Differential Coding of Conspecific Vocalizations in the Ventral Auditory Cortical Stream

Makoto Fukushima, Richard C. Saunders, David A. Leopold, Mortimer Mishkin, and Bruno B. Averbeck

Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892

The mammalian auditory cortex integrates spectral and temporal acoustic features to support the perception of complex sounds, including conspecific vocalizations. Here we investigate coding of vocal stimuli in different subfields in macaque auditory cortex. We simultaneously measured auditory evoked potentials over a large swath of primary and higher order auditory cortex along the supratemporal plane in three animals chronically using high-density microelectrocorticographic arrays. To evaluate the capacity of neural activity to discriminate individual stimuli in these high-dimensional datasets, we applied a regularized multivariate classifier to evoked potentials to conspecific vocalizations. We found a gradual decrease in the level of overall classification performance along the caudal to rostral axis. Furthermore, the performance in the caudal sectors was similar across individual stimuli, whereas the performance in the rostral sectors significantly differed for different stimuli. Moreover, the information about vocalizations in the caudal sectors was similar to the information about synthetic stimuli that contained only the spectral or temporal features of the original vocalizations. In the rostral sectors, however, the classification for vocalizations was significantly better than that for the synthetic stimuli, suggesting that conjoined spectral and temporal features were necessary to explain differential coding of vocalizations in the rostral areas. We also found that this coding in the rostral sector was carried primarily in the theta frequency band of the response. These findings illustrate a progression in neural coding of conspecific vocalizations along the ventral auditory pathway.

Key words: auditory cortex; ECoG; evoked potential; LFP; monkey; multielectrode

Introduction

The functional organization of auditory cortex and subcortical nuclei underlies our effortless ability to discriminate complex sounds (King and Nelken, 2009). Vocalizations are an important class of natural sounds that are critical for conspecific communication in a wide range of animals (Doupe and Kuhl, 1999; Ghazanfar and Hauser, 2001; Petkov and Jarvis, 2012), including macaques (Hauser and Marler, 1993; Rendall et al., 1996). In primates, it is thought that information about vocalizations is extracted in a ventral auditory stream responsible for processing sound identity (Romanski et al., 1999; Rauschecker and Scott, 2009), analogous to a ventral visual stream that processes visual object identity (Ungerleider and Mishkin, 1982; Kravitz et al., 2013). The ventral auditory pathway consists of several highly interconnected subdivisions on the supratemporal plane (STP) (Romanski and Averbeck, 2009; Hackett, 2011). The auditory

response latency estimated from single-unit recordings systematically increases from caudal to rostral areas, suggesting that auditory processing progresses caudorostrally along the STP (Kikuchi et al., 2010; Scott et al., 2011).

The rostral auditory subdivisions have been shown to have selectivity for conspecific vocalizations over other classes of stimuli by both functional imaging and electrophysiological measurements including single-unit recordings and local field potentials (Poremba et al., 2004; Petkov et al., 2008; Perrodin et al., 2011). The selectivity is defined as increased responses to conspecific vocalizations, relative to other classes of stimuli. This leaves open several questions about the nature of vocal stimulus coding in these areas. First, to what extent does this selectivity reflect the capacity of neurons in this area to discriminate among distinct vocalizations? Note that selectivity for a particular class of stimuli does not necessarily result in increased discriminability among stimuli within the favored class, as illustrated in a recent functional magnetic resonance imaging (fMRI) study of macaque face patches (Furl et al., 2012). For neural responses in the rostral STP to be useful for communication, they must signal not only the issuance of a vocalization but also the unique spectral and temporal features that differentiate distinct calls. Second, how does complex auditory selectivity arise in the ventral auditory stream? How do neural populations discriminate among conspecific vocal stimuli as information flows from primary areas to higher order auditory cortex? More specifically, what is the degree to which spectral or temporal acoustic features explain neural coding of natural vocalizations in the primary and higher areas?

Received Sept. 16, 2013; revised Feb. 10, 2014; accepted Feb. 24, 2014.

Author contributions: M.F., M.M., and B.B.A. designed research; M.F., R.C.S., and B.B.A. performed research; M.F. and B.B.A. analyzed data; M.F., R.C.S., D.A.L., M.M., and B.B.A. wrote the paper.

This research was supported by the Intramural Research Program of the National Institute of Mental Health, National Institutes of Health, and Department of Health and Human Services. We thank K. King for audiologic evaluation of the monkeys' peripheral hearing, and M. Mullarkey, R. Reoli, and D. Rickrode for technical assistance. This study utilized the high-performance computational capabilities of the Helix Systems (<http://helix.nih.gov>) and the Biowulf Linux cluster (<http://biowulf.nih.gov>) at the National Institutes of Health, Bethesda, MD.

The authors declare no competing financial interests.

Correspondence should be addressed to Makoto Fukushima, Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Building 49, Room 1B80, 49 Convent Drive, Bethesda, MD 20892. E-mail: makoto_fukushima@me.com.

DOI:10.1523/JNEUROSCI.3969-13.2014

Copyright © 2014 the authors 0270-6474/14/344665-12\$15.00/0

Here we addressed these questions by monitoring neural responses to a range of macaque vocalizations with a recently developed microelectrocorticography (μ ECoG) method (Fukushima et al., 2012) that enabled high-density recording from 96 electrodes distributed along the STP. The μ ECoG arrays are well suited to examining spatiotemporal activation profiles from a large expanse of cortex with millisecond temporal resolution, although they have lower spatial resolution than is provided by single neuron recordings. Using this approach, we could evaluate simultaneously recorded electrophysiological activity along the caudal to rostral progression from primary to high-level auditory areas.

Materials and Methods

Subjects. Recordings were performed on three adult male rhesus monkeys (*Macaca mulatta*) weighing 5.5–10 kg. All procedures and animal care were conducted in accordance with the Institute of Laboratory Animal Resources Guide for the Care and Use of Laboratory Animals. All experimental procedures were approved by the National Institute of Mental Health Animal Care and Use Committee.

Multielectrode arrays and implant surgery. We used custom-designed μ ECoG arrays to record field potentials from macaque auditory cortex (NeuroNexus Technologies). The array was machine fabricated on a very thin polyimide film (20 μ m). Each array had 32 recording sites, 50 μ m in diameter, on a 4×8 grid with 1 mm spacing (i.e., 3×7 mm rectangular grid; Fukushima et al., 2012). We implanted four or five μ ECoG arrays in each of the three monkeys (Monkey M, five arrays in the right hemisphere; Monkeys B and K, four arrays each in the left hemisphere). Three arrays in each monkey were placed on the STP in a caudorostrally oriented row (Fig. 1a). The fourth array was positioned over the parabelt on the lateral surface of the superior temporal gyrus (STG) adjacent to A1. The fifth array in monkey M was placed on the lateral surface of STG just rostral to the fourth array (data recorded from the lateral-surface arrays are not reported in this paper). The implantation procedure was described in detail previously (Fukushima et al., 2012). Briefly, we removed a frontotemporal bone flap extending from the orbit ventrally toward the temporal pole and caudally behind the auditory meatus and then opened the dura to expose the lateral sulcus. The most caudal of the three μ ECoG arrays on the STP was placed first and aimed at area A1 by positioning it just caudal to an (imaginary) extension of the central sulcus and in close proximity to a small bump on the STP, both being markers of A1's approximate location. Each successively more rostral array was then placed immediately adjacent to the previous array to minimize interarray gaps. The arrays on the lateral surface of the STG were placed last. The probe connector attached to each array was temporarily attached with cyanoacrylate glue or Vetbond to the skull immediately above the cranial opening. Ceramic screws together with bone cement were used to fix the connectors to the skull. The skin was closed in anatomical layers. Post-surgical analgesics were provided as necessary, in consultation with the National Institute of Mental Health veterinarian.

Auditory stimuli. To estimate the frequency preference of each recording site, we used 180 different pure-tone stimuli (30 frequencies from 100 to 20 kHz equally spaced logarithmically, each presented at six equally spaced intensity levels from 52 to 87 dB; Fukushima et al., 2012). For the main experiment, we used 20 vocalizations (VOC) and two sets of synthetic sounds derived from VOC stimuli (envelope-preserved sound, EPS; and spectrum preserved sound, SPS). The VOC stimulus set consisted of 20 macaque vocalizations used in a previous study (Kikuchi et al., 2010; Fig. 2a). From each VOC stimulus, we derived two synthetic stimuli, one EPS stimulus and one SPS stimulus, yielding a set of 20 EPS stimuli and a set of 20 SPS stimuli (Fig. 5a).

For EPS stimuli, we first estimated the envelope of a vocalization by calculating the amplitude of the Hilbert transform from the original VOC stimulus. We then multiplied this amplitude envelope by broadband white noise to create the EPS stimulus. Therefore, all 20 EPS stimuli had the same flat spectral content. Thus, these stimuli could not be discriminated using spectral features, whereas the temporal envelopes (and thus the durations) of the original vocalizations were preserved. For SPS stimuli, we generated broadband white noise with a duration of 500 ms

and computed its Fourier transform. Then the amplitude in the Fourier domain was replaced by the average amplitude of the corresponding VOC stimulus. We then transformed back to the time domain by the inverse-Fourier transform. This created a sound waveform that preserved the average spectrum of the original vocalization with a flat temporal envelope, random phase, and a duration of 500 ms. We then imposed a 2 ms cosine rise/fall to avoid abrupt onset/offset of the sound. Therefore, all 20 SPS stimuli had nearly identical, flat temporal envelopes, such that these stimuli could not be discriminated using temporal features, while the average spectral power of the original vocalizations was preserved.

We presented these 60 stimuli in pseudorandom order with an interstimulus interval of 3 s. Each stimulus was presented 60 times. The sound pressure levels of the stimuli measured by a sound level meter (2237A; Brüel & Kjaer) ranged from 65 to 72 dB at a location close to the animal's ear.

Electrophysiological recording and stimulus presentation. During the experiment, the monkey was placed in a sound-attenuating booth (Bioacoustics Instruments). We presented the sound stimuli while the monkey sat in a primate chair and listened passively with its head fixed. We monitored the monkey's behavioral state through a video camera and microphone connected to a PC. Juice rewards were provided at short, random intervals to keep the monkeys awake. The sound stimuli were loaded digitally into an RZ2 base station (50 kHz sampling rate, 24 bit D/A; Tucker Davis Technology) and presented through a calibrated free-field speaker (Reveal 501A; Tannoy) located 50 cm in front of the animal. The auditory evoked potentials from the 128 channels of the μ ECoG array were bandpassed between 2 and 500 Hz, digitally sampled at 1500 Hz, and stored on hard-disk drives by a PZ2–128 preamplifier and the RZ2 base station.

Data analysis. MATLAB (MathWorks) was used for off-line analysis of the neural data. There was little significant auditory evoked power above 250 Hz. Therefore, we low-pass filtered and resampled the data at 500 Hz to speed up the calculations and reduce the amount of memory necessary for the analysis. The field potential data from each site was re-referenced by subtracting the average of all sites within the same array (Kellis et al., 2010). The broadband waveform was obtained by filtering the field potential between 4 and 200 Hz. For the analysis with bandpassed waveform, the field potential was bandpass filtered in the following conventionally defined frequency ranges: theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), low-gamma (30–60 Hz), and high-gamma (60–200 Hz) (Leopold et al., 2003; Edwards et al., 2005). We filtered the field potential with a Butterworth filter. This was done by convolving the field potential waveforms with a kernel, which results in a phase shift of the convolved waveform. We achieved a zero-phase shift by processing the data in both forward and reverse directions in time ("filtfilt" function in MATLAB), because the phase shift induced by filtering in the forward direction is canceled out by filtering in the reverse direction. The 96 sites on the STP were grouped based on the characteristic frequency maps obtained from the high-gamma power of the evoked response to a set of pure-tone stimuli (Fig. 1b; Fukushima et al., 2012). The four sectors were estimated to correspond to the following subdivisions within the auditory cortex: Sec (Sector) 1, A1 (primary auditory cortex)/ML (middle lateral belt); Sec 2, R (rostral core region of the auditory cortex)/AL (anterior lateral belt region of the auditory cortex); Sec 3, RTL (lateral rostrotemporal belt region of the auditory cortex); Sec 4, RTp (rostrotemporal pole area).

Classification analysis. We performed classification analysis using a linear multinomial classifier. The predictor variable for the classifier ($\mathbf{x}$) was the evoked waveform following the onset of the stimulus. This waveform was used to predict which stimulus was presented.

Classification analysis was always performed only within an individual stimulus set (e.g., the VOC stimulus trials were always analyzed separately from the EPS stimulus trials). There were 20 different stimuli to be classified within each of the three different stimulus types (VOC, EPS, and SPS). Therefore, chance performance in terms of fraction correct for each type was 0.05.

We used a multinomial regression model with a softmax-link function for estimating the posterior probability of the stimuli, given the evoked

Figure 1. Implanted locations of μ ECoG arrays and delineation of the four sectors. *a*, Lateral view of the right hemisphere reproduced from the postmortem brain of Monkey M after removing the upper bank of the lateral sulcus and frontoparietal operculum to visualize the location of the three arrays on the STP. The black rectangles with yellow borders represent the three arrays implanted on the STP. Magnified view of the STP shows locations of μ ECoG arrays. *b*, Delineation of sectors based on the tonotopic maps obtained from evoked power in the high-gamma band to pure-tone stimuli. Frequency indicates the characteristic frequency to which the site is tuned. The caudorostral location of the array is plotted on the horizontal axis.

responses. That is, the posterior probability of stimulus k given evoked response vector $\mathbf{x}$ was modeled as the following multinomial distribution:

$$P(\mathbf{Y}|\mathbf{x};\theta) = \prod_{k=1}^{20} (\mu^k)^{y_k}$$

where $\mu^k = \frac{\exp(\theta_k^T \mathbf{x})}{\sum_{l=1}^{20} \exp(\theta_l^T \mathbf{x})}$, $\mathbf{x} = (x_0, x_1, \dots, x_N)^T$, $\theta_k = (\theta_k^0, \theta_k^1, \dots, \theta_k^N)$, $\mathbf{Y} = (Y^1, Y^2, \dots, Y^{20})$, $Y^j = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases}$ for the stimulus k , $x_0 = 1$ is the dummy variable for the intercept term.

The evoked response vector ($\mathbf{x}$) for decoding from a single site is the evoked waveform from that site, and therefore N is the number of data points in the waveform. For decoding from multiple sites, $\mathbf{x}$ is the concatenated waveforms from multiple sites, and thus $N = (\text{the number of the waveform data points}) \times (\text{the number of sites in a sector})$. In addition, each element of this vector was standardized, across trials, to have zero mean and unit variance. The parameter vector θ represents the model coefficients.

Therefore, the log likelihood of the parameter vector θ is

$$l(\theta) = \sum_{m=1}^M \sum_{k=1}^{20} Y_m^k \log \mu_m^k,$$

where (Y_m^k, μ_m^k) is (Y^k, μ^k) for the m th trial in the training dataset (the total number of trials in the training dataset is M).

Parameters were estimated by maximizing the log-likelihood using cross-validated early stopping to control for overfitting (Kjaer et al., 1994; Skouras et al., 1994; Kay et al., 2008; Cruz et al., 2011; Pasley et al., 2012). This allowed us to evaluate the classification performance for combined sites over a large time window on a single trial basis. The parameter estimation proceeded as follows. First, the dataset was divided into the following three sets: a training set (two-thirds of all trials), a validation set (one-sixth of all trials), and a stopping set (one-sixth of all trials). The training set was used to update the parameters such that the log-likelihood was improved on each iteration through the data. The stopping set was used to decide when to stop the iterations on the training data. The iteration was stopped when the log-likelihood function value calculated with the stopping set became smaller than the value in the

previous iteration. Then this classifier was used to classify the data from the validation set trial by trial. We repeated this for all six possible divisions of the data. The fraction of trials correctly classified in all validation datasets was then used as the measurement of classification performance. To estimate the classification performance from the evoked power (Fig. 8*b*), we used squared waveforms and repeated the same procedure.

ANOVA. All ANOVAs were performed as mixed models with monkey modeled as a random effect. They were implemented in JMP 10.0 (SAS Institute Inc.). Fixed effects varied depending upon the analysis. Specifically, for Figure 4*a*, vocalization category (Grunt, Bark, Scream, Girney/Warble, Coo + other, or Coo) and stimulus number were fixed effects, with stimulus number nested under vocalization category. For Figure 4, *b* and *c*, stimulus number and group (Coo or non-Coo) were fixed effects, with stimulus number nested under group. For Figure 6, *a* or *b*, stimulus number (1–20) and type (1–3: VOC, EPS, or SPS) were modeled as fixed effects to evaluate difference in performance among types. To evaluate the magnitude of the classification performance difference in Figure 6*c*, type of performance difference ($P_{\text{VOC}} - P_{\text{EPS}}$ or $P_{\text{VOC}} - P_{\text{SPS}}$) and sector (1–4) were fixed effects. For Figure 6*d*, type and sector were fixed effects. For Figure

7*a*, frequency band (theta, alpha, beta, low-gamma, or high-gamma) was a fixed effect. The tests for Figures 6, *c* and *d*, and 7 were performed separately for each of the four sectors. For Figure 8*a*, type and frequency band were fixed effects. For Figure 8*b*, predictor type (waveform or power) and frequency band were fixed effects. All *post hoc* comparisons were performed using Tukey's HSD. For simplicity, although all main effects were significant, we report only the Tukey's HSD results, as these are valid without significant main effects in an ANOVA (Zar, 2009).

Results

We recorded auditory evoked potentials simultaneously from 96 sites across three chronically implanted μ ECoG arrays in the lateral sulcus of three monkeys (Fig. 1*a*). The simultaneous recording of all sites allowed us to compare the pattern of neural responses across multiple cortical areas, controlling for factors that vary across time, including monkey vigilance, cortical excitability, and/or history of stimulus presentations. As in our previous study (Fukushima et al., 2012), the 96 sites on the STP were grouped into four sectors, based on frequency reversals in tonotopic maps obtained from responses to a set of pure-tone stimuli (Fig. 1*b*; see Materials and Methods).

Combined sites within each sector produced better classification performance for vocalizations than individual sites

The auditory stimuli included 20 conspecific macaque vocalizations (Fig. 2*a,b*). The vocalization stimuli evoked robust responses in the μ ECoG arrays (Fig. 2*c*). To quantify the information about the stimuli coded in the evoked responses, we performed classification analysis using the trial-by-trial evoked waveforms recorded from each sector to predict the identity of the vocalization stimulus presented on each trial. The measurement of information reported is the fraction of correctly classified trials (chance level = 0.05 as there were 20 stimuli in each set).

Figure 2. VOC stimuli and example evoked responses to vocalization stimuli. **a**, Spectrograms of 20 monkey vocalizations in the VOC stimulus set. These stimuli can be categorized as a Grunt (1,2), Bark (3–8), Scream (9–12), Girney/warble (13, 14), Coo (15–18), and Coo combined with other sounds (Grunt + Coo, 19 and Scream + Coo, 20). **b**, Bar plot indicating duration of each of the 20 vocalizations. **c**, Example of an evoked response to a vocalization stimulus (Stimulus 1, bark). Top, Spectrograms of evoked potentials for Stimulus 1 recorded from Monkey M's Sector 1 (left) and Sector 4 (right). These spectrograms are normalized to the baseline activity before the onset of the stimulus. Bottom, Trial-averaged waveform of the evoked responses from those two sites.

We performed the classification analysis using either (1) each of the 96 sites separately or (2) all of the sites within each of the four sectors simultaneously. First, we examined the temporal accumulation of information by carrying out the analysis with 20 different window lengths (50–1000 ms after stimulus onset in 50 ms steps; Fig. 3*a*). Analyses that included each site separately assessed the contribution of individual sites to the population information, whereas analyses that included all sites within a sector estimated the information present simultaneously within an entire sector. Note that this analysis with combined sites takes advantage of activation patterns across multiple sites in single trials, and thus it does not simply average activity across sites within a region (see Materials and Methods). This procedure could also be referred to as “pattern classification” or “multi-site analysis.” Recording with μ ECoG arrays yielded high fidelity signals that resulted in robust classification performance (Fig. 3*a*). The classification performance from the combined sites was always higher than for any individual channel for all sectors in all monkeys (Fig. 3*a,b*). This shows that the regularized classifier can

extract the information about vocalizations distributed across recording sites within each of the four sectors. We also found that the classification performance generally increased for larger time windows and reached peak performance at ~ 600 ms (Fig. 3*a*, red dots). Henceforth, unless otherwise noted, the results we report are based on performance from the combined sites in each of the four sectors at the optimal window size.

Systematic decrease in classification performance from caudal to rostral sectors

Although robust decoding performance well above chance level was obtained in all four sectors, the average classification performance decreased systematically from caudal to rostral sectors (Fig. 3*c*). To identify the source of the decrease in performance, we examined the classification performance for each of the 20 VOC stimuli individually (Fig. 3*d*). Interestingly, despite the low average performance level in the rostral sectors, the classification performance for some of the VOC stimuli remained elevated, often being quite high for the best-classified stimulus, while it

Figure 3. Classification performance (fraction correct) estimated from a single channel site and from all sites combined within a sector for different time window sizes for each of the three monkeys. *a*, Performance as a function of the time window, estimated from each channel individually (thin curves) and from all channels combined (*Figure legend continues*.)

dropped for others (Fig. 3*d*, red and cyan lines for Sectors 3 and 4, respectively). This suggests that the rostral, higher auditory cortex does not represent all types of conspecific vocalizations equally: it maintains higher discriminability for a particular subset of vocalizations relative to others.

Enhanced discrimination for individual “Coo” vocalizations

We next examined whether the difference in performance across stimuli found in Sector 4 might be related to categorical differences among the vocalizations (Fig. 4). The 20 VOC stimuli can be grouped into six categories (Grunt, Bark, Scream, Girney/Warble, Coo, and Coo combined with other sounds; Fig. 2*a*; Poremba et al., 2004; Kikuchi et al., 2010). Classification performance among these categories differed significantly in Sector 4 ($F_{(5,38)} = 4.012$, $p = 0.0051$; see Materials and Methods) as well as in Sectors 2 ($F_{(5,38)} = 5.33$, $p < 0.001$) and 3 ($F_{(5,38)} = 7.58$, $p < 0.001$). In particular, classification performance was highest for Coo and “Coo combined with others” categories in Sector 4 (Fig. 4*a*). Directly comparing these Coo groups and the “non-Coo” groups (Fig. 4*b*), we found significant differences in the rostral sectors (Sector 3, $F_{(1,38)} = 13.23$, $p < 0.001$; Sector 4, $F_{(1,38)} = 19.15$, $p < 0.001$; see Materials and Methods) but not in the caudal sectors ($p > 0.05$). Thus, the large performance difference within Sector 4 is related to the categorical identity of the stimuli.

Note that the mean duration of stimuli in the Coo group (641 ± 8 ms, mean $\pm$ SEM) is longer than that in the non-Coo group (400 ± 5 ms). To examine whether this explained the specific increase in classification performance for the Coo groups, we performed an additional classification analysis with a shorter time window (200 ms), to eliminate the effect of stimulus duration. In this analysis there was still a significant difference in the performance between Coo and non-Coo groups in Sector 4 ($F_{(1,38)} = 14.82$, $p < 0.0004$; Fig. 4*c*; see Materials and Methods). Thus, these analyses suggest that the population activity in rostral STP contains more information about Coo calls.

Figure 4. Variability in the performance is explained by the difference in vocalization category. *a*, Classification performance within each sector for each of six vocalization categories (Grunt, $n = 2$ stimuli; Bark, $n = 6$; Scream, $n = 4$; Girney Warble, $n = 2$; Coo + other, $n = 2$; Coo, $n = 4$; mean $\pm$ SEM, $N = n \times 3$ monkeys). *b*, Mean classification performance ($P \pm$ SEM) for the three monkeys by sector for both the non-Coo group of vocalizations (green bars, $n = 14$ stimuli consisting of Grunts, Barks, Screams, and Girney Warble) and the Coo group (yellow bars, $n = 6$ stimuli consisting of Coos and Coos + other). *c*, Mean classification performance ($P \pm$ SEM) based on the first 200 ms of the evoked potentials. Brackets connect pairs that differed significantly (see Materials and Methods).

Conjoined spectral and temporal features are necessary to explain neural discrimination in the rostral sector

Any vocalization is a specific combination of spectral and temporal features. To examine which of these two features was being coded within each sector, we synthesized two sets of stimuli, each of which retained only one or the other feature of the original vocalizations. Specifically, one type was EPS, and the other, SPS (see Materials and Methods; Fig. 5*a*). For all 20 EPS stimuli, the spectral content of the original calls was replaced with a flat spectral distribution, while for all 20 SPS stimuli, the temporal envelopes of the original calls were replaced with a flat temporal envelope (Fig. 5*a*). Thus, discriminating the stimuli within each synthetic sound set, EPS or SPS, could only be accomplished using the preserved auditory dimension, temporal or spectral, respectively. Like the vocalization stimuli, the synthetic stimuli evoked robust responses throughout the caudal and rostral sectors (Fig. 5*b*). We performed the same classification analysis for each of these synthetic stimulus sets as we had for the VOC set. Then we compared classification performance across all three sets (see Materials and Methods).

In the two caudal sectors (Sectors 1 and 2), the vocalizations and synthetic sound sets were classified with similar, high levels of accuracy (Fig. 6*a,b*). In contrast, classification performance in the two rostral sectors (Sectors 3 and 4) differed between the original vocalizations and the matched synthetic sound sets. In both rostral sectors, and particularly in Sector 4, the neural classification among vocal stimuli was much higher than among either of the two synthetic sound sets (Fig. 6*a,b*).

To evaluate this relative enhancement of classification performance statistically, we proceeded in two steps. First, we examined

(Figure legend continued.) (bold black curve). The red dot on each curve indicates maximum performance. The black dotted line indicates the chance level of performance (0.05). *b*, Maximum performance for combined and single sites. Black bars, Maximum performance obtained by combining all sites in a sector, same as that indicated by the red dot on the bold black curve in *a*. White bars, Mean maximum performance ($\pm$ SEM) obtained from single sites in a sector. Black dot, Highest performance among the single sites. The gray dotted line indicates the chance level of performance (0.05). *c*, Maximum performance obtained from combined sites averaged across the three monkeys. The black dotted line indicates the chance level of performance (0.05). *d*, Classification performance for each of the 20 VOC stimuli, sorted by classification performance for the vocalization, in each sector, averaged across the three monkeys (Sectors 1–4 shown in blue, black, red, and cyan, respectively).

Figure 5. Synthetic sound stimuli and example evoked responses. **a**, Monkey vocalizations and two sets of synthetic stimuli. Top, Original VOC Stimulus 10; middle, EPS-derived from stimulus 10 by replacing the original power spectrum with a flat one; bottom, SPS-derived from Stimulus 10 by replacing the original temporal envelope with a flat one. Left, Waveforms; right, average power spectrum. **b**, The trial-averaged evoked waveforms to 60 stimuli (20 each for three sound types: VOC, EPS, and SPS) recorded from Monkey M's Sector 1 (top) and Sector 4 (bottom).

whether the average classification performance (P) for the VOC set (P_{VOC}) of 20 stimuli was higher than the average performance for each synthetic sound set of 20 stimuli (P_{EPS} , P_{SPS}) in each area separately (Fig. 6a). We found that the classification performance for the VOC set was higher than that for the EPS set in Sectors 1 (Tukey's HSD, $p < 0.001$, see Materials and Methods), 2 ($p = 0.0077$), 3 ($p < 0.001$), and 4 ($p < 0.001$), and higher than that for the SPS set in Sectors 3 ($p < 0.001$) and 4 ($p < 0.001$). The classification performance for the SPS set was significantly higher than that for the EPS set only in Sector 1 ($p < 0.0004$).

Next, we examined whether the magnitudes of this relative increase in classification performance differed among the sectors (Fig. 6c). We found that both $P_{\text{VOC}} - P_{\text{EPS}}$ and $P_{\text{VOC}} - P_{\text{SPS}}$ were greater in Sectors 3 and 4 than in either Sectors 1 or 2 (Tukey's HSD, $p < 0.001$ for all except $P_{\text{VOC}} - P_{\text{EPS}}$ in Sectors 3 vs 2, in which the p value was 0.0063; Fig. 6c, orange bars). Thus, the rostral sectors, but not the caudal sectors, coded vocalizations better than the synthetic sound sets. Therefore, neural discrimination in the higher rostral cortex cannot be explained by the isolated temporal or spectral features of the vocalizations, suggesting that the natural conjunction of these features in vocalizations is an essential aspect of neural discrimination in the rostral auditory cortex.

We also examined the degree to which the difference in classifier performance across individual vocalizations could be explained by the difference in performance of the synthetic stimuli. To quantify this difference across individual stimuli, we calculated the variance of the performance within each sector (Fig. 6d). Small variance within a stimulus set indicates similar performance across stimuli, whereas large variance indicates elevated

performance for a subset of those stimuli. We found that the variance for the VOC set systematically increased from Sector 1 to Sector 4. However, the variance for the synthetic stimuli peaked in Sector 3 and dropped in Sector 4 (Fig. 6d). As a result, the variance in classification performance for VOC stimuli was significantly higher than that for the synthetic sound sets only in Sector 4 (Tukey's HSD, $p = 0.025$ for VOC vs EPS, $p = 0.012$ for VOC vs SPS). Thus, whereas the difference in performance across individual stimuli for the VOC set in Sectors 1, 2, and 3 can be explained by the difference in either the spectral or the temporal features preserved in that particular synthetic sound set, the high variance in performance of the VOC set in Sector 4 cannot be explained in the same way.

Note also that the stimuli in the EPS set have exactly the same duration distribution as those in the VOC set. Therefore, the small performance variability of the EPS set in Sector 4 suggests that the difference in stimulus duration alone (Fig. 2b) cannot explain either the difference in performance across individual VOC stimuli or the specific increase in classification performance for a subset of VOC stimuli (e.g., Co; Fig. 4). Thus, these results suggest that the most rostral sector (Sector 4) is distinctly different from other sectors in combining spectral and temporal features in vocalizations.

Dominance of theta band in rostral STP vocalization coding

The above results were obtained from broadband (4–200 Hz) evoked waveforms. Examination of the spectrograms of the evoked responses indicated that there was greater power at high frequencies (e.g., high-gamma band, 60–200 Hz) in the caudal than in the rostral STP (Fig. 2c). On the other hand, low-frequency power (e.g., theta band, 4–8 Hz) was present in both

caudal and rostral sites. These observations suggest that the vocalization-specific coding in rostral areas may rely most heavily on low-frequency response components. To test this hypothesis, we compared the information estimated from the broadband evoked waveforms with that present in individual frequency bands (theta, alpha, beta, low-gamma, and high-gamma; Fig. 7).

The enhanced coding for vocal stimuli in rostral areas was expressed predominantly in the theta frequency range. While classification performance using different frequency bands was similar in the caudal Sectors 1 and 2, it differed significantly across these bands in the rostral sectors: Sector 3 ($F_{(4,8)} = 7.98$, $p = 0.007$) and Sector 4 ($F_{(4,8)} = 21.26$; $p < 0.001$; Fig. 7a). In Sector 3, the classification performance of the theta band was significantly higher than the performance of the low-gamma and high-gamma bands (Tukey's HSD, $p = 0.016$ for low-gamma, $p = 0.008$ for high-gamma), whereas in Sector 4, theta band performance was significantly higher than that of all other bands ($p = 0.017$ for alpha, $p = 0.001$ for beta and low-gamma, and $p < 0.001$ for high-gamma).

Although the performance in Sector 4 was highest in the theta band, high-frequency components in Sector 4 still yielded significant classification performance. We examined how much of this could be explained by either the temporal or the spectral features of the vocalizations and found that the classification performance for the VOC set (P_{VOC}) was significantly greater than that for either of the synthetic sound sets in both the theta band (Tukey's HSD for VOC vs EPS and VOC vs SPS, $p < 0.001$ for each) and the alpha band (Tukey's HSD for VOC vs EPS, $p = 0.006$; VOC vs SPS, $p = 0.014$), but not in the beta, low-gamma, or high-gamma bands ($p > 0.05$; Fig. 8a). This suggests that the classification performance for the VOC set from the slow-wave components (theta or alpha) cannot be explained simply by the difference in acoustic features available in the synthetic sound sets (e.g., the slowly modulated envelope which is present in the EPS set), and thus these components contribute significantly to the coding of vocalizations. On the other hand, the difference in acoustic features in the synthetic sounds could explain the classification performance from the high-frequency components (beta, low-gamma, and high-gamma) applied to the VOC set. This suggests that information about conjoined spectral and temporal features in vocalizations in the rostral sector are selectively carried by the slow-wave components of the evoked potentials.

To examine which feature of the evoked waveform contributed to coding vocalizations in Sector 4, we compared the classi-

Figure 6. Classification performance for the vocalization and synthetic sounds within each sector. **a**, The fraction of the 20 stimuli in each of the three stimulus sets (blue, VOC; green, EPS; red, SPS) in each of the four sectors that are correctly classified, in descending rank order, averaged across the three monkeys. The black dotted line indicates the chance level of performance (0.05). **b**, The same data shown in **a** are plotted as a function of stimulus number defined in Figure 2a. **c**, Difference in classification performance between the VOC (P_{VOC}) and EPS sets (P_{EPS} ; orange bars) and between the VOC and SPS sets (P_{SPS} ; blue bars). Brackets connect sectors that differ significantly (see text). The performance difference was obtained by averaging for 20 stimuli for each stimulus set in each of three monkeys (mean $\pm$ SEM, $N = 3$ monkeys $\times$ 20 stimuli). **d**, Variance in classification performance across the 20 stimuli for each of the three stimulus sets, in each sector, averaged across the three monkeys (mean $\pm$ SEM).

fication performance obtained from the power of the evoked response to that obtained from the evoked waveform (Fig. 8b; see Materials and Methods). We found that the information in the waveform was consistently higher than the information in the power, except at the highest frequency band (i.e., high-gamma). The classification performance of the broadband waveform was significantly different from that of the power (Tukey's HSD, $p < 0.0001$). The theta band was the only component that showed a significant difference in performance between the evoked waveform and the power (Tukey's HSD, $p = 0.0005$). The similarity in performance of the broadband and theta band activity again indicates the importance of the contribution that the theta band activity makes to the coding of vocalizations. Also, the significant

Figure 7. Classification performance obtained from the bandpassed waveform. *a*, The mean fraction correct for each of five frequency bands (theta, θ ; alpha, α ; beta, β ; low-gamma, L γ ; high-gamma, H γ) for each of the four sectors (mean $\pm$ SEM, $N = 3$ monkeys $\times$ 20 stimuli). The red dotted line indicates the mean performance from the broadband waveform. The black dotted line indicates the chance level of performance (0.05). The brackets connect pairs that differ significantly (see text). *b*, The ratio between the classification performance from the bandpassed waveform and that from the broadband waveform. One hundred percent indicates that classification performance from that specific bandpassed waveform equals the performance from the broadband waveform (mean $\pm$ SEM, $N = 3$ monkeys $\times$ 20 stimuli).

difference in performance between the waveform and the power points to a role for the phase of the theta-band waveform in coding vocalizations in Sector 4.

Discussion

In the current study, we investigated neural coding of conspecific vocalizations by simultaneously recording auditory evoked potentials from multiple auditory areas in the ventral stream. We then used a multivariate regularized classifier to decode the evoked potentials and estimate information about vocalizations within each area. We found a gradual decrease in the level of overall classification performance from the caudal to rostral sectors (Fig. 3c). Despite the decreased performance in the rostral sectors, the performance for the best-classified stimulus in Sector 4 remained high (~ 0.7 ; Fig. 3d, cyan line), and thus there was a considerable difference in classifier performance across stimuli in the rostral sectors. Further analysis showed that different vocalization categories exhibited different levels of classification performance (Fig. 4).

Several of these results are consistent with previous observations in the visual system. First, decoding studies in fMRI have shown a decrease in classification performance along the visual cortical hierarchy from V1 to V3 (Miyawaki et al., 2008), which is similar to the performance decrease we found from the caudal to rostral auditory areas. Second, rostral, high-level visual cortex contributes to coding the semantic content of images rather than low-level visual features (Naselaris et al., 2009). Our results also suggest that coding of auditory stimuli in higher auditory cortex might not be based simply on low-level auditory features, as we

found a difference in classification accuracy in the rostral STP across vocalization categories (i.e., Coo calls were better represented than other calls).

One caveat is that the ECoG electrode arrays recorded field potentials from the cortical surface. Therefore, we cannot say for certain whether the same results would hold if one recorded from a large group of neurons within each cortical area, along the extent of the STP. However, the arrays do reveal a consistent map of characteristic frequency, and this is known to be a feature of single neurons in auditory cortex (Fukushima et al., 2012). ECoG array recordings also reflect the retinotopic map in V1, similar to the one that is obtained from single-unit spikes recorded from depth electrodes (Rols et al., 2001; Bosman et al., 2012). Also, simultaneous recording with ECoG and depth electrodes has demonstrated a high correlation between evoked potentials recorded with ECoG and depth electrodes at the same cortical locations (Toda et al., 2011).

Conjoined spectral and temporal features are necessary to explain coding property in the rostral auditory cortex

The data also showed another trend in the caudorostral direction: neural classification performance in rostral STP was significantly better for vocalizations than for the synthetic stimuli (EPS or SPS; Fig. 6c).

This difference was smaller in primary auditory cortex but increased gradually along the ventral auditory pathway. The high classification performance in primary auditory cortex can be understood in terms of the functional properties of this area. Neurons there tend to respond to simple and natural stimuli (Wang et al., 1995; Kikuchi et al., 2010) and reliably follow the temporal modulation of acoustic stimuli (Bendor and Wang, 2007; Scott et al., 2011). This pattern of responses would produce distinct population-response profiles to sounds with different spectrotemporal content, regardless of stimulus type. This is consistent with our results, which reflect nonselective processing of vocalizations and synthetic stimuli in the primary auditory cortex (Fig. 6a, Sectors 1 and 2). Evidently, in this area, differences in either the temporal or the spectral features of the original vocalizations are sufficient to drive unique population representations. For the rostral sectors, on the other hand, the vocalizations and synthetic stimuli are not equally represented. This suggests that conjoined spectral and temporal features are necessary to produce a level of neural discrimination as high as that produced by the original vocalizations. This is consistent with the hypothesis that the most rostral auditory areas act as detectors for complex spectrotemporal features (Bendor and Wang, 2008).

Enhanced discrimination for particular sound classes in rostral auditory cortex

The enhanced discrimination we found for a subset of vocalizations (specifically, Coo calls) in rostral auditory cortex raises questions of why such an enhancement exists, what could drive

Figure 8. Comparison of classification performance in Sector 4 across the different band-passed field potential frequencies. *a*, Difference in performance (mean $\pm$ SEM, $N = 3$ monkeys $\times$ 20 stimuli) between the VOC set and each of the two synthetic sound sets (red, EPS; blue-green, SPS). Asterisks indicate that the performance scores obtained from theta (θ) and alpha (α) waves differ between VOC and synthetic sound sets (** $p < 0.0001$, * $p < 0.015$). *b*, Classification performance obtained from the power of the evoked waveform (yellow bar; mean $\pm$ SEM, $N = 3$ monkeys $\times$ 20 stimuli). The performance indicated by the orange bars is identical to the data for Sector 4 in Figure 7*a*. The black dotted line indicates chance level performance (0.05). Brackets indicate that the mean fraction correct for the power differs significantly from that for the waveform (see text). beta, β ; low-gamma, Low γ ; high-gamma, High γ .

such an enhancement, and whether the enhancement is specific to vocalizations. The better representation of harmonically structured Coo calls over other vocalizations, such as broadband grunts, might be rooted in their ecological value with respect to information about the individual caller. For example, previous behavioral studies have demonstrated that Coo calls convey behaviorally relevant information such as body size, which could be related to individual identity (May et al., 1989; Rendall et al., 1998; Ghazanfar et al., 2007). Another possible explanation for better classification performance for Coo calls would be an increased sensitivity to harmonically structured sounds (not exclusive to vocalizations) in anterior auditory cortical fields. The features of harmonic sounds are coded in specific regions in human and monkey auditory cortex (Bendor and Wang, 2005; Norman-Haignere et al., 2013). The harmonic-phase coding is also a feature thought to be essential for stimulating ventrolateral prefrontal cortex (Averbeck and Romanski, 2004), a region of the brain closely connected to the animals' behavioral responses. Whether these or still other mechanisms underlie enhanced discrimination of Coo sounds in anterior sectors of the STP requires further investigation.

The rostral auditory cortex may also be able to expand the neural representation of behaviorally relevant nonvocalization sounds, although we did not examine this in our study. In humans, it has been shown previously that speech-sensitive regions of the left posterior superior temporal sulcus become sensitive to artificial sounds with which subjects have been trained in a behavioral task (Leech et al., 2009). There have been similar findings in the visual system, where it has been shown that expertise on nonface objects recruits face-selective areas (Gauthier et al., 1999, 2000). It would be interesting to test whether training monkeys with nonvocalization sounds would improve neural discrimination selectively in rostral, higher auditory cortex. A previous study reported that the number of neurons responsive to white noise increased in the rostral auditory cortex in monkeys trained to release a bar in response to white noise to obtain a juice reward (Kikuchi et al., 2010).

Coding of vocalization is supported by theta band activity

Our analysis showed that while information from the theta band was robust throughout all four sectors of the STP, it dominated the selective discrimination of vocalizations in the rostral sectors (Fig. 7). It is important to note that there was also less power in higher frequency bands such as gamma in the rostral sector (Fig. 2*c*), and this may account for some of our results. However, in Sector 4, classification performances for all frequency bands were still significantly higher than chance, and we showed that the classification performance for VOC stimuli was significantly different from that for synthetic stimuli only in the theta band (Fig. 8*a*). This supports the idea that conjoined features in vocalizations are coded in theta components in the rostral sector. This cannot entirely be explained by a general reduction of evoked power in high-frequency components, suggesting the importance of theta oscillation in coding vocalizations in the rostral sector.

In the rostral auditory cortex, the temporal profile of neural responses does not precisely follow temporal modulations in acoustic stimuli (Bendor and Wang, 2010; Scott et al., 2011). Temporally modulated sounds such as acoustic flutter can, however, be well encoded by firing rate in rostral areas, suggesting that there is a transformation from temporal to rate code as one progresses from the caudal to rostral auditory cortex (Bendor and Wang, 2007). This transformation to a rate code implies a reduction in stimulus-driven synchronized spiking activity among local populations of neurons in the rostral auditory cortex. Interestingly, recent studies have suggested that increases in higher frequency power can be regarded as an index of the degree of synchronization in local populations of neurons (Buzsáki et al., 2012). Thus, our finding of a reduction of high-frequency evoked power in the rostral sector might reflect the rate coding of sounds in this sector.

It has also been suggested that the theta band is an intrinsic temporal reference frame that could increase the information encoded by spikes (Panzeri et al., 2010; Kayser et al., 2012). This mechanism also has the potential to add information to spiking activity and thereby help the coding of sounds in the rostral area where spike timing relative to stimulus onset is not as reliable as that in primary auditory cortex (Bendor and Wang, 2010; Kikuchi et al., 2010; Scott et al., 2011).

Thus, one interpretation of our results is that the caudorostral processing pathway selectively extracts conjoined high-level features of vocalizations and represents them with a temporal structure.

ure that matches slow rhythms important for conspecific interaction (Singh and Theunissen, 2003; Averbek and Romanski, 2006; Cohen et al., 2007; Hasson et al., 2012; Ghazanfar et al., 2013).

References

- Averbek BB, Romanski LM (2004) Principal and independent components of macaque vocalizations: constructing stimuli to probe high-level sensory processing. *J Neurophysiol* 91:2897–2909. [CrossRef Medline](#)
- Averbek BB, Romanski LM (2006) Probabilistic encoding of vocalizations in macaque ventral lateral prefrontal cortex. *J Neurosci* 26:11023–11033. [CrossRef Medline](#)
- Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. *Nature* 436:1161–1165. [CrossRef Medline](#)
- Bendor D, Wang X (2007) Differential neural coding of acoustic flutter within primate auditory cortex. *Nat Neurosci* 10:763–771. [CrossRef Medline](#)
- Bendor D, Wang X (2008) Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J Neurophysiol* 100:888–906. [CrossRef Medline](#)
- Bendor D, Wang X (2010) Neural coding of periodicity in marmoset auditory cortex. *J Neurophysiol* 103:1809–1822. [CrossRef Medline](#)
- Bosman CA, Schoffelen JM, Brunet N, Oostenveld R, Bastos AM, Womelsdorf T, Rubehn B, Stieglitz T, De Weerd P, Fries P (2012) Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron* 75:875–888. [CrossRef Medline](#)
- Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nat Rev Neurosci* 13:407–420. [CrossRef Medline](#)
- Cohen YE, Theunissen F, Russ BE, Gill P (2007) Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. *J Neurophysiol* 97:1470–1484. [CrossRef Medline](#)
- Cruz AV, Mallet N, Magill PJ, Brown P, Averbek BB (2011) Effects of dopamine depletion on information flow between the subthalamic nucleus and external globus pallidus. *J Neurophysiol* 106:2012–2023. [CrossRef Medline](#)
- Doupe AJ, Kuhl PK (1999) Birdsong and human speech: common themes and mechanisms. *Annu Rev Neurosci* 22:567–631. [CrossRef Medline](#)
- Edwards E, Soltani M, Deouell LY, Berger MS, Knight RT (2005) High gamma activity in response to deviant auditory stimuli recorded directly from human cortex. *J Neurophysiol* 94:4269–4280. [CrossRef Medline](#)
- Fukushima M, Saunders RC, Leopold DA, Mishkin M, Averbek BB (2012) Spontaneous high-gamma band activity reflects functional organization of auditory cortex in the awake macaque. *Neuron* 74:899–910. [CrossRef Medline](#)
- Furl N, Hadj-Bouziane F, Liu N, Averbek BB, Ungerleider LG (2012) Dynamic and static facial expressions decoded from motion-sensitive areas in the macaque monkey. *J Neurosci* 32:15952–15962. [CrossRef Medline](#)
- Gauthier I, Tarr MJ, Anderson AW, Skudlarski P, Gore JC (1999) Activation of the middle fusiform “face area” increases with expertise in recognizing novel objects. *Nat Neurosci* 2:568–573. [CrossRef Medline](#)
- Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci* 3:191–197. [CrossRef Medline](#)
- Ghazanfar AA, Hauser MD (2001) The auditory behaviour of primates: a neuroethological perspective. *Curr Opin Neurobiol* 11:712–720. [CrossRef Medline](#)
- Ghazanfar AA, Tureson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK (2007) Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr Biol* 17:425–430. [CrossRef Medline](#)
- Ghazanfar AA, Morrill RJ, Kayser C (2013) Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. *Proc Natl Acad Sci U S A* 110:1959–1963. [CrossRef Medline](#)
- Hackett TA (2011) Information flow in the auditory cortical network. *Hear Res* 271:133–146. [CrossRef Medline](#)
- Hasson U, Ghazanfar AA, Galantucci B, Garrod S, Keysers C (2012) Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends Cogn Sci* 16:114–121. [CrossRef Medline](#)
- Hauser MD, Marler P (1993) Food-associated calls in rhesus macaques (*Macaca mulatta*): I. Socioecological factors. *Behav Ecol* 4:194–205. [CrossRef](#)
- Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. *Nature* 452:352–355. [CrossRef Medline](#)
- Kayser C, Ince RA, Panzeri S (2012) Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. *PLoS Comput Biol* 8:e1002717. [CrossRef Medline](#)
- Kellis S, Miller K, Thomson K, Brown R, House P, Greger B (2010) Decoding spoken words using local field potentials recorded from the cortical surface. *J Neural Eng* 7:056007. [CrossRef Medline](#)
- Kikuchi Y, Horwitz B, Mishkin M (2010) Hierarchical auditory processing directed rostrally along the monkey’s supratemporal plane. *J Neurosci* 30:13021–13030. [CrossRef Medline](#)
- King AJ, Nelken I (2009) Unraveling the principles of auditory cortical processing: can we learn from the visual system? *Nat Neurosci* 12:698–701. [CrossRef Medline](#)
- Kjaer TW, Hertz JA, Richmond BJ (1994) Decoding cortical neuronal signals: network models, information estimation and spatial tuning. *J Comput Neurosci* 1:109–139. [CrossRef Medline](#)
- Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M (2013) The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn Sci* 17:26–49. [CrossRef Medline](#)
- Leech R, Holt LL, Devlin JT, Dick F (2009) Expertise with artificial non-speech sounds recruits speech-sensitive cortical regions. *J Neurosci* 29:5234–5239. [CrossRef Medline](#)
- Leopold DA, Murayama Y, Logothetis NK (2003) Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. *Cereb Cortex* 13:422–433. [CrossRef Medline](#)
- May B, Moody DB, Stebbins WC (1989) Categorical perception of conspecific communication sounds by Japanese macaques, *Macaca fuscata*. *J Acoust Soc Am* 85:837–847. [CrossRef Medline](#)
- Miyawaki Y, Uchida H, Yamashita O, Sato MA, Morito Y, Tanabe HC, Sadato N, Kamitani Y (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60:915–929. [CrossRef Medline](#)
- Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63:902–915. [CrossRef Medline](#)
- Norman-Haignere S, Kanwisher N, McDermott JH (2013) Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J Neurosci* 33:19451–19469. [CrossRef Medline](#)
- Panzeri S, Brunel N, Logothetis NK, Kayser C (2010) Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33:111–120. [CrossRef Medline](#)
- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251. [CrossRef Medline](#)
- Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. *Curr Biol* 21:1408–1415. [CrossRef Medline](#)
- Petkov CI, Jarvis ED (2012) Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front Evol Neurosci* 4:12. [CrossRef Medline](#)
- Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK (2008) A voice region in the monkey brain. *Nat Neurosci* 11:367–374. [CrossRef Medline](#)
- Poremba A, Malloy M, Saunders RC, Carson RE, Herscovitch P, Mishkin M (2004) Species-specific calls evoke asymmetric activity in the monkey’s temporal poles. *Nature* 427:448–451. [CrossRef Medline](#)
- Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724. [CrossRef Medline](#)
- Rendall D, Rodman PS, Emond RE (1996) Vocal recognition of individuals and kin in free-ranging rhesus monkeys. *Anim Behav* 51:1007–1015. [CrossRef](#)
- Rendall D, Owren MJ, Rodman PS (1998) The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. *J Acoust Soc Am* 103:602–614. [CrossRef Medline](#)
- Rols G, Tallon-Baudry C, Girard P, Bertrand O, Bullier J (2001) Cortical mapping of gamma oscillations in areas V1 and V4 of the macaque monkey. *Vis Neurosci* 18:527–540. [CrossRef Medline](#)
- Romanski LM, Averbek BB (2009) The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu Rev Neurosci* 32:315–346. [CrossRef Medline](#)

- 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. [CrossRef](#) [Medline](#)
- Scott BH, Malone BJ, Semple MN (2011) Transformation of temporal processing across auditory cortex of awake macaques. *J Neurophysiol* 105:712–730. [CrossRef](#) [Medline](#)
- Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114:3394–3411. [CrossRef](#) [Medline](#)
- Skouras K, Goutis C, Bramson MJ (1994) Estimation in linear models using gradient descent with early stopping. *Statist Comput* 4:271–278. [CrossRef](#)
- Toda H, Suzuki T, Sawahata H, Majima K, Kamitani Y, Hasegawa I (2011) Simultaneous recording of ECoG and intracortical neuronal activity using a flexible multichannel electrode-mesh in visual cortex. *Neuroimage* 54:203–212. [CrossRef](#) [Medline](#)
- Ungerleider LG, Mishkin M (1982) Two cortical visual systems. In: *Analysis of visual behavior* (Ingle MA, Goodale MI, Masfield RJW, eds), pp 549–586. Cambridge, MA: MIT.
- Wang X, Merzenich MM, Beitel R, Schreiner CE (1995) Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J Neurophysiol* 74:2685–2706. [Medline](#)
- Zar JH (2009) *Biostatistical analysis*, Ed 5. New York: Pearson.