h(k, j) G(t - k - \tau, j), \qquad (5)$$

where $\hat{r}(t)$ is the estimated response of the neuron in time bin t, K is the number of lagged time bins, J (= 10) is the number of stimulus classes or cepstral coefficients, and τ is the response latency of the neuron (always 50 ms, although the results are not sensitive to small changes in this value). G is either the log-likelihood output by the set of HMMs as a function of time or the cepstrum coefficients, depending on the analysis. Because the cepstrum is a linear function of the log transform of the smoothed spectrogram, fitting a linear model to the cepstrum coefficients is equivalent to fitting a linear model to the log of the smoothed spectrogram, and as such, our model is equivalent to previous spectral-temporal receptive field (STRF) models, except that we used log-scaling for the width of our frequency filters, whereas absolute difference scaling was used previously (Theunissen et al., 2000). We varied the number of lagged time bins K between 1 and 6, and the number of bins that resulted in the best prediction was used.

Because each stimulus was presented multiple (usually 10) times, we were able to compute a poststimulus time histogram for each call for each neuron. Although the model was fit using raw neural responses on individual trials, the results reported are the fraction of the variance accounted for, normalized by the fraction of the variance accounted for by the poststimulus time histogram (PSTH). Specifically,

$$\beta = \frac{\langle (r(t) - \mu_r)^2 \rangle - \langle (r(t) - \hat{r}(t))^2 \rangle}{\langle (r(t) - \mu_r)^2 \rangle - \langle (r(t) - r_{\mathrm{PSTH}}(t))^2 \rangle}. \qquad (6)$$

Where r_{psth} is the response estimated as the average across all trials, without cross-validation, and $\hat{r}(t)$ was estimated using Eq. 5, with cross-

Table 2. Classification table for HMM

	Predicted category									
Call category	AG	CO	CS	GK	GY	GT	HA	SB	SC	WB
AG	44	3	0	1	2	8	0	0	0	0
CO	2	26	0	0	2	0	1	0	0	0
CS	0	0	52	4	12	2	1	0	11	1
GK	0	0	0	24	0	0	0	0	1	0
GY	3	4	0	0	18	1	0	1	0	1
GT	2	1	0	1	1	41	0	2	0	0
HA	0	2	1	0	0	0	14	1	2	0
SB	1	1	0	0	1	1	0	17	0	0
SC	0	0	6	4	2	0	2	0	31	3
WB	0	3	0	0	0	0	0	0	0	2

Correct, 0.74. WB, Warble; CO, coo; GK, gekker; GY, girney; SB, shrill bark; CS, copulation scream; SC, submissive scream; AG, aggressive call; GT, grunts; HA, harmonic arches.

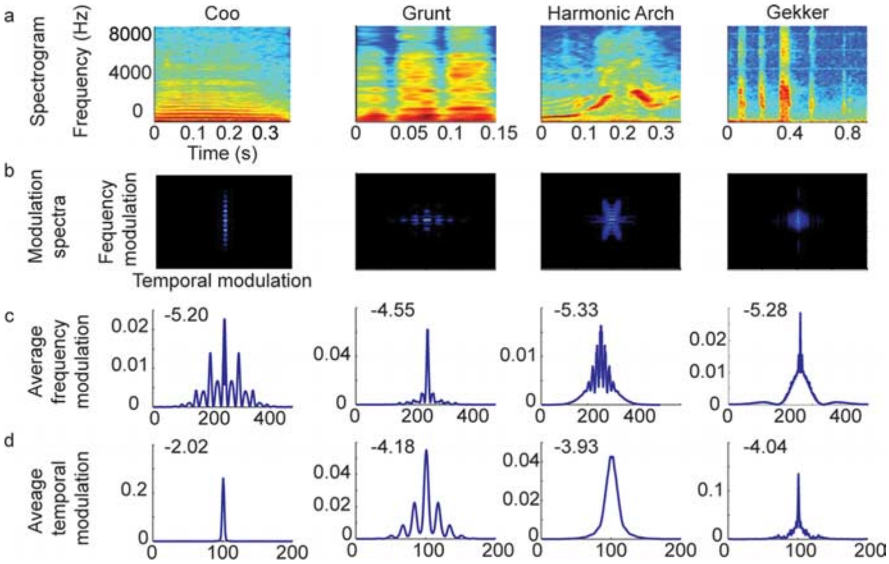

Figure 2. Characterization of spectral and temporal contrast. ***a***, Spectrograms showing time frequency representation of the sounds. ***b***, Modulation spectra computed by taking 2D Fourier transform of spectrograms. ***c***, Average frequency modulation for each call in ***a***. These are the average across the time axis of the spectrograms. ***d***, Average temporal modulation for each call in ***a***. These are the average across the frequency axis.

validation. Thus, the variance in the second term of the numerator was always estimated on the half of the dataset not used to estimate the model, whereas the variance in the denominator was estimated using all of the data. The normalization in the denominator does not affect our comparison between models, it simply allows us to compare the performance of our model to the best possible performance, which would be obtained using the PSTH. There is currently no data suggesting that noise, correlated or uncorrelated, unrelated to the average response is carrying information (Averbeck and Lee, 2006). The significance of the fit was calculated using a permutation test. The test was performed by generating 100 bootstrap datasets in which the relation between the responses, $r(t)$, and the cepstrum or the log-likelihood output by the HMM, $G(t,j)$, were shuffled. We then computed β in each of the bootstrap datasets, and estimated where the β in our original, unshuffled dataset fell in this distribution. If the original β was beyond the 99th percentile of the βs in the random distribution, we considered it a significant fit.

Finally, there is an important difference between our model fits and those of most previous researchers (except Machens et al., 2004). We did not optimize our model to get accurate estimates of the receptive field itself, we optimized our model to predict the response of the neuron. With limited data and complex models, these two goals will not arrive at the same solution.

Results

Classification of calls with static estimates of spectral and temporal contrast

We performed analyses on a set of 367 macaque vocalizations, which had been previously assigned to one of 10 categories based on the behavioral context in which they were produced as well as their acoustic features (see Materials and Methods). We began our analyses by examining features of these vocalizations that might be useful for classifying them to one of the 10 categories. The first features considered were the average spectral and temporal contrast of the calls. To perform this analysis, we measured the average spectral and temporal contrast of each call (Fig. 2) and used these features to classify the individual calls to one of the 10 categories. Although these are average measures of the spectral-temporal features of the calls, the temporal contrast characterizes the dynamics of the spectra, and therefore it could potentially capture the relevant features of the calls. We found that, based on measures of the spectral and temporal contrast, we could classify the calls into their correct categories only 37% of the time (Fig. 3*a*). This was because of the overlap in the distributions of the calls corresponding to each category (Fig. 3*c*) rather than a limitation of the linear decoding algorithm we used for classification. Thus, the average spectral and temporal contrasts alone do not provide a good basis for classifying calls to their correct class. This is not to say that there is not more information in the modulation spectra about the calls than what we extracted using the average spectral and temporal contrast, it is just not straightforward to extract this information.

The average spectral and temporal contrast measures were extracted from the modulation spectra of individual calls. These measures reflect only averages across the individual spectral and temporal dimensions of the modulation spectra and, thus, ignore much of the information that is present in the modulation spectra. To extend the analysis, we factored the modulation spectra matrices using non-negative matrix factorization (Lee and Seung, 2001) (Fig. 3*d*). This allowed us to extract much more information from the modulation spectra. We found that using additional dimensions of the modulation spectra improved our ability to classify the calls into their correct categories (Fig. 3*a*,*b*). However, we were still only able to achieve ~42% correct classification (Table 1). Interestingly, we did find that there was more information in the temporal features of the vocalizations than in the spectral features (Fig. 3*b*). This is consistent with studies of human speech recognition (Drullman et al., 1994a,b; Drullman,

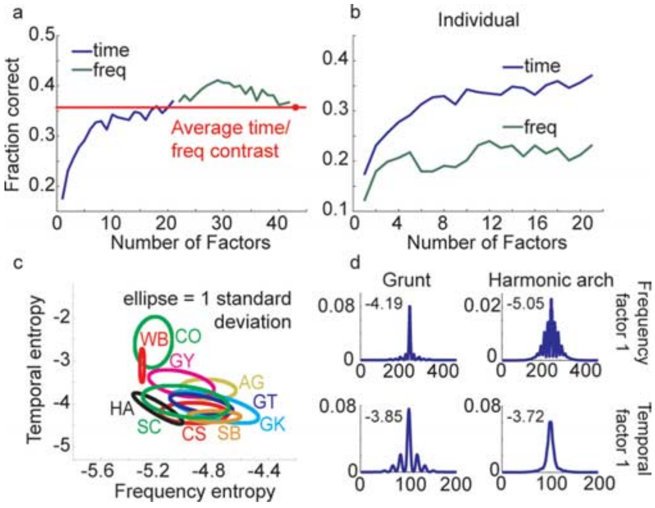

Figure 3. Classification with time and frequency contrast. ***a***, Classification with average time and frequency contrast is shown by a red line. Performance as a function of the number of factors extracted by NNMF is shown by the blue and green lines. Classification performance was assessed as factors were added to the model. Time factors were added first, followed by frequency factors. The first point in both curves is the average. ***b***, Individual performance of time and frequency factors, as a function of the number of factors. Again, the first point in both curves is the average. Subsequent points were derived by computing entropy on the NNMF factors extracted from the modulation spectra matrix. ***c***, Distribution of samples from each of the 10 categories in the average spectral and temporal contrast space. The large overlap in the distributions indicates that these features do not separate the groups well. ***d***, Example factors for two call types [grunts (GT) and harmonic arches (HA)] extracted by NNMF. The first factor was generally similar to the average. WB, Warble; CO, coo; GK, gekker; GY, girney; SB, shrill bark; CS, copulation scream; SC submissive scream; AG, aggressive call.

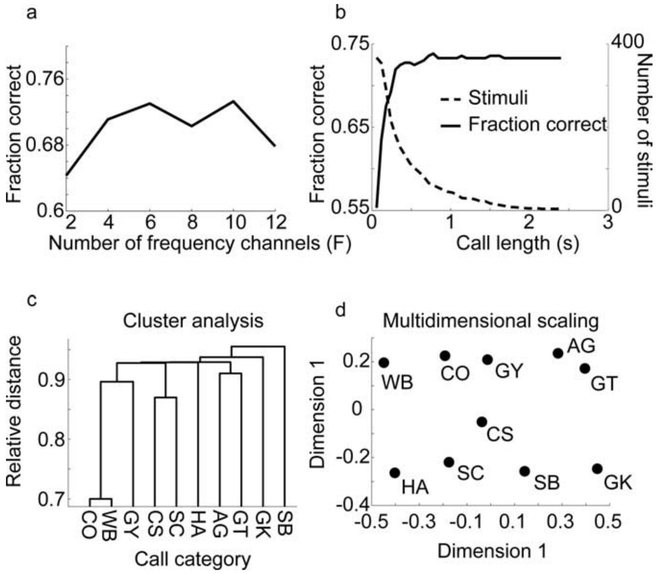

Figure 4. Classification characteristics of HMM. ***a***, Performance of HMM as a function of number of frequency channels (*F*; see Materials and Methods) used to prefilter vocalization. ***b***, Performance of HMM as a function of time; the fraction correct of calls as a function of time. This shows that as time evolves, calls are more easily discriminated, and most of the information has been extracted by a few hundred milliseconds. Also shown is the number of calls at least as long as the time indicated. ***c***, Clusters derived from the classification matrix. This plot shows that coos (COs) and warbles (WBs) were often confused, as well as submissive screams (SCs) and copulation screams (CSs). ***d***, Multidimensional scaling representation of the same data. This plot shows the categorical relationships in a continuous space. Abbreviations are the same as in Figure 3.

1995; Shannon et al., 1995) as well as monkey call recognition (Ghazanfar et al., 2002).

Classification of calls with the HMM

Having determined that static estimates of the spectral and temporal contrasts of calls were not highly effective features for classification, we turned to a hidden Markov model (Rabiner, 1989), which allowed us to model the dynamics of the spectral features of the calls more explicitly. HMMs are general purpose tools for modeling the statistics of time-varying stimuli, and they have been shown to be effective in modern speech recognition algorithms (Huang et al., 2001). As we show below, they are also useful for discriminating among macaque vocalizations. The HMM characterizes the time-varying statistics of the vocalizations using a set of hidden states and a transition probability matrix (see Materials and Methods). The hidden states correspond, in a probabilistic manner, to the spectral features of the calls in a single time slice, and the transition probability matrix models the transition from one hidden state to the next. One can think, approximately, of the different hidden states as the typical spectral patterns that occur within a class of calls. Classification was performed by first fitting a separate HMM to each of the 10 classes of calls. Each individual HMM attempts to capture the relevant spectral and temporal features of a single class of calls. Thus, the HMM works with the spectral and temporal features of the calls, as does the spectral and temporal contrast, but it does so using a much more powerful statistical characterization of the features. Individual calls were classified by passing them through each of the 10 HMMs and calculating the probability that each individual HMM gave rise to the call in question. The more closely the time-varying statistics of an individual call was predicted by the HMM from a particular class, the more probable the call was produced by that HMM, and correspondingly, the more probable that the call came from that class. Ultimately, the call is classed to the category that corresponds to the HMM that most probably generated the call, or more specifically, the HMM which gives the highest log likelihood for the call.

Before classification with the HMM was performed, the vocalizations were prefiltered into a small number of frequency channels using a filter bank (see Materials and Methods). We found that the classification performance of the HMM was relatively constant with four to 10 frequency channels, and decreased when we used only two or >10 channels (Fig. 4*a*). The rest of the results are based on the HMM with 10 frequency channels, as this was the band with the best performance. Furthermore, using 10 frequency channels allowed us to predict the neural responses using a representation with the same dimensionality as the input of the HMM, thus facilitating comparison (see below). We found that the HMM was able to classify the calls at just under 75% correct (Table 2), which is better than the classification performance achieved by static estimates of the spectral and temporal contrast. Furthermore, information on the call category was available very early during the call, with the initial performance, based on only 60 ms of the call, exceeding 50% correct (Fig. 4*b*). Investigation of the classification matrix (Table 2) and the MDS and cluster characterizations of the stimuli (Fig. 4*c*,*d*) show that the HMM was capturing the spectral-temporal features of the vocalizations obvious in the spectrograms. The cluster and MDS analyses are both derived from the confusion matrix, and they show which call categories tend to be more similar to the HMM. Thus, warbles and coos were often confused by the model, as these calls are both harmonic calls with relatively little temporal modulation. This leads to warbles and coos clustering together in the cluster analysis (Fig. 4*c*) and locating adjacent to each other in the MDS analysis (Fig. 4*d*). Gekkers and harmonic arches, however, inhabit very different clusters and locate far apart in the MDS space and, as such, they were rarely confused by the HMM algorithm. The classification matrix was relatively sparse, such that when particular call classes were confused, the confusions were often caused by heterogeneity in the call classes. For example, submissive screams have been subdivided into as many as five call classes on the basis of particular acoustic features (Gouzoules et al., 1984), with each of the five subclasses bearing some resemblance to calls of other classes that also contain these acoustic features. This shows up in the classification matrix by the fact that submissive screams are often misclassified as copulation screams or gekkers. Examinations of individual misclassified calls showed that submissive screams that were misclassified as gekkers had the staccato noisy structure characteristic of gekkers (data not shown). Some of these limitations in the classification accuracy of the HMM could be overcome by subdividing the categories into smaller subcategories, or using hierarchical HMMs, but the performance of the HMM was better than the performance of the spectral and temporal contrast, and as such, the HMM more effectively models the probabilistic features of the call classes.

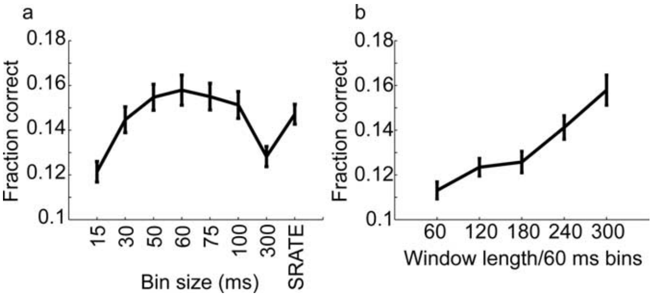

Figure 5. Classification performance of neurons as a function of bin width. ***a***, Average and SEM ($n = 122$ neurons) classification performance as a function of the bin width. All analyses were performed in the same 300 ms window. SRATE indicates classification performance achieved by computing the spike rate across a response window equal to the length of the call. ***b***, Classification performance as a function of time, at a bin width of 60 ms. Classification starts out relatively low, just above chance. As bins are accumulated in the analysis, the performance increases, as expected.

HMM and the Encoding of Vocalizations in PFC Neural Responses

Having established the HMM as a useful basis for classifying the macaque vocalizations into their appropriate categories, we wanted to test the hypothesis that the output of the HMM was, for the prefrontal cortex, similar to the output of the cochlea for brain areas early in the auditory processing stream. Specifically, we wanted to see whether linear functions of the probabilistic output of the HMM were better able to predict the responses of prefrontal auditory neurons than linear functions of the time-frequency representation, which are effective in early auditory areas. As a preliminary step in this analysis, however, we first examined the time scale on which the spike trains of PFC neurons were carrying information about the vocalizations. To do this, we used the responses of single neurons to predict which call had been played on individual trials using linear discriminant analysis (Romanski et al., 2005). Because vocalizations vary in length, we used a fixed 300 ms response window. To assess the relevant time scale, we divided this window into smaller bins which were 15–100 ms in width, and performed the decoding analysis using all of the bins simultaneously. Thus, if we used 60 ms bins, five individual bins were used for prediction. If classification improved when the 300 ms window was divided into three 100 ms bins, there was information at the 100 ms time scale that was lost when the response was averaged over the entire bin. We found that the maximum amount of information was extracted at a bin width of 60 ms (Fig. 5*a*). Thus, in the analyses that follow, we will predict the responses of the neurons using 60 ms bins. We also found that, on average, most of the information in the 300 ms window began to accumulate toward the end of the window (Fig. 5*b*).

When an individual call was processed with the set of HMMs, each HMM produced the probability as a function of time that the call came from the corresponding class. Thus, much like the spectrogram, which is a time-frequency representation of the call, processing with the set of HMMs produces a time-probability representation of the call, with one probability corresponding to each of the call classes. Furthermore, similar to an STRF, a linear probabilistic receptive field (LPRF) can be used to predict the time-varying firing rate of the neuron as a function of the output of the HMM (Fig. 6). In this case, we are predicting the response of the neuron using the HMM output or the cepstral representation based on the specific set of calls that was presented to the neuron, not the entire set of calls that were used in the call classification analyses given above. We fit LPRFs and STRFs to the responses of 122 neurons that had been shown previously to have a significant response to the vocalizations in an ANOVA (Romanski et al., 2005) ($p < 0.05$), and assessed their ability to predict the neural response in 60 ms time bins using twofold cross-validation. Thus, we had already confirmed that these neurons were carrying information about the calls in their average firing rate, and we wanted to see whether we could predict their time-varying responses.

In all cases, a single STRF and LPRF was fit across all calls. The STRF was fit using a time-frequency representation with 10 frequency channels, and the LPRF was fit to the 10 classes of calls. Thus, at each point in time, both predictors had the same degrees of freedom. We found that in 66% (50%) of the neurons the LPRF (STRF) better predicted the time-varying response than expected by chance ($p < 0.01$, permutation test), showing that these models were statistically significantly predicting the temporal profile of the response in these cases. Additionally, in 48% of the neurons the responses were significantly predicted by both the LPRF and the STRF. Thus, there was considerable overlap in the population of neurons whose responses were well predicted by the two models, and an additional 34% of neurons could not be well predicted by either model. Subsequent analyses were based on the population of 81 neurons that were significant for one of the models. The average number of bins that was best for the LPRF was 2.72 (77%, three lags or less), and the average number that was best for the STRF was 1.72 (89%, three lags or less). The number of significant lags was small because of the relatively small number of trials available for estimating the models, but at a bin width of 60 ms, three lags represent a window of integration of 180 ms.

In a few cases the LPRF provided a highly accurate prediction of the average time-varying response of individual neurons to individual calls (Fig. 7*a*). In many cases, both the LPRF and the STRF were able to provide reasonable predictions of the responses (Fig. 7*b*). The quality of fit for these example neurons is indicated in Figure 8. At the population level, the LPRF outperformed the STRF in 72% of the significant neurons (Fig. 8). A *t* test on the distribution of differences in β between the LPRF and the STRF (mean $\beta_{\mathrm{lprf}} - \beta_{\mathrm{strf}} = 0.06$) showed that it was significantly different from zero ($p < 0.05$). Thus, the time courses of the responses of more of the neurons were better predicted by the LPRF. We also examined the difference in the model performance ($\beta_{\mathrm{lprf}} - \beta_{\mathrm{strf}}$) as a function of call category (Fig. 9*a*). We found that 7/10 of the categories were better predicted by the LPRF. Furthermore, the response of the neurons to the harmonic arches and the shrill barks were predicted much better by the LPRF than by the STRF. Because the lengths of the calls differ, we were also interested in whether neurons which were better fit by one model or the other tended to have their strongest responses to particular classes of calls. We found that there was heterogeneity in the call class to which the neurons responded most strongly, and neurons that responded most strongly to different classes of calls were better fit by either the LPRF or the STRF (Fig. 9*b*). However, this heterogeneity was only marginally significantly different between the two models ($\chi^2 = 16.72$; df = 9; $p = 0.053$). Thus, neurons that responded strongly to particular classes of calls did not robustly tend to be fit better by one model or the other.

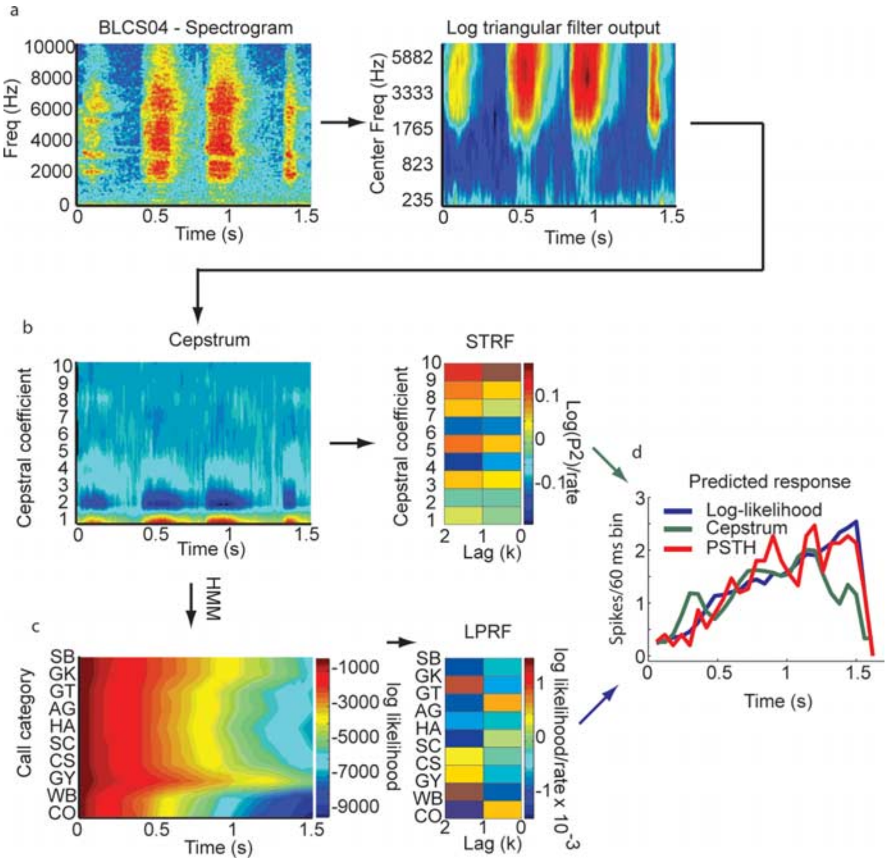

Figure 6. Processing steps in analysis. ***a***, Spectrogram and log-filter representation of example CSc. ***b***, Cepstral representation of example call and STRF estimated across all calls for this neuron. ***c***, Time-probability representation generated by processing with the 10 HMMs and LPRF for this neuron. ***d***, Estimate of response based on STRF (cepstrum) and LPRF (log-likelihood) shown along with average response (PSTH). Abbreviations are the same as in Figure 3.

Discussion

There were three main findings from our study. First, static estimates of the spectral and temporal contrasts were not highly effective features for discriminating among the call classes, and as such, if the auditory system was encoding only these features, it would not be able to discriminate well among the calls. This is not to say that these features are not important for other auditory perceptual tasks. Second, we found that the HMM was more effective at discriminating among the call classes, reaching a performance of almost 75% correct. Finally, we found that the responses of prefrontal auditory neurons could be predicted more effectively as linear functions of the probabilistic output of the HMM than as linear functions of the spectral features of the calls.

This work is complementary to other ongoing approaches we are using to identify specific features of the vocalizations that are being encoded in the responses of prefrontal auditory neurons (Averbeck and Romonski, 2004). In the present work, we have dealt only with the second-order statistics of the calls, because we computed only spectrograms and not bispectra. Using the higher-order statistics of the calls such as the bispectra, which we have analyzed in a previous study (Averbeck and Romanski, 2004), would allow us to improve our classification performance above the 74% we achieved with the HMM. However, using higher-order statistics for classification can be difficult because of the explosion in the number of dimensions that must be analyzed. A more important feature of the higher-order statistics not explored here is their ability to separate ecologically relevant stimuli, such as vocalizations, from background noise, which tends to be Gaussian (Nelken et al., 1999). Furthermore, HMMs could be fit to calls that have been filtered with only a few independent components, and the classification performance could be compared with the performance of the HMMs fit to the unfiltered calls. In this way, we could examine the hypothesis that the independent components represented features that were useful for discriminating among the classes of calls (Bartlett et al., 2002).

A number of studies have used behavioral assays and statistical analyses of the calls to examine which features of primate vocalizations the animals use to discriminate among the calls from various classes (Zoloth et al., 1979; May et al., 1988; Hauser, 1991; Hauser and Marler, 1993; Hauser et al., 1998; Le Prell and Moody, 2000; Ghazanfar et al., 2001a,b; Le Prell et al., 2002). These studies have demonstrated that, for example, the interpulse interval in noisy calls as well as the temporal direction of the calls can be relevant for call discrimination (Hauser et al., 1998; Ghazanfar et al., 2001a). Such features are well modeled by the HMMs because the duration of the interpulse interval can be modeled by how long the HMM would remain in a state related to the interpulse interval, and as such, if this interval was shortened, the HMM could detect the call as no longer coming from the appropriate class. It would be interesting to compare the performance of the HMM to the performance of animals in either laboratory or natural situations, because it is likely that animals occasionally misperceive calls. Another important question is whether or not calls that differ in meaning along important dimensions are more dissimilar. For example, it would be unfortunate and maladaptive if an animal of higher ranking confused a submissive call, for example a scream, with an aggressive call. Indeed, our analyses (Table 2) showed that none of the aggressive calls were confused with screams, and vice-versa.

In this study we have not endeavored to find a minimal set of features that allow us to discriminate among categories (Hauser,

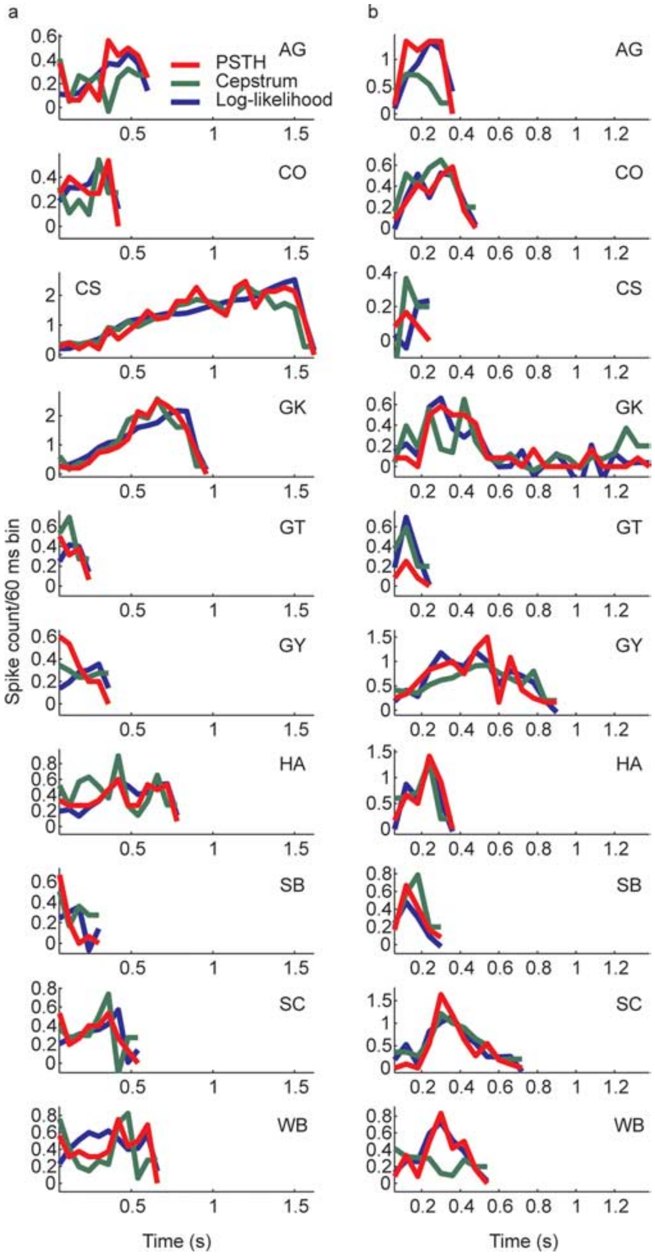

Figure 7. Predicted responses of two example neurons across all call categories. Abbreviations are the same as in Figure 3.

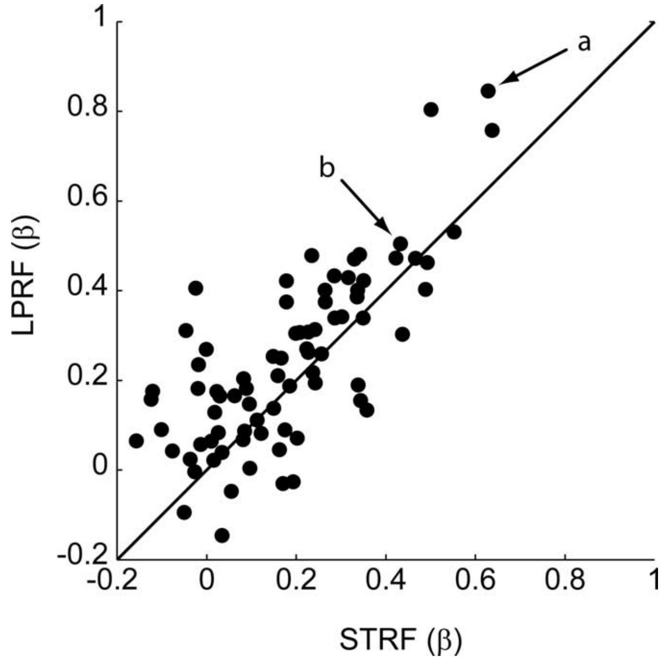

Figure 8. Fraction of variance accounted for by linear models, normalized by variance accounted for by PSTH. All analyses were done with twofold cross-validation. Only neurons whose response was significantly predicted by one of the models are shown. Arrows indicate example neurons in Figure 7, *a* and *b*.

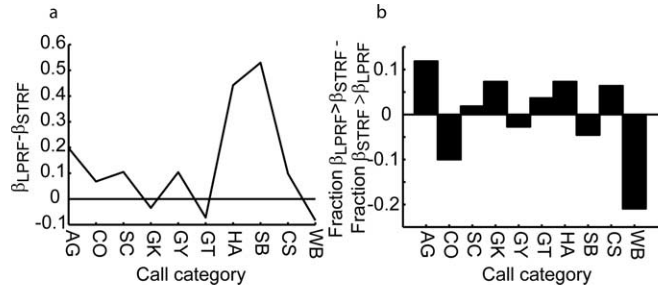

Figure 9. Relation of best model to neuron response properties. ***a***, Performance difference as a function of call category. ***b***, Histogram shows the difference in the proportion of neurons better modeled by the LPRF than the STRF as a function of call category to which the neuron responded most strongly. Neurons which responded strongly to coos, girneys, shrill barks, and warbles were better modeled by the STRF, although the effect was only marginally statistically significant ($p = 0.053$). Abbreviations are the same as in Figure 3.

1991; Hauser and Marler, 1993), although our analyses suggest that the average spectral and temporal contrast are not highly useful features. We also found that the performance of the HMM was reasonable with between four and 10 spectral channels, but using only two or $>$10 channels led to a decrease in performance. These results are consistent with studies of speech perception that have demonstrated relatively high quality speech recognition with only a few spectral channels (Drullman et al., 1994a,b; Drullman, 1995; Shannon et al., 1995). In principal, we could also filter the temporal dimension in the spectrogram, before fitting the HMM, to determine how much temporal resolution is necessary for discriminating among the calls.

Other work has examined the representation of call categories in the responses of prefrontal cortex neurons (Gifford et al., 2005). However, there are significant differences between their work and ours. First, they showed that the summed population response of the VLPFC neurons did not differentiate between harmonic arches and warbles, which are semantically similar, but that the responses did differentiate between these calls and grunts, which are semantically different. In the current study, we did not group calls from different categories together based on whether or not they had similar semantic meaning. We retained the original call category information in all cases and, thus, harmonic arches and warbles were treated distinctly. Our study does not explicitly address the question of whether or not the responses to harmonic arches and warbles are different, although implicitly, we assume that the responses are different because we are trying to predict the time-varying responses of the neurons using the time-varying cepstral (STRF) or HMM (LPRF) outputs, which would be different for the two call classes. Second, we took a probabilistic approach and assumed that the responses of prefrontal neurons represented the likelihood that a particular call came from each of the classes, again maintaining the distinction between calls from different categories that have similar semantic meaning. In additional studies, Cohen et al. (2006) have shown that there is more information in the responses of single neurons about different calls within the category nonfood than within the category food, but they did not indicate the amount of information about stimuli across categories. Another important point is that categorical representations are nonlinear functions of the acoustics (Nearey, 1997), as we have shown here with the HMM model, so that distinctions between categorical and acoustic representations come down to distinctions between linear and nonlinear representations of acoustic information, which are difficult distinctions to make in the brain, because most representations are nonlinear.

We found that the responses of a large proportion of prefrontal auditory neurons could be better predicted as linear functions of the time-varying probabilities (LPRF) produced by the HMM than by STRFs. Although we were able to well predict the responses of some of the neurons using either STRFs or LPRFs, we did not build our models using, for example, white noise stimuli, and then try to predict the responses to vocalizations. If we had done this, which is not even possible for the HMM, our prediction performance would likely have been much lower. Our analyses were motivated by theoretical work which has examined ways in which probabilistic information can be encoded in neural responses (Zemel et al., 1998; Barber et al., 2003a,b; Sahani and Dayan, 2003). The STRF is commonly used at earlier levels of sensory processing (STRF) (Aertsen and Johannesma, 1981). Because the STRF predicts the responses of neurons as a linear function of the spectrogram, it is an appropriate model for neurons that are only a few synapses removed from the auditory nerve, because the auditory nerve represents sounds spectrotemporally (Galambos, 1943). However, neural representations further from the periphery will likely be further removed from the spectral representation, and their responses will be highly nonlinear functions of the spectrotemporal representation (Nelken et al., 2003). The main question is, what sort of nonlinear transformations should we be looking at?

Perception is, in general, a probabilistic process, and there is considerable evidence that perceptual processing is closely related to optimal inference (Knill and Richards, 1996; Knill, 1998; Kersten, 1999; Ernst and Banks, 2002; Jacobs, 2002; Pouget et al., 2003; Knill and Pouget, 2004). What this means is that, because of noise and variability in the environment, sensory stimuli cannot be unambiguously mapped to a sensory percept, they can only be mapped to a probability distribution over possible percepts. In the case of vocalization perception by macaques, the calls would map to a probability distribution over the possible call classes. Thus, at some stage of the neuronal processing hierarchy, a probabilistic representation of stimuli or actions should be generated, as has been shown in the visual-motor (Shadlen and Newsome, 2001), and cognitive-motor (Averbeck et al., 2006) systems. The HMM can approximate the transformation from the acoustic features to the probabilistic representation. As such, the responses of neurons which were involved in representing probabilistic information about the vocalization categories should be described as relatively simple functions of the probabilistic representation produced by the HMM, just like the responses of neurons in the auditory nerve can be reasonably well described as linear functions of the spectral-temporal representation of the sound produced by the cochlea. This is, in fact, what we found for the auditory responses of the prefrontal neurons we studied.

References

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

Averbeck BB, Lee D (2006) Effects of noise correlations on information encoding and decoding. J Neurophysiol 95:3633–3644.

Averbeck BB, Romanski LM (2004) Principal and independent components of macaque vocalizations: constructing stimuli to probe high-level sensory processing. J Neurophysiol 91:2897–2909.

Averbeck BB, Crowe DA, Chafee MV, Georgopoulos AP (2003) Neural activity in prefrontal cortex during copying geometrical shapes II. Decoding shape segments from neural ensembles. Exp Brain Res 150:142–153.

Averbeck BB, Sohn JW, Lee D (2006) Activity in prefrontal cortex during dynamic selection of action sequences. Nat Neurosci 9:276–282.

Barber MJ, Clark JW, Anderson CH (2003a) Neural representation of probabilistic information. Neural Comput 15:1843–1864.

Barber MJ, Clark JW, Anderson CH (2003b) Generating neural circuits that implement probabilistic reasoning. Phys Rev E Stat Nonlin Soft Matter Phys 68:041912.

Barbour DL, Wang X (2003) Contrast tuning in auditory cortex. Science 299:1073–1075.

Bartlett MS, Movellan JR, Sejnowski TJ (2002) Face recognition by independent component analysis. IEEE Trans Neural Net 13:1450–1464.

Cheney DL, Seyfarth RM (1988) Assessment of meaning and the detection of unreliable signals by vervet monkeys. Anim Behav 36:477–486.

Cohen YE, Hauser MD, Russ BE (2006) Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex. Biol Lett 2:261–265.

Dittus WPJ (1984) Toque macaque food calls: semantic communication concerning food distribution in the environment. Anim Behav 32:470–477.

Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97:585–592.

Drullman R, Festen JM, Plomp R (1994a) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053–1064.

Drullman R, Festen JM, Plomp R (1994b) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95:2670–2680.

Duda RO, Hart PE, Stork DG (2001) Pattern classification, Ed. 2 New York: Wiley.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429–433.

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

Evans EF (1972) The frequency response and other properties of single fibres in the guinea-pig cochlear nerve. J Physiol (Lond) 226:263–287.

Galambos R (1943) The response of single auditory-nerve fibers to acoustic stimulation. J Neurophysiol 6:39.

Ghazanfar AA, Hauser MD (1999) The neuroethology of primate vocal communication: substrates for the evolution of speech. Trends Cogn Sci 3:377–384.

Ghazanfar AA, Hauser MD (2001) The auditory behaviour of primates: a neuroethological perspective. Curr Opin Neurobiol 11:712–720.

Ghazanfar AA, Smith-Rohrberg D, Hauser MD (2001a) The role of temporal cues in rhesus monkey vocal recognition: orienting asymmetries to reversed calls. Brain Behav Evol 58:163–172.

Ghazanfar AA, Flombaum JI, Miller CT, Hauser MD (2001b) The units of perception in the antiphonal calling behavior of cotton-top tamarins (*Saguinus oedipus*): playback experiments with long calls. J Comp Physiol [A] 187:27–35.

Ghazanfar AA, Smith-Rohrberg D, Pollen AA, Hauser MD (2002) Temporal cues in the antiphonal long-calling behaviour of cottontop tamarins. Animal Behav 64:427–438.

Gifford III GW, Maclean KA, Hauser MD, Cohen YE (2005) The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. J Cogn Neurosci 17:1471–1482.

Gouzoules H, Gouzoules S, Tomaszycki M (1998) Agonistic screams and the classification of dominance relationships: are monkeys fuzzy logicians? Anim Behav 55:51–60.

Gouzoules S, Gouzoules H, Marler P (1984) Rhesus monkey (*Macaca mulatta*) screams: Representational signaling in the recruitment of agonistic aid. Anim Behav 32:183–193.

Hauser MD (1991) Sources of acoustic variation in rhesus macaque (*Macaca mulatta*) vocalizations. Ethology 89:29–46.

Hauser MD (1998) Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. Anim Behav 55:1647–1658.

Hauser MD, Marler P (1993) Food-associated calls in rhesus macaques (*Macaca mulatta*): I. Socioecological factors. Behav Ecol 4:194–205.

Hauser MD, Agnetta B, Perez C (1998) Orienting asymmetries in rhesus monkeys: the effect of time-domain changes on acoustic perception. Anim Behav 56:41–47.

Hsu A, Woolley SM, Fremouw TE, Theunissen FE, Amin N, Shaevitz SS, Fremouw T, Hauber ME, Singh NC, Sen K, Doupe AJ (2004) Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J Neurosci 24:9201–9211.

Huang X, Acero A, Hon H-W (2001) Spoken language processing. A guide to theory, algorithm, and system development. Upper Saddle River, NJ: Prentice Hall.

Jacobs RA (2002) What determines visual cue reliability? Trends Cogn Sci 6:345–350.

Kersten D (1999) High level vision as statistical inference. In: The new cognitive neurosciences (Gazzaniga MS, ed). Cambridge, MA: MIT.

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

Knill DC (1998) Surface orientation from texture: ideal observers, generic observers and the information content of texture cues. Vision Res 38:1655–1682.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719.

Knill DC, Richards W (1996) Perception as Bayesian inference. New York: Cambridge UP.

Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems (Dietterich TG, Becker S, Ghahramani Z, eds), pp 556–562. Cambridge, MA: MIT.

Le Prell CG, Moody DB (2000) Factors influencing the salience of temporal cues in the discrimination of synthetic Japanese monkey (*Macaca fuscata*) coo calls. J Exp Psychol Anim Behav Process 26:261–273.

Le Prell CG, Hauser MD, Moody DB (2002) Discrete or graded variation within rhesus monkey screams? Psychophysical experiments on classification. Anim Behav 63:47–62.

Macedonia JM, Evans CS (1993) Variation among mammalian alarm call systems and the problem of meaning in animal signals. Ethology 93:177–197.

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

May B, Moody DB, Stebbins WC (1988) The significant features of Japanese macaque coo sounds: a psychophysical study. Anim Behav 36:1432–1444.

Nearey TM (1997) Speech perception as pattern recognition. J Acoust Soc Am 101:3241–3254.

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

Nelken I, Rotman Y, Bar Yosef O (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397:154–157.

Nelken I, Fishbach A, Las L, Ulanovsky N, Farkas D (2003) Primary auditory cortex of cats: feature detection or something else? Biol Cybern 89:397–406.

Owren MJ, Rendall D (2001) Sound on the rebound: bringing form and function back to the forefront in understanding nonhuman primate vocal signalling. Evol Anthropol 10:58–71.

Pillow JW, Paninski L, Uzzell VJ, Simoncelli EP, Chichilnisky EJ (2005) Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. J Neurosci 25:11003–11013.

Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. Annu Rev Neurosci 26:381–410.

Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286.

Romanski LM, Goldman-Rakic PS (2002) An auditory domain in primate prefrontal cortex. Nat Neurosci 5:15–16.

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2:1131–1136.

Romanski LM, Averbeck BB, Diltz M (2005) Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J Neurophysiol 93:734–747.

Sahani M, Dayan P (2003) Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. Neural Comput 15:2255–2279.

Seyfarth RM, Cheney DL, Marler P (1980) Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. Science 210:801–803.

Seyfarth RM, Cheney DL, Bergman TJ (2005) Primate social cognition and the origins of language. Trends Cogn Sci 9:264–266.

Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol 86:1916–1936.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304.

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

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

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273–1276.

Zemel RS, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Comput 10:403–430.

Zoloth SR, Petersen MR, Beecher MD, Green S, Marler P, Moody DB, Stebbins W (1979) Species-specific perceptual processing of vocal sounds by monkeys. Science 204:870–873.

Zuberbühler K, Noë R, Seyfarth RM (1997) Diana monkey long-distance calls: messages for conspecifics and predators. Anim Behav 53:589–604.

Zuberbühler K, Cheney DL, Seyfarth RM (1999) Conceptual semantics in a nonhuman primate. J Comp Psychol 113:33–42.