Principal and Independent Components of Macaque Vocalizations: Constructing Stimuli to Probe High-Level Sensory Processing

Bruno B. Averbeck¹ and Lizabeth M. Romanski²

¹Center for Visual Science, Department of Brain and Cognitive Sciences, and ²Department of Neurobiology and Anatomy, and Center for Navigation and Communication Sciences, University of Rochester, Rochester, New York 14627

Submitted 14 November 2003; accepted in final form 8 January 2004

Averbeck, Bruno B. and Lizabeth M. Romanski. Principal and independent components of macaque vocalizations: constructing stimuli to probe high-level sensory processing. J Neurophysiol 91: 2897–2909, 2004; 10.1152/jn.01103.2003. Neurons in high-level sensory cortical areas respond to complex features in sensory stimuli. Feature elimination is a useful technique for studying these responses. In this approach, a complex stimulus, which evokes a neuronal response, is simplified, and if the cell responds to the reduced stimulus, it is considered selective for the remaining features. We have developed a feature-elimination technique that uses either the principal or the independent components of a stimulus to define a subset of features, to which a neuron might be sensitive. The original stimulus can be filtered using these components, resulting in a stimulus that retains only a fraction of the features present in the original. We demonstrate the use of this technique on macaque vocalizations, an important class of stimuli being used to study auditory function in awake, behaving primate experiments. We show that principal-component analysis extracts features that are closely related to the dominant Fourier components of the stimuli, often called formants in the study of speech perception. Conversely, independent-component analysis extracts features that preserve the relative phase across a set of harmonically related frequencies. We have used several statistical techniques to explore the original and filtered stimuli, as well as the components extracted by each technique. This novel approach provides a powerful method for determining the essential features within complex stimuli that activate higher-order sensory neurons.

INTRODUCTION

Receptive fields of cortical neurons become more complex along the sensory processing hierarchy. Whereas the receptive fields of neurons in primary sensory cortical areas can often be modeled well as linear transforms of the sensory stimuli (Jones and Palmer 1987), receptive fields in subsequent stages of the hierarchy are related to more complex stimulus features (Kobatake and Tanaka 1994). Often neurons within higher-order areas are selective for a particular class of stimuli. For example, it has been shown in nonhuman primates that a population of neurons in the anterior superior temporal sulcus are responsive selectively to faces (Baylis et al. 1985; Desimone et al. 1984; Gross et al. 1972), and neurons in ventral lateral prefrontal cortex (vlPFC) are responsive selectively to species-specific vocalizations (Romanski and Goldman-Rakic 2002). In correspondence with the animal experiments, human functional magnetic resonance imaging (fMRI) studies have shown activation of a specific region of extrastriate cortex, the fusiform gyrus, by face stimuli (Kanwisher et al. 1997; Puce et al. 1995), and speech stimuli have been shown to activate temporal and frontal lobe language regions more strongly than non-speech stimuli (Davis and Johnsrude 2003; Vouloumanos et al. 2001).

To understand the response properties of neurons that are selective for complex stimuli one would like to determine which features of the sensory stimulus drive the neural responses. An understanding of these features could provide important insights into the perception of complex stimuli. Although a number of studies have addressed this issue it remains problematic for several reasons. First, stimuli used to define receptive fields in primary sensory areas, including pure tones and long white noise sequences, often do not produce reliable responses in higher-order sensory areas. For example, cortical neurons located in the lateral belt auditory region are often unresponsive to pure tones but can be driven with band-passed noise and species-specific vocalizations (Rauschecker 1998a,b; Rauschecker and Tian 2000; Rauschecker et al. 1995). Furthermore, auditory neurons located in the macaque vlPFC, which respond well to human and monkey vocalizations, often show no response to short sequences of white noise or pure tones (Romanski and Goldman-Rakic 2002). Second, the selectivity of neurons in higher-order sensory areas for complex stimuli implies that the responses are strongly nonlinear functions of the sensory inputs (Bar-Yosef et al. 2002; Lau et al. 2002; Mechner et al. 2002; Rauschecker et al. 1995; Sahani and Linden 2002; Salinas et al. 2000; Tanaka et al. 1991). Therefore reverse correlation techniques (Marmarelis and Marmarelis 1978), which in practice are limited to approximating arbitrary nonlinearities using first- or second-order polynomial expansions, may not be effective in approximating the real nonlinearities. Furthermore, accurate estimation of the terms in even a second-order expansion requires stimulus sequences that are prohibitively long and problematic to use in awake, behaving animal experiments. Thus defining receptive fields, or the features to which these neurons are responsive, is particularly challenging.

An alternative approach is to carry out feature elimination. In this approach, one first searches through a set of complex stimuli to find a stimulus to which the neuron responds strongly. After the stimulus to which the neuron has the strongest response has been identified, “features” are removed from this preferred stimulus, and the effect on the neural response is assessed. If the neuron responds as strongly to the reduced stimulus, it is concluded that the neuron was responding to the...
remaining features. In the auditory system, feature elimination has been done by high- or low-pass filtering of the stimulus, or by eliminating the spectral structure, while retaining the temporal envelope of the power (Rauschecker et al. 1995). In the visual system, this has been done by replacing images with simplified versions composed of oriented bars (Tanaka et al. 1991), by high- or low-pass filtering stimuli such as faces (Rolls et al. 1987), or by finding the simplest geometrical shape that could drive a neural response (Kayaert et al. 2003).

Although feature elimination is somewhat ad hoc, behavioral studies can provide insight into the dimensions or features that are relevant to neurons in higher-order cortical areas. Assuming that as we move up the sensory hierarchy the responses of neurons are more closely related to reportable percepts (Crick and Koch 1995; Sheinberg and Logothetis 1997), we might expect auditory neurons in vlPFC to be related to the animal’s perception of communication calls. The acoustic features that are relevant in human speech perception have been studied intensely (Pickett 1999). These studies have found that the prominent spectral features of phonemes, known as formants, play an important role in speech perception (Nearey 1989). In other studies, researchers who have studied the electronic representation and transmission of speech have found that the relative phase between the frequency components in a vocalization can carry much of the information of the original sound (Oppenheim and Lim 1981).

Guided by these behavioral insights, as well as theoretical considerations described in the Discussion, we have developed a feature-elimination approach for exploring the features to which auditory neurons in the prefrontal cortex might respond. We have extracted principal components (PCs) and independent components (ICs) from a set of rhesus macaque vocalizations, which have been shown to be effective at driving neurons in the lateral belt auditory cortex (Rauschecker 1998a; Rauschecker et al. 1995; Tian et al. 2001) and vlPFC (Romanski and Goldman-Rakic 2002; Romanski et al. (2003). SFN Abstract 722.13). The PCs and ICs extracted from each stimulus allow us to define features based on the second-order (in the case of PCs) and higher-order (in the case of ICs) statistics of each vocalization. Each PC or IC corresponds to a feature of the vocalization, and due to the way the components are defined; the complete set of PCs or ICs corresponds to all of the features present in the stimuli. By selecting a subset of the components, we can create a filtered version of the vocalization, which is composed of only those features that correspond to the components retained. In this paper, our goal is to understand and illustrate the features extracted by each technique, and the statistics of the stimuli produced by filtering with subsets of the PCs and ICs. We will also illustrate the use of the approach to examine the selectivity of an auditory cortical neuron recorded in vlPFC. Examination of the features extracted by each technique will show that the PCs correspond closely to the main Fourier features of the sounds, which are related to the formants of the vocalizations. Conversely, the ICs correspond to features that preserve the relative phase across a set of frequencies (Bell and Sejnowski 1996). Because the features extracted by the 2 techniques can be characterized well, we can directly relate neural responses to PC- and IC-filtered stimuli to specific behavioral (Nearey 1989; Nossair and Zahorian 1991) and theoretical (Lewicki 2002; Linsker 1988) hypotheses. Finally, although this technique was developed and is illustrated on macaque vocalizations, it could easily be applied to visual images for studying the visual system.

METHODS

Estimating spectra and bispectra

We will use a number of signal processing techniques, along with principal-component analysis (PCA) and independent-component analysis (ICA), to examine the statistical features extracted by each technique. To begin the analyses, all calls were preprocessed by filtering and down-sampling to 20 kHz, which retained most of the information in most of the calls. Time–frequency distributions were estimated by calculating a windowed spectrogram (Cohen 1995), with a window width of 256 samples. These spectrograms were smoothed with a symmetric Gaussian window, with a SD of 5 samples.

We will also use bispectra to analyze the vocalizations. Although bispectra have not been used commonly to analyze auditory data, they provide insight into characteristics of the sounds beyond that which can be shown by spectrograms. For example, the spectrograms are only a representation of the second-order statistics of the sounds, whereas the bispectra are a measure of the third-order statistics. To understand the features to which the bispectrum is sensitive, note that when complex numbers are multiplied, as in the case of \( F(\omega_1) \) and \( F(\omega_2) \), their product has a phase equal to the sum of the phases of each component. When this is multiplied by the conjugate of the frequency equal to the sum of the frequencies, the result is the phase difference between the product of the frequencies, and the frequency equal to their sum. If the phase difference is consistent across the segments of the vocalization, the frequencies are said to be phase coupled. Therefore if the relation

\[
\phi(\omega_1) + \phi(\omega_2) - \phi(\omega_1 + \omega_2) = C
\]

is constant across samples, where \( \phi \) indicates the phase at the indicated frequency and \( C \) is a constant, there will be a peak in the bispectrum at \( \omega_1, \omega_2 \). Thus the bispectrum indicates the consistency of the phase relationship between a pair of frequencies and their sum. This fact can be used to illustrate a key difference between the spectrogram and the bispectra. As an example, consider 2 signals, both of which have power at 5, 10, and 15 Hz. The first signal has a constant relative phase between the frequencies as a function of time, whereas the second signal has a random relative phase between the frequencies as a function of time. Both signals will have the same spectrograms, but only the first signal will show a peak in the bispectra, whereas the second signal will not have any peaks in its bispectra.

The bispectra were estimated by calculating the expected value of the triple product for pairs of frequencies and their sum (Nikias and Mendel 1993). The equation for estimation of the bispectra is

\[
b_i(\omega_1, \omega_2) = E[F^*(\omega_1 + \omega_2)F(\omega_1)F(\omega_2)]
\]

where \( F \) is the Fourier transform of a segment of the vocalization, * indicates the complex conjugate, and \( E \) is the expectation operator, with the expectation taken in the complex domain. The estimate is computed by first segmenting the vocalization into overlapping pieces, calculating the Fourier transform of each segment, and then calculating the triple product within the expectation operator, on each segment of data. The complex average of this triple product is the estimated bispectrum. Usually smoothing in the 2-dimensional frequency space is also used, to reduce noise in the estimate. In the plots shown in this paper we split the vocalization into 20 overlapping segments, used a 512-point fast Fourier transform and smoothed in the frequency domain using a Parzen window.

All results presented in the paper will be presented in terms of the bicoherence, which is directly derived from the bispectrum. The
bicoherence normalizes the power at each pair of frequencies in the bispectrum, by the power in the spectrum at each frequency, and is defined as

$$bic(\omega_1, \omega_2) = \frac{bi(\omega_1, \omega_2)}{S(\omega_1) S(\omega_2)}$$

where $S(\omega)$ is the power at frequency $\omega$. The bicoherence reveals consistent phase relations between frequencies, independent of the power at those frequencies, whereas the power in the bispectrum is a function of both the phase and the power at the relevant frequencies. Therefore the bicoherence more clearly shows the higher harmonics, which are not as strong in the bispectra, because there is less power at high frequencies. As we will see in the following text, the representation extracted by the ICs is most clearly seen in the bicoherence plots.

**Principal and independent components**

Both PCA and ICA have been treated extensively in the literature (Hyvarinen et al. 2001; Johnson and Wichern 1998). Both models assume that the observed variables to be modeled are the result of linear mixing among latent variables, which can be written as

$$s = Wv$$

where $v$ refers to the latent variables, $W$ is a mixing matrix, and $s$ are the observed variables. In our case the $s$ are the samples taken from the vocalizations. The objective of both approaches is to estimate the mixing matrix $W$, and the latent variables $v$, using only measurement of the observed variables $s$. PCA extracts a mixing matrix $W$, which has orthogonal columns, and leads to latent variables that are uncorrelated. ICA attempts to discover a mixing matrix $W$, with columns that are not necessarily orthogonal, and leads to a distribution of the latent variables that is independent and non-Gaussian, with the particular distribution specified by the nonlinearity used in the estimation algorithm (Roweis and Ghahramani 1999).

Both models were fit to a data matrix compiled by selecting a filter order, which established the frequency resolution of the analysis, and building trials out of time-shifted samples from the vocalizations. We used a filter order of 512, equivalent to 25.6 ms, which resolved the important features of the vocalizations. Each row of the matrix represents a separate random variable, and each column represents a trial. Therefore each column of the data matrix consisted of 512 samples from the stimulus, and subsequent trials were extracted from the vocalization by shifting one time step to the right. Specifically, the data matrix was given by

$$s(1) \ s(2) \ ... \ s(j)$$
$$s(2) \ s(3) \ ... \ s(j + 1)$$
$$\vdots \ \vdots \ \vdots$$
$$s(512) \ s(513) \ ... \ s(j + 512)$$

where $s(j)$ is sample $j$ from the stimulus. The calls varied in length, so the total number of samples varied. For example, the grunt illustrated in Fig. 1 is relatively short, and at about 200 ms it would have about 4,000 samples.

The PCs were found using the princomp function of MATLAB (The Mathworks, Natick MA). The ICs of the vocalizations were found using the fast-fixed point algorithm of Hyvarinen et al. (Hyvarinen and Oja 1997; Hyvarinen et al. 2001), using a cubic nonlinearity, which corresponds to maximizing the kurtosis of the distribution of the ICs. (This algorithm is available on the World Wide Web.
at http://www.cis.hut.fi/projects/ica/fastica/.) We will give a brief outline of the algorithm here, following Hyvarinen and Oja (1997). The first step is the whitening of the data. This step is followed by maximization of the kurtosis of the whitened data, which for a single vector \( w \) gives one of the ICs. The kurtosis of an IC defined by the weight vector \( w \) is

\[
k(w^TS) = E[(w^TS)^4] - 3E[(w^TS)^2]^2
\]

where \( w \) is a weight vector that defines one of the ICs, \( S \) is a column of the data matrix \( S \), and the expectation is taken over all columns of the data matrix. In practice, this function has to be constrained during optimization or the weight vector \( w \) will grow without bound. Thus the constrained function to be optimized takes the form

\[
J(w) = E[(w^TS)^4] - 3E[(w^TS)^2]^2 + F(|w|^2)
\]

where \( F \) is a function of the length of the weight vector. Because the data were prewhitened, \( E = ((w^TS)^2) = |w|^2 \). Using this and taking the derivative of Eq. 7 with respect to the weight vector \( w \) gives

\[
J'(w) = E[(w^TS)^3] - 3|w|^2F'(|w|^2)w
\]

and thus a maximum or minimum of the kurtosis can be found by solving for the fixed point of Eq. 8, which gives

\[
E(w^TS)^3 - 3|w|^2w + F'(|w|^2)w = 0
\]

where \( a \) is a scalar. In the algorithm, the estimation of \( w \) given by Eq. 10, is followed by explicit normalization of the weight vector

\[
w = \frac{w}{|w|}
\]

In practice, the algorithm is iterated, such that a new vector \( w \) is calculated according to Eqs. 10 and 11, until convergence.

Because the ICA algorithm is subject to local maxima, it was run 4 times on each sound, and the best run for each sound was retained, with best defined below. Also, unlike the eigenvectors and their associated eigenvalues in the PC analysis, the ICs are not generated with any predefined rank. Because our goal was to use the ICs to extract the higher-order features from the sounds, we defined a cost function to measure the quality of the ICs extracted from the sounds. Our measure assessed the amount of power in the original bispectrum, which was preserved in the IC features selected. The cost function was

\[
G = \sum_{i,j} [B_i(\alpha_i, \alpha_j) - |B^{PC}(\alpha_i, \alpha_j)|]
\]

where the vertical bars indicate absolute value, \( B_i \) is the bispectrum of the original call, and \( B^{IC}_i \) is the bispectrum of the original call with a subset of the ICs. This allowed us to assess the amount of power in the bispectrum that was preserved by each, or a subset of ICs. Because the ICs are not actually independent, the best combination of 10 components is not the 10 components that work best when assessed in isolation. Therefore to find a subset which minimized \( G \), we selected a set of ICs by first choosing the single IC that minimized \( G \) in Eq. 12, then selected a second IC that most decreased \( G \) when added to the first IC, and so forth. This is similar to a forward stepwise regression procedure. Although not optimal, this approach yielded reasonable results in a feasible amount of time. We were able to produce an IC-filtered call in about 4 days, with most of the time taken to repeatedly evaluate the cost function in the stepwise selection procedure.

**Filtering with the principal and independent components**

We had to develop a technique for filtering our original sounds using the PC and IC features. We did this by first projecting each column of our original data matrix, \( S \), given in Eq. 5, into a subspace defined by the corresponding algorithm. The subspace is defined by placing each of the retained components into the columns of a smaller matrix \( W_r \), where the number of columns of \( W_r \) is equal to the number of components retained, and the number of rows is equal to the filter order, in our case 512. In other words, \( W_r \) contains a subset of the columns of \( W \). This is equivalent to setting the value of the latent variable \( v \), given in Eq. 4, to zero, for some components. The projection is a matrix multiplication, where the data matrix given in Eq. 5 is projected into the column space of \( W_r \). This is given explicitly by

\[
S_{filtered} = W_r [W^T W_r]^{-1} W^T S
\]

After this, the original signal was reconstructed by using the method of windows (Coifman and Donoho 1995; Hyvarinen et al. 2001). The method of windows was carried out by averaging all values in \( S_{filtered} \) that corresponded to a particular time. For example, the second sample of the filtered vocalization was computed by taking the average of all the values that corresponded to \( s(2) \) from the matrix \( S_{filtered} \). (Refer to \( S \) given in Eq. 5 to see how the values of the original stimulus are copied into that matrix; the corresponding positions in \( S_{filtered} \) are averaged.) This gave us a filtered sound that contained only the features of the original sound that corresponded to each retained component. This approach eliminates edge effects that would occur if nonoverlapping portions of the vocalization were projected into the subspace. Because this is an averaging procedure, it does lead to the reduction of power in highly localized features, which often correspond to high frequencies. However, the effect is small, and the procedure has been found to be quite effective heuristically (Coifman and Donoho 1995; Hyvarinen et al. 2001). In all cases, we arbitrarily chose to retain 10 components, which provided a starting point for the analyses. Ten components corresponded to a massive reduction in dimensionality, given that 512 components were defined in both cases. However, we found that even 10 components were in many cases able to retain interesting features of the calls.

**Electrophysiological recording methods**

We recorded extracellular neuronal activity from the frontal lobe of awake, behaving macaque monkeys (Macaca mulatta) in response to auditory stimuli, which included species-specific vocalizations. Single-unit and multunit activity were recorded from chronically implanted recording chambers centered over the vIPFC auditory region.

**RESULTS**

The data set of sounds that were analyzed using PC and IC analyses consisted of a set of macaque vocalizations that were used in neurophysiology experiments to explore the nonhuman
primate auditory system (Rauschecker et al. 1995). We began by examining the spectral and bispectral statistics of the vocalizations in the frequency domain. Following that, we looked closely at the PCs and ICs extracted from an example vocalization. These components corresponded to statistically prevalent features of the call, and so they give us insight into the call’s statistical structure, as well as insight into the features of the call to which each technique is sensitive. For the final analyses, we filtered the original calls using subsets of either the PCs or ICs, and examined the similarities and differences in the second- and third-order statistics, between the original calls and their filtered counterparts.

Second- and third-order statistics of vocalizations in the frequency domain

We began by estimating the spectral and bispectral statistics of the set of vocalizations. In the first analysis, we explored the spectrograms of the vocalizations, an approach that has been applied extensively in the auditory domain. In Fig. 1, the spectrograms of 3 example vocalizations—a coo, a girney, and a grunt—are shown. The coo and the girney have the clearest higher-order harmonics, which can be seen as parallel horizontal lines in the spectrogram plots. The girney is characterized by more energy in the higher-order harmonics than the coo. The grunt, in contrast, is noisier. The harmonic structure of these calls is common in mammalian vocalizations, attributed to the anatomy and physiology of the vocal apparatus (Fant 1960). The oscillation of the larynx during the production of voiced phonemes, like the coo and the girney, produces a series of air pressure pulses similar to a sawtooth function. This sequence produces the harmonic structure, with the spacing between the harmonics controlled by the fundamental frequency of the oscillation of the larynx. This basic feature of the calls is responsible for many of the aspects of the vocalizations explored below, given that it leads to phase locking across harmonically related frequencies. The grunt is more similar to an unvoiced phoneme, and thus has a less-distinct harmonic structure.

In Fig. 2, we show bicoherence plots for the same 3 calls shown in Fig. 1. The bicoherence plots show power where there is phase coupling across frequencies (see METHODS), and therefore they are a measure of the consistency of the phase across the harmonics seen in Fig. 1. The bicoherence plots of the coo and the girney show a grid of power at all multiples of the fundamental frequency. This is due to the strong phase locking across the harmonics. Other calls have a less-regular structure. For example, the grunt shows phase locking among a number of higher frequencies, but not the regular pattern of the coo. The girneys, as a class of calls, are more variable than the coos. Some show the strong harmonic regularity evident in this example whereas others are less regular. The coos on the other hand consistently have the regular harmonic structure.

Principal and independent components of macaque vocalizations

Figure 3 shows the first 8 PC filters (i.e., those that describe the most variance), derived from the coo. For a stationary Gaussian process, and a sufficiently long filter window, the PCs would be equal to the Fourier components (Fuller 1996). We see a close correspondence to the Fourier components in the PCs of the coo. Most of the power is concentrated at a single frequency, and the PC filters are matched 90° phase-shifted (sine/cosine or derivative) pairs. Deviation between the PC filters and Fourier components is a consequence of nonstationarities in the calls, as well as the truncation of the filter length at 512, although it can be seen that each PC filter is dominated by a single frequency.

Figure 4 illustrates 8 IC filters derived from the same coo vocalization. The IC filters for this coo vocalization have a different structure than the PC filters. Instead of being localized in frequency, they show peaks in the spectrogram at all harmonic frequencies present in the coo. In fact, the power spectra of all the ICs appear to be quite similar. The interesting features of the IC filters are in the relative phases among the harmonic frequencies. In Fig. 5a a plot of the phase of the first 4 harmonics is shown, for the 10 ICs that maximized the cost function given in Eq. 12. The 1st harmonic is represented at several different phases. However, the 2nd harmonic follows the phase of the 1st harmonic. This is seen more clearly in Fig.
5B, which shows a plot of the phase difference of the first 4 harmonics. This plot shows a regular pattern across the ICs. If the phase of the 2nd harmonic is further ahead of the phase of the 1st harmonic, then the phase of the 3rd harmonic will be correspondingly farther ahead of the phase of the 2nd harmonic. Presumably this reflects a statistical regularity of the phase structure of this call, and may be attributable to invariances in the oscillations of the larynx.

In the next analysis, the original calls were filtered using either the first 10 PCs or the first 10 ICs, with the order of the ICs defined by Eq. 12. Figure 6 shows the spectrograms of the 3 example calls, after filtering with the PCs and ICs. These plots can be compared with those in Fig. 1, for the unfiltered calls. It can be seen that the PCs pull out the dominant spectral components of the calls, also known as the formants. Because the PCs are extracted by estimating the average covariance matrix of the calls, they will pull out the spectral components that are, on average, strongest. The PC-filtered call, however, loses its higher harmonics, which have little power. The call filtered with the ICs, on the other hand, pulls out some of the power in the main spectral component, but it also pulls out power at higher-order harmonics. This is related to the structure of the individual IC filters. Whereas the PC filters tend to be isolated in frequency, and so only extract specific frequencies, the IC filters can extract multiple harmonics, with each filter representing a different phase relation, as was shown in Fig. 5.

The phase relation in the filtered calls can be seen by looking at the bicoherence, shown in Fig. 7, for the PC- and IC-filtered calls. These should be compared with the bicoherence of the unfiltered calls, shown in Fig. 2. The phase locking at the frequencies related to the formants is preserved in the PC-filtered calls; however, the phase locking at the highest frequencies is gone because the power at these frequencies has been removed. However, in the case of the ICs, the phase locking across all frequencies has been retained. Figure 8 shows a small portion of the bicoherence of the unfiltered and the 2 filtered versions of the call at larger scale, for comparison. These plots show clearly that the PCs contain no features at frequencies above 1,500 Hz, whereas the ICs have a bicoherence that is similar to the unfiltered call’s bicoherence. The examples in Figs. 7 and 8 show a fundamental difference in the way the PC and IC filters extract information from the vocalizations. It can be seen that the PC representation is not blind to phase in the original call; it simply cannot detect phase regularities across all frequencies within the call, and localize
these regularities into a small number of dimensions. Therefore
the phase information is distributed in the PC space, whereas
the ICs are able to localize the phase information in a small
number of dimensions.

We also carried out the PC and IC component-filtering
technique on 2 human vocalizations. The vocal apparatus
of the macaque has some similarity to the human vocal tract, but
it lacks much of the complexity; therefore it is worth demon-
strating the generality of the filtering approaches. Figure 9
shows spectrograms for the unfiltered, the PC-filtered, and the
IC-filtered versions of the word “then.” The unfiltered call
shows spectral peaks at 90 and 180 Hz. These are present in the
PC-filtered calls, and to some extent in the IC-filtered calls.
The IC-filtered calls, however, preserve more of the high-
frequency harmonics of this example. This shows that the
filtering techniques are generally valid, and can extract similar
features from human speech sounds or monkey vocalizations.

We quantified the ability of the PC filters and the IC filters
to preserve the variance and the phase of the original calls.
Figure 10A shows the percentage of the variance retained by
the first 10 principal and independent components for a set of
macaque vocalizations, as well as the human speech sounds,
“then” and “oh.” The PCs are more effective at retaining the
variance of the original calls, in some cases retaining nearly all
the variance with only 10 components. The ICs do not preserve
the variance as well. Figure 10B illustrates the ability of the
principal and independent components to preserve the phase
information. The data plotted are a measure of the similarity in the bicoherence of the original and the filtered calls, as a function of frequency. It can be seen that the ICs are better at retaining the phase structure of the original calls, even at high harmonics, whereas the PCs are able to retain only the phase structure at low frequencies. This corresponds to the fact that the only structure in the call retained by the PCs is the low-frequency structure. Thus there is dissociation between the PCs and the ICs, in that the PCs retain more of the variance in the original call, while the ICs retain more of the phase consistency across frequencies. To some extent, this is a reflection of how the 10 PC and IC filters were selected. However, phase is distributed across the PCs, so no subset can be selected that would preserve the global phase structure. Furthermore, the PCs are optimized to concentrate the variance, over the length of the filter, of the original call, in as few dimensions as possible, and the ICs cannot do better. Therefore although the cost functions may exaggerate the differences in the 2 techniques, the individual strengths of the techniques for selecting particular features of the original calls can be clearly demonstrated (Fig. 10).

Neurophysiological responses to filtered calls

The manipulations were developed for isolating essential features of macaque vocalizations, to which auditory neurons might be responsive. As an example of the application of this approach, Fig. 11 shows neural responses to unfiltered and filtered versions of a shrill bark vocalization, as well as the spectrograms for the calls. This example should not be considered representative of the population because individual neurons in vIPFC differ widely in their response to complex auditory stimuli. The spectrograms show that the 2 filtering techniques extract different features from the calls. The neural response to the IC-filtered call is almost the same as the neural response to the unfiltered call, whereas the neural response to the PC-filtered call is smaller. This is the case for this particular single unit, even though the PCs conserve more of the total variance of the original call (see Fig. 10A, Shrill Bark).

DISCUSSION

We have shown that principal and independent components can be used to extract statistical features from a set of macaque vocalizations. These features not only provide insights into the statistical structure of the calls, but are also useful for generating filtered vocalizations that retain subsets of the features of the original vocalizations. We will consider the relation of these techniques to experimental and behavioral findings as well as theoretical models that have been used to study sensory cortical processing.

Behavioral studies in audition as well as vision, have explored the ability of subjects to make perceptual discriminations of stimuli that have been altered to preserve either their phase (Oppenheim and Lim 1981) or their spectral power (Nearay 1997). As we have shown above, the sounds generated by filtering with the PCs retain the prominent spectral features of the unfiltered sounds, which are often referred to as the formants (Fant 1960), in human language studies. Formants, defined as the dominant peaks in the power spectra of phonemes, have been shown to be strong perceptual cues to the identification of phonemes and syllables (Nearay 1989, 1997). Therefore use of PC-filtered vocalizations in auditory neurophysiology can help determine whether formants, which appear to be important behaviorally for sound recognition, are the essential features driving vocalization-responsive neurons. A related approach, which has been used in auditory neurophysiology, involves replacing a sound with pure tones at the
dominant frequencies (Bar-Yosef et al. 2002; who also used a number of other manipulations). However, this approach also disrupts the relative phase between the formants. Although the PCs do not preserve the global phase structure of the vocalizations, they do preserve the relative phase of the frequencies that they extract. Therefore they can be used to test the hypothesis that it is not only the power at the formants, but the relative phase at the formants that is important.

Another feature that might be important for auditory neurons is the relative phase between frequencies. In visual processing, phase has been shown to be important for object recognition (Oppenheim and Lim 1981; Piotrowski and Campbell 1982), although only a crude phase resolution was found to be important. We have shown that the relative phase structure across frequencies can be preserved by the ICs. Neural responses in cortical auditory areas, including A1, are often nonlinear (Rauschecker et al. 1995; Sahani and Linden 2002) and therefore can be sensitive to the higher-order statistical structure of the vocalizations. The ICs are sensitive to this structure as well, as can be seen by their ability to preserve third-order correlations, measured with the bicoherence, in the vocalizations. This makes ICs plausible candidates for higher-order neural representations of these sounds.

It is important to point out that the PC- and IC-filtered vocalizations differ considerably from vocalizations that could be produced by either scrambling the phase, or whitening the

**FIG. 7.** Bicoherence plots for the same filtered calls shown in Fig. 6. PC-filtered calls retain the relative phase information between the low frequencies, but do not retain the phase structure of the higher frequencies. ICs retain the phase structure across frequencies. These plots should be compared with the bicoherence plots shown in Fig. 2.
power spectra while retaining the relative phase. These manipulations would test a more specific hypothesis, that it was only the phase, or only the power spectrum, that was essential. As we have shown, the PCs do not scramble the phase, but rather retain only the relative phase at the frequencies corresponding to the PCs retained. Furthermore, the ICs do not whiten the power spectra while retaining the phase structure, given that the ICs do not introduce power at frequencies that were not present in the original stimuli. Although some aspects of these techniques are novel, others are similar to previously used approaches. For example, it is possible to design a zero-phase linear filter that would produce calls similar to those produced by the retention of a set of PCs, although this is not exactly true, given that the PCs do not correspond exactly to Fourier components. We have mainly developed the similarity between PCA and the Fourier domain to help interpretation of the features extracted by these techniques. It is more difficult to draw a direct analog between the ICs and previously used approaches.

FIG. 8. Bicoherence plots for the coo, at increased scale. Frequencies between 1,500 and 4,000 Hz are shown for the unfiltered call, and the PC- and IC-filtered calls. These plots show more clearly that the PC-filtered call has essentially no features at these frequencies, and that the IC-filtered call contains features at the same locations as the original call. Features in the IC-filtered calls generally tended to be less localized in frequency.

FIG. 9. Spectrogram and time–amplitude plots for the human speech sound “then.” As with the macaque vocalizations, it can be seen that the PCs extract the dominant spectral peaks, which show up as intense bands in the spectrograms between 200 and 400 ms. ICs extract some power at these peaks, but they also extract higher-frequency harmonics.
analyses because a linear filter cannot produce the same filtered calls that the ICs produce. Furthermore, both the PC and IC approaches provide not only a means of filtering out “interesting” features from the calls but also a means of defining, in a consistent way, what is interesting.

The PC- and IC-filtered vocalizations can also be related to reverse correlation methods used in the study of sensory processing. Linear filters, or the first-order component of the Volterra series, are not sensitive to relative phase between frequencies, but only to the amount of power in the frequencies that they pass. However, second-order filters, which also correspond to spectral temporal receptive fields (Aertsen and Johannesma 1981), can be sensitive to the relative phase between frequencies (i.e., the relative timing of power at various frequencies is as important as the amount of power). Therefore a cell’s relative sensitivity to either the preservation of the principal spectral or phase components of a call is related to the order of the nonlinearity of the cell’s receptive field.

As a final motivation for the use of these techniques to explore sensory responses, principal and independent components (Bell and Sejnowski 1997; Linsker 1988), as well as related methods known under the general heading of generative models, have been used to model sensory processing. These models have been developed under the assumption that the goal of sensory processing is to maximize the mutual information between the peripheral sensory processors and the central representations of the sensory input (Nadal and Parga 1994). These models have also been developed under the related assumption that feedback connections in sensory cortices are used to model the sensory input (Hinton and Ghahramani 1997; Mumford 1994). Previously, these approaches have been applied to data sets consisting of members of an entire class of sensory stimuli, for example, natural images or sounds. When they are applied in this way, the filters generated by these approaches have been presented as models for the receptive fields of early stages of sensory processing.

The analyses presented also provide a useful description of the higher-order statistics of the vocalizations. Bispectra as well as the bicoherence are insensitive to Gaussian noise because Gaussian signals have no cumulants beyond second order, and the bispectra can be computed from the third-order cumulant (Papoulis 1991). Therefore the features of the calls that are represented by the bispectra can be easily detected even in colored Gaussian noise. It has been shown that neural responses to background Gaussian noise can be suppressed by tones that are played on top of this noise (Nelken et al. 1999). This could be important for extracting information from the acoustic biotope, given that many of the unimportant sounds may simply add together to form a Gaussian background, attributed to the central limit theorem. As we have seen, however, nonhuman primate vocalizations contain rich higher-order structure, and therefore could be easily filtered out of the background, by focusing on the higher-order aspects of the sounds. This of course assumes that the structure evident in the bicoherence can be used to differentiate the calls, a subject of ongoing work.

Defining the essential features of a stimulus that drive neuronal responses is an important goal in sensory neurophysiology. Assuming that some neurons in higher-order cortical processing regions are selectively responsive to vocalizations, we can, in principle, restrict our search to only the dominant features of the vocalizations. PCA allows us to construct our search to the portion of the distribution of sounds spanned by the prominent second-order statistics of the vocalizations, and ICA allows us to construct our search to the portion of the distribution of sounds spanned by the prominent higher-order statistics of the sounds. By filtering the sounds with these features we can test neurons for residual responses to the sounds defined by these features. These techniques provide
powerful tools for exploring the feature space to which high-level auditory or visual sensory neurons are responsive.

ACKNOWLEDGMENTS

We thank M. D. Hauser and P. Rakic for vocalization stimuli and A. A. Ghashanfar, D. Lee, I. Nelken, and W. E. O’Neill for helpful comments on the manuscript.

REFERENCES


FIG. 11. Example of a vIPFC neuronal response to filtered and unfiltered macaque vocalizations. Note that the time axes between A and B do not correspond. A: spectrogram plots of unfiltered, PC-filtered, and IC-filtered shrill bark vocalization. B: raster and spike-density plots of single-unit activity recorded from a neuron in the vIPFC of an awake, fixating macaque presented with the corresponding sounds. Each row of tic marks in the raster plot corresponds to a single presentation of the sound. Onset of the sound is at time 0; the spike-density function is overlaid onto the raster plot. Unfiltered and IC-filtered versions of the shrill bark vocalization evoked greater responses than the PC-filtered version in this particular cell. C: mean and variance of spike count for each call type.