Sound-identity processing in early areas of the auditory ventral stream in the macaque

Pawl Kuśmierek,1 Michael Ortiz,1,2 and Josef P. Rauschecker1

1Department of Neuroscience, Georgetown University Medical Center, Washington, District of Columbia; and 2International Max Planck Research School, Tübingen, Germany

Submitted 29 August 2011; accepted in final form 24 November 2011

Kuśmierek P, Ortiz M, Rauschecker JP. Sound-identity processing in early areas of the auditory ventral stream in the macaque. J Neurophysiol 107: 1123–1141, 2012. First published November 30, 2011; doi:10.1152/jn.00793.2011.—Auditory cortical processing is thought to be accomplished along two processing streams. The existence of a posterior/dorsal stream dealing, among others, with the processing of spatial aspects of sound has been corroborated by numerous studies in several species. An anterior/ventral stream for the processing of nonspatial sound qualities, including the identification of sounds such as species-specific vocalizations, has also received much support. Originally discovered in anterolateral belt cortex, most recent work on the anterior/ventral pathway has been performed on far anterior superior temporal (ST) areas and on ventrolateral prefrontal cortex (VLPFC). Regions of the anterior/ventral stream near its origin in early auditory areas have been less explored. In the present study, we examined three early auditory regions with different anteroposterior locations (caudal, middle, and rostral) in awake rhesus macaques. We analyzed how well classification based on sound-evoked activity patterns of neuronal populations replicates the original stimulus categories. Of the three regions, the rostral region (rR), which included core area R and medial belt area RM, yielded the greatest classification success across all stimulus classes or between classes of natural sounds. Starting from ~80 ms past stimulus onset, clustering based on the population response in rR became clearly more successful than clustering based on responses from any other region. Our study demonstrates that specialization for sound-identity processing can be found very early in the auditory ventral stream. Furthermore, the fact that this processing develops over time can shed light on underlying mechanisms. Finally, we show that population analysis is a more sensitive method for revealing functional specialization than conventional types of analysis.

auditory cortex; rostral area; auditory object; cluster analysis

THE CONCEPT OF TWO STREAMS in auditory cortical processing, similar to those in visual cortex (see, e.g., Mishkin et al. 1983) was proposed more than a decade ago (Rauschecker et al. 1997; Rauschecker 1998; Rauschecker and Tian 2000). It was supported by contrasting patterns of anatomical connections in the macaque from anterior/ventral and posterior/dorsal belt regions of auditory cortex to segregated domains of lateral prefrontal cortex (Romanski et al. 1999) and by different physiological properties of these regions. In particular, the anterior lateral belt (area AL) exhibited enhanced selectivity for the identity of sounds (monkey vocalizations), the caudal lateral belt (CL) was particularly selective to sound location, and the middle lateral belt area (ML) fell in between and showed no specific preference for either of these properties (Tian et al. 2001).

Although refinements of the dual-pathway hypothesis have been proposed (e.g., Rauschecker and Scott 2009; Rauschecker 2011), its core concept has persisted and has been supported by numerous studies. Perhaps the most massive evidence for dual auditory processing streams comes from functional imaging in humans (e.g., Alain et al. 2001; Arnott et al. 2004; Binder et al. 2000; Chevillet et al. 2011; Leaver and Rauschecker 2010; Maeder et al. 2001). In nonhuman primates, Recanzone and colleagues have presented extensive data confirming the enhanced selectivity of neurons in caudal regions of macaque auditory cortex (especially area CL) to sound location (Recanzone et al. 2000a, 2000b, 2010; Woods et al. 2006), thus providing further support for the existence of a caudal “where”-stream.

The concept of an anterior “what”-stream has been tested in various monkey studies as well. After the initial demonstration of increased selectivity to monkey calls in area AL by Tian et al. (2001), most of the evidence for stimulus-identity coding in the anterior pathway came from studies of far anterior regions of the superior temporal (ST) cortex and of ventrolateral prefrontal cortex (VLPFC) (Cohen et al. 2009; Kikuchi et al. 2010; Petkov et al. 2008; Poremba et al. 2004; Romanski et al. 2005). Fewer data are available from the earliest stages of the anterior stream, that is, from areas adjacent to primary auditory cortex (A1). Recanzone (2008) found no difference in monkey call selectivity between the rostral core area (R) and A1, suggesting that feature selectivity, or at least vocalization selectivity, may not emerge until the level of the belt. The same study failed to find specificity in belt area ML, consistent with Tian et al. (2001). However, Recanzone’s (2008) recordings did not extend to belt area AL, where Tian et al. (2001) found the earliest signs of selectivity to vocalizations, nor to other anterior belt regions, a fact noted both by the author himself (Recanzone 2008) and by subsequent commentators (Bizley and Walker 2009).

Indirect confirmation for the existence of an anterior “what”-stream in primates came from measures of temporal integration, which increase in areas anterior to A1 (Bendor and Wang 2008; Kuśmierek and Rauschecker 2009; Scott et al. 2011). Temporal integration of acoustic features is necessary for auditory stimulus identification (Rauschecker and Tian 2000). Combined with increasing spectral integration (Rauschecker et al. 1995), it provides the equivalent to the gradual increase in receptive field complexity along the ventral visual stream (see, e.g., Connor et al. 2007; Desimone et al. 1984).
As the anterior stream processes auditory structure in a hierarchical fashion (Chevillet et al. 2011), selectivity for sound stimuli is expected to develop gradually along the stream, which would make it more difficult to find indications of selectivity closer to A1. Still, the beginnings of selectivity constituting initial primitives of sound identification should be present even at relatively early processing stages. Coding may take place across larger populations of neurons, as has been found in posterior regions: While single neurons carry little information about sound location in auditory areas adjacent to A1 on the posterior side, analyzing neural populations allowed extraction of more precise spatial information (Miller and Recanzone 2009). Population analysis has also been successful in studies of call selectivity in rhesus monkey VLPFC (Romanski et al. 2005) and of periodicity discrimination in ferret auditory cortex (Bizley et al. 2010). Thus we decided to investigate how stimulus identity is represented by population activity patterns in early areas of the auditory ventral stream.

Furthermore, many previous studies focused on selectivity of single neurons for stimuli within a stimulus class, e.g., monkey calls (e.g., Ku¨smierek and Rauschecker 2009; Recanzone 2008; Romanski et al. 2005; Russ et al. 2008; Tian et al. 2001). In the present study, we examined how responses of neural populations can be used to discriminate between stimulus classes.

MATERIALS AND METHODS

This article presents population analysis of data collected in two single-unit/multunit recording experiments from four male rhesus monkeys (experiment 1: monkeys S and L; experiment 2: monkeys B and N). Data from monkeys S and L have also been used in another study, but for a different purpose and analyzed in a different way (Ku¨smierek and Rauschecker 2009). Experiments 1 and 2 were conceptually very similar but differed in some details. For the sake of clarity, detailed information on methodological differences between experiments 1 and 2 was moved to the last section of MATERIALS AND METHODS. In earlier sections, we describe only those differences that are crucial to data interpretation.

Animals. Each animal was implanted with a plastic recording chamber (Crist Instruments, Hagerstown, MD) over left auditory cortical areas. Implant locations were confirmed by 3 T MRI with 1-mm³ voxel size. Monkeys were water restricted to provide adequate drive in a fluid-rewarded task. All experiments were conducted in accordance with National Institutes of Health guidelines and approved by the Georgetown University Animal Care and Use Committee.

Stimuli and task. Monkeys were seated in a monkey chair (Crist Instruments) in a sound-attenuated chamber (IAC, Bronx, NY) measuring 2.6 m × 2.6 m × 2.0 m (W × L × H).

The stimuli of interest included artificial sounds: pure tones (PT), 1/3-octave and 1-octave band-pass noise bursts (1/3-oct BPN, 1-oct BPN), and two classes of natural sounds: rhesus monkey calls (MC) and environmental sounds (ES). Duration of tones and noise bursts was 500 ms (experiment 1) or 300 ms (experiment 2). Because of limitations of the presentation system in experiment 2, the range of PT and BPN frequencies was reduced compared with experiment 1 and the number of MC and ES stimuli was 7 in each class instead of 10. (Throughout this article, “BPN frequency” denotes BPN center frequency.) All MC used in experiment 2 were previously used in experiment 1, whereas only five of seven ES used in experiment 2 were previously used in experiment 1. Spectrograms of the MC and ES stimuli used in experiment 1 were published previously [see Fig. 1 in Ku¨smierek and Rauschecker (2009)]. For experiment 2, three of these MC (coo2, coo3, scream2) and five of the ES [cage (sound of monkey swinging), cage divider, monkey chair latch close, VCR and TV turning on, and water running in sink] were dropped, while two new ES (cage latch and water dripping) were added. Of the retained ES, some were shortened compared with experiment 1, but this did not affect the first 160-ms period, which was analyzed in the present study.

Data from each experiment were analyzed separately. Then, in addition, responses to stimuli that were common to experiments 1 and 2 were pooled and analyzed together. The respective analyses/results are labeled as “experiment 1,” “experiment 2,” and “combined.” The stimulus sets used in the analyses were as follows: experiment 1: nine PT, nine 1/3-oct BPN, nine 1-oct BPN (frequencies of PT and BPN: 0.125–32 kHz), ten MC, and ten ES; experiment 2: seven PT, seven 1/3-oct BPN, seven 1-oct BPN (frequencies of PT and BPN: 0.25–16 kHz), seven MC, and seven ES; combined: seven PT, seven 1/3-oct BPN, seven 1-oct BPN (frequencies of PT and BPN: 0.25–16 kHz), seven MC, and five ES.

Stimulus duration ranged from 151 to 2,614 ms. Thus, to ensure that only stimulus-driven activity contributed to the results, all analyses covered the first 160 ms of neural responses (the approximate duration of the shortest stimulus). Similarly, when acoustic properties of stimuli were examined, only the first 160 ms of each stimulus period was used. The stimulus presentation level was set to ~50 dB and ~30 dB above the macaque hearing threshold (Jackson et al. 1999) for experiments 1 and 2, respectively. The stimulus equalization procedure has been described previously (Ku¨smierek and Rauschecker 2009).

The behavioral task was go/no-go auditory discrimination: A bar-release response to an infrequent (~10–15%) auditory target was rewarded by a small amount of juice, water, or balanced electrolyte drink (Prang, BioServ, Frenchtown, NJ). The purpose of the task was to keep the animals at an approximately constant level of attention. A block of all stimuli (including several repetitions of the behavioral target) was presented 10–13 times (experiment 1) or 60 times (experiment 2) in random order within each block presentation. Each trial started with a 300- to 400-ms pretrial period, during which the animal had to keep its hand on the bar. All trials, irrespective of behavioral response, were used in the analyses. Neuronal recordings. Single-unit and multiunit recordings were obtained by advancing one or two Epoxyite- or glass-insulated 1- to 3-MΩ tungsten electrodes (FHC, Bowdoin, ME or NAN Instruments, Nazareth Illit, Israel) into the auditory cortex by means of a micro-positioner (model 650, David Kopf, Tujunga, CA or FlexMT/EPIS, Alpha Omega, Nazareth Illit, Israel). A stainless steel guide tube was used to puncture the dura. A 1 mm × 1 mm spacing grid (Crist Instruments) provided a repeatable spatial reference for electrode location. The electrode signal was amplified and filtered [model 1800, A-M Systems, Sequim, WA, and PC1, TDT (Alachua, FL), or MCP Plus, Alpha Omega]. In experiment 1, neural activity was isolated with a window discriminator (SD1, TDT), and spike time stamps were recorded with a custom-made program (“Fiordiligi”; Ku¨smierek and Rauschecker 2009), which also presented stimuli and controlled the behavioral task. In experiment 2, Power1401mk2 (CED, Cambridge, UK) interface and Spike2 program (v. 6 or 7, CED) running custom-made scripts were used to record and isolate neural activity, present stimuli, and control behavior; in most cases, units were isolated post hoc by principal component analysis. As the recording electrode was lowered, the surface of auditory cortex was determined from recording depth (in reference to MRI images) and from the presence of a “silent gap” corresponding to the lateral sulcus. To drive the neurons (whether identified by baseline activity or silent when not stimulated), we used the same stimuli as those in the formal testing and/or natural sounds produced ad hoc (knocking, hissing, key jingling, clapping, etc.). Only auditory-responsive units were tested further.

Cortical regions. Since the main purpose of the study was to investigate antero-posterior differences of acoustic stimulus representation in core and medial belt, recordings from core area R were
pooled with recordings from neighboring medial belt area RM; similarly, recordings from A1 were pooled with those from MM. Total number of elements being an important factor in this type of study (Bizley et al. 2010; Miller and Recanzone 2009), pooling enabled us to increase the size of analyzed neuronal populations. We have shown previously that response properties of MM and RM neurons are quite similar to responses of cells in A1 and R, respectively (Kušmierek and Rauschecker 2009).

Thus the recorded neural population was divided into three cortical regions for analysis: a rostral region (rR) comprising the rostral core area R and rostromedial area RM, a middle region (rM) consisting of the primary core area A1 and middle medial area MM, and a caudal region (rC), which consisted of the caudo-medial area CM (Fig. 1). Region rR data originated from experiment 1 only, rC data from experiment 2 only, and rM data from both experiments 1 and 2. To account for the fact that experiments 1 and 2 differed in several respects, rM data sets coming from experiments 1 and 2 were analyzed separately and labeled as rM1 and rM2, respectively. This enabled us to distinguish differences between regions from differences between experiments.

Antero-posterior parcellation was based on best-frequency reversals, which were clear in both experiments 1 (Kušmierek and Rauschecker 2009) and 2. In experiment 2, medio-lateral delimitation of area CM from CL was also required. In monkey N, CL was separated from CM on the basis of longer latencies and higher selectivity to azimuth (Woods et al. 2006; Kušmierek and Rauschecker, unpublished observations). The resulting boundary ran approximately along the midline of the superior temporal plane, consistent with the placement of the CM/CL boundary in the anatomical literature (Smiley et al. 2007). In monkey B, caudal recordings were performed only in the medial half of the superior temporal plane, and neither long latencies nor increased azimuth selectivity was found laterally in the recorded area. Thus all recordings caudal to the best-frequency reversal at the caudal end of A1 in monkey B were considered to be from CM. The total number of neurons per cortical region was 159, 262, 95, and 69 for rR, rM1, rM2, and rC, respectively.

Neural data analysis. Spike time stamps were exported to MATLAB (MathWorks, Natick, MA) and processed with custom-made MATLAB scripts. Spike times were corrected for sound travel time from the loudspeaker to the monkey’s ear and, in experiment 1, for spike discriminator delay.

The analysis of neural data was performed in several variants. First, as mentioned above, it was done separately for experiment 1 or 2 or for both experiments combined. Second, we analyzed all cortical regions together or each of them separately to detect any between-region differences. Third, we performed computations on data from a single 160-ms temporal window starting at the stimulus onset, compared with a 160-ms window immediately preceding sound onset (pretrial), or we analyzed eight consecutive 20-ms windows within the stimulus starting at the stimulus onset and compared them with a 20-ms pretrial window. Fourth, analyses were performed for all stimuli (PT, BPN, MC, and ES) or for subsets of stimuli (PT and BPN or MC and ES).

The first stage of analysis followed the method of Kiani et al. (2007). Specifically, for each unit and each stimulus, average firing rate within 20-ms or 160-ms temporal windows was calculated across all stimulus presentations. For each unit, values of firing rate in response to the entire stimulus set were treated as a vector that was normalized by subtracting the mean and dividing by the vector’s Euclidean length. Correlation coefficients (r) between normalized population responses were calculated and visualized as similarity matrices. For natural stimuli MC and ES, representation of stimulus classes in the similarity matrices was quantified by comparing within-class correlation coefficients to between-class correlation coefficients (both between the given class and the other natural stimulus class and between the given class and all artificial stimuli, that is, PT and BPN) with a t-test.

Next, the normalized responses were arranged into a units × stimuli matrix, and hierarchical clustering of stimuli based on a measure of neural distance (1 − r) was calculated and visualized with dendrograms.

To quantify and compare the representation of stimulus classes in population responses, we assigned stimuli to k a priori categories of stimuli. The choice of actual k values and of categories is described in RESULTS. Next, we clustered the normalized firing rates into k clusters with the k-means procedure. The main measure obtained in this analysis was classification success, that is, the proportion of stimuli that were clustered into their a priori classes: proportion of correct classifications (PCC).

Different numbers of neural units were available for different cortical regions. This could skew the results of clustering because the quality of stimulus representation by a neural population may depend on the population size (Miller and Recanzone 2009). To avoid this potential confound, we performed k-means clustering on a subset of neurons from each region. The size of the subset was set to the number of units in the least numerous region of the analysis. Clustering was repeated 50 times with subsets drawn randomly from each region every time. The mean PCC (or mode PCC, see below) from these 50 repeats was taken as the representative value for a region.

The statistical significance of PCC values was assessed in two ways. First, to evaluate whether quality of clustering was higher than the baseline, the mode PCC obtained in each temporal window during the stimulus was compared with the mode PCC derived from the pretrial with a one-way Fisher exact probability test. When cortical regions were analyzed separately, their separate mode PCCs were compared with one pretrial mode PCC from all regions combined.

Second, when regions were analyzed separately, we assessed whether quality of clustering in a particular region deviated significantly from the “reference range.” The reference range was estimated by randomly reassigning the neurons to regions and repeating the k-means analysis in an identical way as described above to obtain reference mean PCCs. The number of reassignments was such that the number of reference PCCs was 400. For example, when 4 regions were analyzed, the analysis was run 100 times, each run producing 4 reference PCCs. The mean PCC of a region was considered significantly (P ≤ 0.05) above or below the reference range if it was outside the middle 95% of reference PCCs.

J Neurophysiol • doi:10.1152/jn.00793.2011 • www.jn.org
Again, this procedure could possibly be confounded by the unequal number of neurons per region. The reference range obtained by drawing neurons from the entire pool in a random fashion would be skewed toward values characteristic for the rM1 region (which contributed 45% of neurons to the pool) and less representative for the rC region (12%). Similarly, the reference range would be skewed toward values obtained from experiment 1, which provided 72% of the analyzed units. Thus the randomized pool of neurons was created by drawing the same number of neurons from each region. The number was equal to the mean number of neurons across the regions. Consequently, in each randomization a random subset of more numerous regions was used, whereas some neurons were redrawn from less numerous regions. Still, the subset size used for k-means clustering of randomized data was the same as for the original data, that is, equal to the number of neurons in the least numerous region.

Analysis of sound stimuli. The purpose of the analysis of sound stimuli was to determine whether classification of the stimuli based on responses of neural populations in the auditory cortex can be matched by classification based on acoustic properties of the stimuli. Only the first 160-ms segment of each stimulus was used, to match the information used for the neural analysis. Additional analyses were performed on 0–80 ms and 80–160 ms segments. To avoid cutoff transient, 5-ms linear on- and off-fades were applied to the segments for the analysis, except for the stimulus onset, which was faded in when the stimuli were originally prepared. All stimuli used in any of the experiments were analyzed: 10 PT, 20 BPN, 10 MC, and 12 ES. Three analysis approaches were used.

First, a log-frequency scale spectrogram was created by splitting each stimulus into 57 1/6-octave frequency bands (center frequencies: 64 Hz – 41.3 kHz) with a 16,384-point FIR filter and measuring RMS value (expressed in dB) in 33 (17 for 80-ms segments) consecutive nonoverlapping 5-ms bins. The Pearson correlation coefficient between spectrograms was used to measure the similarity between pairs of stimuli, which was then visualized in the form of a similarity matrix. Between-/within-class correlations of natural sounds were quantified, same as for the neural data. The similarity matrix was converted to a dissimilarity measure by subtracting from 1, and the dissimilarity matrix was subjected to multidimensional scaling (MDS) with the number of dimensions set to 4. As a result, each subject’s stimulus set was represented by a 4D vector, and between-subject distances were calculated. Two-dimensional Fourier transformation of this autocorrelation matrix was used as a basis for the clustering procedure.

The last step of the analysis of sounds was an attempt to classify the stimuli based on calculated acoustic parameters in a similar way as for the neural responses. To this end, each of eleven acoustic parameters (3 from direct measurements, 4 from MDS based on spectrogram dissimilarity, 4 from MDS based on modulation spectrum dissimilarity) was converted to Z scores, and k-means clustering (k = 4 for all sounds, k = 2 for natural sounds, and k = 2 for PT/BPN only) was performed 1) separately on each of three direct parameters (spectrum center of gravity, mean harmonicity, standard deviation of intensity); 2) on all four parameters derived from spectrum dissimilarity (combined, i.e., used as 4 variables in a single clustering procedure); and 3) on all four parameters derived from MS dissimilarity.

Furthermore, the clustering procedure was performed 1) on all 3 direct parameters combined; 2) on all 8 parameters derived from spectrum and MS dissimilarity combined; and 3) on all 11 parameters combined (3 direct measures, 4 parameters derived from spectrum dissimilarity, and 4 parameters derived from MS dissimilarity). As for the neural data, classification quality was quantified as PCC, and the PCC values were compared with those obtained from the neural data in three temporal ranges: 0–160 ms, 0–80 ms, and 80–160 ms.

Differences between experiments 1 and 2 in more detail. In experiment 1, a 19-mm-diameter round recording chamber was used, while in experiment 2 the chamber was oval and measured 19 × 38 mm. In experiment 1, auditory stimuli were played with an Audiophile 192 (M-Audio, Irwindale, CA) sound card, PA4 attenuator (TDT), SE 120 amplifier (Hafler, Tempe, AZ), and Reveal-6-two-way studio “monitor” loudspeaker (Tannoy, Coatbridge, UK), located 1.7 m in front of the monkey. The stimuli were played at 96-kHz sampling frequency, 16-bit resolution.

In experiment 2, auditory stimuli were played with a Power1401 MKII laboratory interface (CED), PA4 attenuator, SE 120 amplifier, and 400-312-10 3.5-in. one-way open-back car speakers (CTS, Elkhart, IN). Because spatial tuning data were also collected in experiment 2 (not reported here), the speakers were arranged in a vertical arc-shaped array (Crist Instruments) that was rotated automatically around the monkey chair with a Unidex 100 controller (Acrotech, Pittsburgh, PA) and a 3008MB3-HM (Acrotech) stepper motor under control of the Spike2 v.6.67 software program (CED). In the first stage of the experiment, a subset of stimuli was presented at azimuths of 0°, 45°, 90°, 135°, 180°, 215°, 270°, and 315° and at elevation 0 to estimate spatial tuning in the horizontal plane. In the second stage, the full set of stimuli was presented at best azimuth and at five elevations of −60°, −30°, 0°, 30°, and 60°. Data from experiment 2 analyzed in this article come from the second stage only. The distance between the loudspeakers and the monkey’s head was ~0.95 m. The stimuli were played at 48-kHz sampling frequency, 16-bit resolution.

The difference in loudspeaker size and quality as well as in sampling frequency (and, consequently, bandwidth) resulted in a noticeably lower playback quality in experiment 2 compared with experiment 1.

In addition to stimuli of interest, in experiment 1, bursts of white noise (equal power per hertz) and pink noise (equal power per octave) were presented, and a short four-note melody was used as behavioral target. In experiment 2, a white-noise burst was used as behavioral target. Although white-noise bursts were used in both experiments, they were excluded from analyses because of differing behavioral contingencies.

In experiment 1 each stimulus block consisted of 49 stimuli and 8 repetitions of the target, whereas in experiment 2 a block consisted of 35 stimuli plus 4 repetitions of the target. The high number of block repeats in experiment 2 (60) resulted from each block being played 12 times at each of 5 elevations.
In experiment 1 the location of the speaker and monkey chair was adjusted to minimize the influence of the room on low-frequency response; this could not be done for experiment 2 because of constraints of the spatial tuning study.

RESULTS

As expected, similarity matrices revealed no correlations in activity of neuronal populations during pretrial, with all correlation coefficients (r) close to zero (Fig. 2, top, Fig. 3). However, a clear correlation structure emerged during the first 160 ms of stimulus presentation (Fig. 2, bottom). Oblique lines of high r values reflect similarity of population responses to PT, 1/3-oct BPN, and 1-oct BPN of the same frequency, that is, frequency tuning. Dark blue colors visible between these lines demonstrate that population responses to distant frequencies were anticorrelated, whether within one bandwidth or across bandwidths. Apparently, stimulus frequency was the main factor determining the population response to artificial sounds, whereas bandwidth played less of a role.

Responses to natural sounds were typically correlated within each natural stimulus class. This was particularly visible for MC in experiment 2 and in combined data, with the first six of seven MC all evoking clearly similar population responses. Response correlation within the ES class was less pronounced but was still clearly noticeable.

These observations were confirmed quantitatively; Fig. 3 shows that within-class correlation coefficients of responses to each of the natural stimulus classes were clearly and significantly higher than between-class r values. This was true both when response similarity within MC or ES class was compared with similarity of responses between this class and artificial stimuli and when compared with similarity of responses between MC and ES classes.

Further confirmation of the findings based on similarity matrices was provided by hierarchical clustering (Fig. 4). Again, during the pretrial period, the stimulus structure was not reflected at all in the population responses (Fig. 4, top), which is by itself trivial but provides a control condition for the computational procedures.

During stimulus presentation, however, clustering of population responses replicated many features of the stimulus structure (Fig. 4, bottom). Responses to artificial stimuli (PT and BPN) always clustered with PT and BPN of the same frequency and usually fell close to PT and BPN of neighboring

Fig. 2. Similarity matrices showing correlations between population responses to pure tone (PT), band-pass noise (BPN), monkey call (MC), and environmental sound (ES) stimuli. Left: data combined from experiments 1 and 2. Middle: experiment 1 only. Right: experiment 2 only. Top: pretrial data from 160 ms preceding stimulus onset. Bottom: data from the first 160 ms of stimulus. Stimuli are coded by symbols along matrices’ edges, with explanation of symbols at bottom. Correlation coefficient (r) value is color-coded in the −0.5 to 1 range. Bottom right: white labels show sets of correlation coefficients that were used for quantitative analysis comparing mean coefficients of correlation between population responses to artificial sounds (A), ES, and MC (see Fig. 3).
frequencies, indicating frequency tuning in the auditory cortex. Bandwidth was not a robust organizing principle of the population responses, as shown by the fact that responses to PT and BPN of the same frequency always clustered very closely together. However, the effect of bandwidth on the response was still detectable. Of all PT/BPN triplets of the same frequency, in only one case did the response to PT and 1-oct BPN (i.e., 2 outlying bandwidths) cluster together first, while 1/3-oct BPN (the intermediate bandwidth) joined at a larger distance (Fig. 4, experiment 1, bottom, 3 lowermost branches). In all other cases, responses to neighboring bandwidths clustered together most closely (PT with 1/3-oct BPN or 1/3-oct BPN with 1-oct BPN), only later joined by the response to the remaining outlying bandwidth of the triplet. In quantitative terms, 1 of 16 (6.25%) response triplets clustered inconsistently with a proportional effect of bandwidth on response...
clustering. The 95% confidence interval of this proportion is <0.01% to 30.31% and is below the chance level of 33.3%.

(Only data from the experiment 1 and experiment 2 dendrograms were included in this calculation, because the combined data are not independent from experiment 1 and experiment 2 data.)

In neither experiment, nor in the combined data, were the responses to MC or to ES (or to natural stimuli in general) assigned to clusters that contained all responses to the class, and only responses to that class. However, in many cases they formed clear subgroups separated from responses to artificial sounds and, to some extent, to each other. In experiment 1, possibly because of a larger number of ES and MC involved, the picture was not unequivocal; still, a certain degree of clustering of responses to ES and MC can be observed. In the data from experiment 2, grouping is much clearer: All responses to MC except one were in a single cluster with two ES responses, with remaining responses to ES forming another big cluster. A similar picture emerged from the combined data, with only one response to ES clustering with most MC responses. The remaining four ES responses grouped closely, also with responses to high-frequency artificial stimuli. This picture contrasts with that seen for pretrial responses, where all classes were mixed (Fig. 4, top).

In summary, analyses of similarity matrices and of results of hierarchical clustering yielded several findings. Population responses to PT and BPN were determined mostly by the stimulus frequency. Bandwidth contributed less to the response. Responses to each of two classes of natural sounds formed distinguishable clusters and differed from responses to PT and BPN, and from responses to each other.

These findings were used to guide the choice of the cluster number (k) for the quantitative analysis of population responses to stimulus classes with k-means clustering. Not only does the procedure require a decision on an a priori number of clusters (k) into which the data will be grouped; also, for the purpose of quantification of clustering quality with PCC, the stimuli must be preassigned to k classes. Stimulus assignments resulting from clustering are then compared with these original assignments.

For natural stimuli, the choice was simple, as they formed two obvious natural categories: MC and ES. Because analyses of similarity matrices and of hierarchical clustering showed that bandwidth of the artificial stimuli only weakly affected population responses in our data set, we decided to split PT and BPN into classes based on frequency only. As frequencies were evenly spaced, the decision had to be partly arbitrary. Thus we split PT and BPN into two classes: low and high frequency, for a total k = 4. Specifically, when analyzing data from experiment 1 (9 PT/BPN frequencies), we placed four PT/BPN frequencies into the “low-frequency” class (125 Hz–1 kHz) and five frequencies into the “high-frequency” class (2–32 kHz), while the seven PT/BPN frequencies of experiment 2 (combined) were divided into ranges of 250 Hz to 1 kHz (3 members) and 2 kHz to 16 kHz (4 members). We have also tested other k values: k = 3 (classes: all PT + BPN, MC, ES) and k = 5, with PT/BPN frequencies split into low, middle, and high range (experiment 1: 125–500 Hz, 1–4 kHz, and 8–32 kHz; experiment 2 (combined): 250–500 Hz, 1–4 kHz, 8–16 kHz). Separate analyses on responses to natural sounds only (MC vs. ES) and to PT and BPN only (low vs. high frequencies) were performed with k = 2, and the latter was also tested with k = 3 (middle frequency range added).

Figure 5A shows mean classification success (PCC) values from experiment 1, experiment 2, and combined experiments, for k = 4. Classification of responses recorded prior to stimulus presentation yielded mean PCC values from 0.366 to 0.395, considered chance value. Clustering of responses recorded during stimulus presentation replicated the original class structure with much higher accuracy of 0.585–0.905. Mode PCC for stimulus responses was significantly higher than mode PCC for pretrial responses in combined data and in experiment 2 (P < 0.05, 1-way Fisher exact probability test), as well as for experiment 1 when analyzed with k = 5 (data not shown). These results demonstrate that when a small number of discrete clusters are imposed on the data (as opposed to the basically continuous approach of hierarchical clustering) classification of population responses can still recreate original stimulus classes with reasonable accuracy.

The next step was to apply k-means clustering to population responses recorded from each cortical region separately. Figure 5B shows PCC values calculated for each region for pretrial (160 ms preceding stimulus onset) and the first 160 ms of stimulus. The result obtained with data pooled across regions is confirmed and strengthened: Here, mode PCC calculated from the 160-ms stimulus period was significantly higher than pretrial PCC also for experiment 1 and for all regions. Still, differences between the regions were small in the analysis that used 160-ms temporal windows, and none of them appeared to be significant.

A more compelling picture emerged from the next analysis, wherein we again calculated PCC for each region separately, but this time in eight 20-ms windows covering the first 160 ms of the stimulus period, with a single 20-ms window preceding stimulus onset used to establish chance level (Fig. 5C).

Classification based on responses from rC increased from chance level faster than classification based on other regions: In the first temporal window of the stimulus (0–20 ms), mode PCC from rC but not from other regions was significantly higher than pretrial PCC. Later in time, from ~60 ms past stimulus onset, classification based on responses from rC tended to be poorer than that in rM2, but the difference was slight and there was much overlap.

The most interesting results, however, emerged from region rR. The development of classification success was notably slower in rR than in the other regions: In the first 20 ms of the stimulus period, PCC from rR remained significantly below the reference range, barely different from pretrial level. However, starting from ~60–80 ms past stimulus onset, clustering based on the population responses from rR became clearly more successful in replicating the original classes than clustering based on response from any other region. For the four temporal windows starting at 80 ms after stimulus onset, the PCC ranges calculated from rR never overlapped with those from region rM1 within the same experiment (experiment 1: rR 0.783–0.915, rM1 0.582–0.742; combined: rR 0.755–0.949, rM1: 0.600–0.717). Within this temporal range, mode PCC calculated from rR was significantly above pretrial level in each window, which was not always the case for rM1. In most cases, mean PCC for rR was above the upper significance limit of the reference range, but this was never true for rM1; actually, in
one case (experiment 1 at 100 ms) mean PCC for rM1 was below the lower significance limit.

From the inspection of similarity matrices (Fig. 2) and dendrograms (Fig. 4) we inferred that artificial stimulus frequency was a powerful factor that determined population responses. In the $k$-means analyses described above, the natural stimuli (ES and MC) were subjected to clustering together with the artificial stimuli (PT and BPN), the latter being classified based on frequency. Thus the question remains whether the high classification success found in rR with $k$-means analyses arose only from very accurate classification of artificial stimuli, or region rR also excelled in classification of ES versus MC stimuli. Many lines of evidence show that the latter was the case. First, the artificial stimuli constituted 57% or 64% (experiment 1 and combined, respectively) of stimuli, whereas >90% of stimuli were correctly classified based on rR responses at some temporal windows. Second, correlations of population responses within the MC and ES classes were very significantly higher than between these classes or between any of these classes and artificial sounds for rR in temporal windows at 80–160 ms, at which classification based on rR responses was particularly successful (Fig. 6, left). Thus there was a significant potential for MC vs. ES classification in rR responses. For comparison, correlations within the ES class did not exceed correlations between MC and ES or between ES and artificial stimuli in rM1 responses as reliably as in rR...
of PCC for all stimuli in rC (Fig. 5C) was caused primarily by accurate classification of PT/BPN frequency, whereas the late advantage of rR over rM in the 80–160 ms window was driven mostly by improved discrimination of MC from ES.

Figure 8 shows the result of test k-means analyses with alternative cluster numbers. Clustering all stimuli with \( k = 3 \) (all PT+BPN, MC, ES; Fig. 8A) resulted in most successful classification in rR at late time windows, similar to \( k = 4 \), although much less pronounced and apparent only late after stimulus onset. Early advantage of rC was almost absent, consistent with the finding that it was mainly driven by frequency discrimination. Clustering of all stimuli with \( k = 5 \) (low-, mid-, high-frequency PT+BPN, MC, ES; Fig. 8B) basically replicated the pattern seen with \( k = 4 \), although both the classification accuracy and separation of regions were weakened, suggesting that the neural populations did not distinguish the middle frequency PT/BPN class very well. This suggestion was confirmed by analysis of clustering of PT/BPN into three frequency classes (Fig. 8C): Compared with results of clustering with \( k = 2 \) (Fig. 7B) both the PCC values and certain effects, such as accurate early classification of PT/BPN in region rC, were diminished. In summary, our choice of \( k = 4 \) for analysis of all stimuli and \( k = 2 \) for analysis of PT/BPN appears to be most appropriate.

The analyses of neural data described above provide evidence that population responses in early auditory cortex, specifically in region rR, carry sufficient information to allow for correct classification of stimuli that evoked these responses with an accuracy exceeding 90\%, at 80 ms or later after stimulus onset. On the other hand, neural responses in region rC supported a very accurate classification of PT/BPN frequency within the first 20 ms after stimulus onset, but not later. It would be interesting to know whether this degree of classification accuracy can be supported by acoustic properties of stimuli. Therefore, we measured a number of acoustic parameters of the stimuli and applied to them classification methods similar to those we used for the neural data.

Correlations between log-frequency spectrograms and MS were visualized as similarity matrices (Fig. 9, A and C). A prominent feature of the similarity matrix calculated from spectrograms (Fig. 9A) was the presence of oblique lines in the upper left area, showing similarity of spectrograms calculated for PT/BPN of the same frequency but different bandwidth. The picture resembled very closely the PT/BPN area seen in similarity matrices derived from neural data (Fig. 3). Correlations within the group of natural stimuli were strong (lower right), but the separation of the MC and ES classes was not clear. Within-MC correlation coefficients were particularly high, but within-ES correlation coefficients appeared to be similar to correlation coefficients between ES and MC. These observations were confirmed quantitatively: Mean within-MC correlation coefficient \( (r) \) values were significantly higher than \( r \) values between MC and PT/BPN and than those between MC and ES. Within-ES \( r \) values, however, were significantly higher than \( r \) values between ES and PT/BPN but did not differ from correlation coefficients between MC and ES (Fig. 9B), in contrast to correlations derived from neural data (Fig. 3).

The similarity matrix calculated from modulation spectra did not show the same characteristic patterns for the PT/BPN stimuli (Fig. 9C) as did similarity matrices calculated from spectrograms or neural data. The connection between PT/BPN

---

**Fig. 6.** Quantification of correlations between population responses to MC and ES for region rR (left) and rM1 (right) in 4 temporal windows from 80 to 160 ms after stimulus onset (rows). Data from combined experiments. For conventions and explanations of symbols, see Fig. 3.

responses (Fig. 6, right), indicating that the difference in general classification ability between rR and rM1 stemmed at least partially from different capability to classify natural sounds.

Finally, Fig. 7A shows PCC for all regions calculated for MC versus ES classification only \( (k = 2) \), for both 160-ms and 20-ms window sizes. In contrast to rM1 (as well as rM2 and rC), only classification based on rR responses was significantly better than classification based on pretrial data for the 0–160 ms window and for all 20-ms windows starting from 80 ms past stimulus onset. In that time period, classification success for responses from rR ranged from 0.867 to 0.917, compared with 0.722 to 0.808 for responses from rM1.

Classification of PT and BPN data (low vs. high frequencies, \( k = 2 \); Fig. 7B) again showed somewhat more accurate classification in rR than in rM1, although the difference was relatively small and not necessarily limited to a particular time range. The effect of experiment was strong, with classification based on experiment 2 data being significantly less accurate than that based on experiment 1 data from ~60 ms on. Interestingly, classification of PT and BPN based on frequency was very accurate in region rC within the first 20-ms period. Comparison with Fig. 7A allows us to suppose that the fast rise
of same frequency but different bandwidths was missed. On the other hand, pure tones seemed to be reasonably well separated from BPN.

At first glance, separation of the ES and MC classes was visible in the similarity matrix. Indeed, both within-MC and within-ES \( r \) values were significantly different from \( r \) values between ES and MC, as well as from \( r \) values between either natural sound class and artificial sounds (Fig. 9D). However, within-MC \( r \) values were actually lower than those between MC and ES, showing that modulation spectra of various MC were on average more similar to modulation spectra of ES than to modulation spectra of other MC. This shows that differentiation of MC as a stimulus class cannot be supported by differences and similarities between MC (as measured with the acoustic parameters that we chose) and is in contrast to the findings from the neural data (Figs. 2 and 3).

Finally, we attempted to classify the stimuli based on acoustic parameters using \( k \)-means clustering. As with neural data, we classified all stimuli into \( k = 4 \) clusters, while natural stimuli (MC and ES) and PT/BPN stimuli were classified into \( k = 2 \) clusters each. The PCC values calculated based on acoustic parameters were compared with PCC values obtained from the neural data in the entire analysis window (0–160 ms past stimulus onset) as well as separately for 0–80 ms and 80–160 ms periods (Fig. 10).

When all stimuli were subjected to \( k \)-means clustering (Fig. 10, top), PCC derived from any acoustic parameter, direct or derived from spectrograms or modulation spectra, or from any

---

**Fig. 7.** Classification of subsets of sound stimuli based on neural population responses. Data from each brain region (rR, rM1, rM2, rC) analyzed separately. Left: pretrial data from 160 ms preceding the stimulus onset on left, stimulus data from the first 160 ms of stimulus on right. Right: pretrial data from 20 ms preceding the stimulus onset are shown on left, followed by stimulus data from 8 consecutive 20-ms windows covering 0–160 ms of stimulus. A: classification of natural (MC and ES) stimuli: mean classification success (proportion of correct classifications) with \( k \)-means clustering (\( k = 2 \)). B: classification of PT and BPN stimuli: mean classification success (proportion of correct classifications) with \( k \)-means clustering (\( k = 2 \)). For detailed legend, see Fig. 5.

**Fig. 8.** Classification of sound stimuli based on neural population response with different \( k \) values: mean classification success (proportion of correct classifications) with \( k \)-means clustering. Within each panel, "pretrial" data from 20 ms preceding the stimulus onset are shown on left, followed by "stimulus" data from 8 consecutive 20-ms windows covering 0–160 ms of stimulus. Data from each brain region (rR, rM1, rM2, rC) are analyzed separately. See Fig. 5 for details. A: classification of all stimuli with \( k = 5 \) (low-frequency PT/BPN; middle frequency PT/BPN; high-frequency PT/BPN, MC, ES). Compare to Fig. 5C, left. B: classification of all stimuli with \( k = 3 \) (all PT/BPN, MC, ES). Compare to Fig. 5C, left. C: classification of PT/BPN stimuli with \( k = 3 \) (low-frequency PT/BPN; middle frequency PT/BPN; high-frequency PT/BPN). Compare to Fig. 7B, right.
combination of the acoustic parameters, remained below 0.7 in each time window tested. In the 0–80 ms time window, the values of the parameters largely overlapped with the ranges of PCC values based on neural data. In the 80–160 ms window, however, the PCC values derived from acoustic parameters at best reached values comparable to those derived from rC and rM recordings, while remaining far below the range of values produced by analysis of rR population response. For clustering of MC and ES stimuli only (Fig. 10, middle), the comparison of PCC derived from acoustic parameters to PCC calculated from neural responses yielded a very similar picture: Some acoustics-based PCC matched neural-based PCC in all regions in the 0–80 ms window as well as in rM and rC in the 80–160 ms window, but they were clearly below rR-based PCC in the 80–160 ms window.

A contrasting pattern emerged from analysis of PT and BPN stimuli (Fig. 10, bottom): PCC derived from several acoustic parameters (spectrum center of gravity, parameters derived from spectrogram differences, or combination thereof with parameters derived from spectrotemporal modulation differences) reached very high values that matched or closely approached (and in some cases exceeded) the ranges of PCC based on neural responses in any region and window. An additional finding is that PCC based on neural responses to PT and BPN in the 80–160 ms window were clearly lower in experiment 2 than in experiment 1.

Taken together, classification of all sounds (or natural sounds) based on acoustic features yielded accuracy values that were at best comparable to the values provided by classification based on population responses from regions rC, rM1, and rM2. On the other hand, classification based on population responses recorded in rR within 80–160 ms past stimulus onset clearly surpassed not only classifications derived from the other regions’ responses but also those obtained from acoustic parameters, whether analyzed separately or in combinations.

DISCUSSION

We applied neural population analysis to study the representation of stimulus identity in rhesus monkey auditory cortex. To our knowledge, such techniques have not been used for this purpose before, although population analysis has proved successful previously in studies of the representation of sound location in auditory cortex (Miller and Recanzone 2009; Recanzone et al. 2010), of the representation of visual stimuli in visual cortex (e.g., Kiani et al. 2007; Kriegeskorte et al. 2008), and of periodicity discrimination in ferret auditory cortex (Bizley et al. 2010).

Methodological considerations. Kriegeskorte et al. (2008) emphasized that the methods they used assumed no or little structure in the data; only after classification was performed was correspondence between resulting clusters and natural categories found. A similar approach was taken by Kiani et al. (2007). We initially followed this principle when we studied similarity matrices and dendrograms. In this way we confirmed the existence of a particular structure in the neural data and that this structure reflected our preconceived notions about stimulus categories. In the next step, however, we used k-means clustering, with a priori structure explicitly sought in the data. This approach allowed us to achieve our main goal: quantifying how well stimulus structure is represented in different cortical areas and at different time points. It has to be pointed out that, should we have imposed

Fig. 9. A: similarity matrix showing correlations between log-frequency spectrograms of all stimuli. B: quantification of correlations between log-frequency spectrograms of MC and ES. C: similarity matrix showing correlations between modulation spectra of all stimuli. D: quantification of correlations between modulation spectra of MC and ES. Note that while in D the mean correlation coefficient within the MC class (MC) differs significantly from the mean correlation coefficient between MC and ES classes (MCvES), the value for MC is lower than for MCvES (gray dashed line). Conventions same as in Fig. 2 and Fig. 3 except for scale in C and D.
an inappropriate structure on the clustering procedure, we would likely not have seen any difference between regions and classification success scores would be low. What we saw instead were clear between-region differences and high classification scores in rR (Fig. 10). Thus the stimulus structure we imposed on k-means clustering appears to match a structure actually represented in the investigated cortical areas and, more importantly, a structure whose representation in anterior areas (region rR) is more accurate than in more posterior areas (regions rM and rC). While we cannot ensure that the structure was the most appropriate, our results, together with the choice of categories, which arguably reflects natural categorization (i.e., vocalizations, environmental sounds, low frequencies, high frequencies), indicate that the structure imposed on k-means clustering was meaningful in terms of cortical sound processing.

In addition to employing k-means clustering to quantify classification accuracy, another important component was to perform classification in successive short temporal windows, in addition to analysis over the entire stimulus duration (in our case, the duration of the shortest stimulus). This has proved very successful: When firing rates were averaged over the first 160 ms of the stimulus, virtually no differences in classification success between regions were detected (Fig. 5B); however, analysis in 20-ms windows not only revealed a clear divergence between region rR and the other regions (Fig. 5C) but allowed us to trace the temporal dynamics of classification ability and discuss its origins (see further below).
A potential confound comes from the fact that the neural data came from two different experiments, with different stimulus presentation techniques and neural recording techniques and slightly different stimulus sets, e.g., data from regions rC and rR came from different experiments. However, our key analyses were replicated within experiment, that is, data from region rC were compared with those from region rM2 (region rM recorded in the same experiment as rC) and data from rR to those from rM1 (region rM recorded in the same experiment as rR; Fig. 5C). In this way not only were the results from the combination of the experiments validated, but thanks to “anchoring” both experiments in region rM we gained insight into the effects of experimental conditions under which neural data were gathered on stimulus classification based on these data. We did not systematically explore the effects of all differences between the experiments on classification, but we can speculate that reduced stimulus quality due to the use of small loudspeakers and/or lower sampling frequency (consequently, lower stimulus bandwidth and/or lower stimulus intensity in experiment 2 resulted in neural data that provided support for somewhat less accurate classification than data from experiment 1. This effect was not offset by supposedly more reliable estimation of firing rates in experiment 2, in which each stimulus was played 60 times compared with 10–13 times in experiment 1. Another potentially significant factor that could have influenced classification accuracy in experiment 2 was spatial variation of stimulus locations. The stimuli were presented at various azimuths (whereas in experiment 1 they were always played only at azimuth 0°)—still, for each neural unit, data used for the analysis were recorded at one azimuth only. Changes of elevation, which did occur during recording from each unit in experiment 2, could have modulated the response and affected classification. However, when k-means analysis was run on a subset of data recorded at 0° elevation only, the resulting PCCs were not better than those obtained from all elevations combined (data not shown). In conclusion, quality of stimulus presentation should be carefully attended to when stimulus classification is investigated.

It may appear surprising that the PCC values derived from the pretrial period were higher than expected chance level (0.25 for the pretrial period were higher than expected chance level attended to when stimulus classification is investigated. However, our key analyses were replicated within experiment, that is, data from region rC were compared with those from region rM2 (region rM recorded in the same experiment as rC) and data from rR to those from rM1 (region rM recorded in the same experiment as rR; Fig. 5C). In this way not only were the results from the combination of the experiments validated, but thanks to “anchoring” both experiments in region rM we gained insight into the effects of experimental conditions under which neural data were gathered on stimulus classification based on these data. We did not systematically explore the effects of all differences between the experiments on classification, but we can speculate that reduced stimulus quality due to the use of small loudspeakers and/or lower sampling frequency (consequently, lower stimulus bandwidth and/or lower stimulus intensity in experiment 2 resulted in neural data that provided support for somewhat less accurate classification than data from experiment 1. This effect was not offset by supposedly more reliable estimation of firing rates in experiment 2, in which each stimulus was played 60 times compared with 10–13 times in experiment 1. Another potentially significant factor that could have influenced classification accuracy in experiment 2 was spatial variation of stimulus locations. The stimuli were presented at various azimuths (whereas in experiment 1 they were always played only at azimuth 0°)—still, for each neural unit, data used for the analysis were recorded at one azimuth only. Changes of elevation, which did occur during recording from each unit in experiment 2, could have modulated the response and affected classification. However, when k-means analysis was run on a subset of data recorded at 0° elevation only, the resulting PCCs were not better than those obtained from all elevations combined (data not shown). In conclusion, quality of stimulus presentation should be carefully attended to when stimulus classification is investigated.

It may appear surprising that the PCC values derived from the pretrial period were higher than expected chance level (0.25 for \( k = 4 \), 0.5 for \( k = 2 \), etc.). We attribute this effect to the method used to calculate PCC. The k-means procedure divides the data into clusters and assigns an arbitrary label to each cluster. It does not, however, link these cluster labels to original class labels (ES, MC, etc.). Such link is necessary to calculate PCC. We solved this problem by permuting the cluster labels and calculating PCC for each possible configuration of links between cluster labels and original labels. The highest of these PCCs was taken as the result of clustering. In case of successful clustering, that is, when stimuli belonging to different original classes are assigned to different clusters, this approach guarantees identifying the classification success as a high PCC value. However, if clustering cannot succeed and stimuli belonging to different classes are randomly mixed in the resulting clusters (this is expected for pretrial data), the method skewes the PCC toward values somewhat higher than expected by chance. It has to be noted that the approach is conservative: It reduces the difference between chance level (defined as PCC based on pretrial) and PCC obtained from successful clustering based on the stimulus period data.

We confirmed that these “above-chance” PCC values seen in the pretrial period are a by-product of the computational procedure and not a result of a systematic error in data collection: A test run of the analysis on data in which random values were substituted for neural firing rate produced PCC values similar to those based on neural recordings from the pretrial period (data not shown).

It has to be noted that our “neural populations” were not recorded simultaneously but the recordings were collected successively over many months, which precludes analysis of correlations between responses of individual units. The actual significance of leaving out analysis of correlations is unclear; they have been shown to influence decoding in auditory cortex, but the direction of that influence depends on the decoding method utilized and on correlation structure, which is largely unknown (Averbeck et al. 2006; Bizley et al. 2010; Jenison 2000). Obviously, to analyze correlations, multiple units must be recorded simultaneously with multiple electrodes, which limits the population size available for the analysis even if difficulties in using multielectrode arrays in cortical areas located within the lateral sulcus were overcome (Hackett et al. 2005). Another approach that we decided not to use in this study is analysis of trial-to-trial variability of neural responses, which has been shown to play a role in cortical processing (e.g., Carandini 2004, Scaglione et al. 2011), and to influence results of neuron-to-neuron correlation analysis (Ventura at al. 2005).

**Band-pass noise.** Bandwidth of BPN (vs. PT) has been shown to influence neural responses in the auditory cortex (e.g., Kajikawa et al. 2011; Kuśmierek and Rauschecker 2009; Petkov et al. 2006; Rauschecker et al. 1995; Rauschecker and Tian 2004); also, discrimination of a tone from a 1/3-octave or 1-octave noise does not seem to be difficult, at least for human listeners (although we are not aware of any formal tests in monkeys; however, see below). It is therefore surprising that only a small effect of bandwidth on population response was found in the present study (see Figs. 2 and 4 and accompanying description in RESULTS). It should be noted that the most robust effects of bandwidth have been found in the lateral belt (Petkov et al. 2006; Rauschecker et al. 1995; Rauschecker and Tian 2004), which was not included in the present study. Our recordings were derived from core and medial belt, and BPN preference in the medial belt over core, although demonstrable, appears to be less pronounced than in the lateral belt (Kajikawa et al. 2011; Kuśmierek and Rauschecker 2009; Petkov et al. 2006). Thus the picture might be different if the present study was replicated in the lateral belt or in cortical areas further anterior or lateral. Furthermore, behavioral data from experiment 1 offer an insight into perception of bandwidth and frequency in rhesus monkeys. In this experiment, the behavioral target was a short “melody” consisting of a rapid succession of four 125-ms pure tones; the first tone’s frequency was 523 Hz, less than a semitone from the frequency of a nontarget 500-Hz stimulus. Consequently, the monkeys often responded to the 500-Hz tone. In addition, they quite often responded to 1/3-octave and 1-octave noise bursts centered at 500 Hz and much more rarely to tones or noises at neighboring frequencies of 250 or 1,000 Hz, let alone more distant frequencies (Fig. 11). Apparently, the perceptual difference between a PT and a BPN burst at the same frequency was smaller than a step to a tone (or...
than in A1 and MM in macaques (Kuśmierek and Rauschecker 2011; also Kuśmierek and Rauschecker, unpublished observations); the same was found by Kajikawa et al. (2005) in the marmoset.

**Stimulus classification in the 0–20 ms window.** Within the first 20 ms after stimulus onset, the classification capability measured by PCC developed differently in the cortical regions (Fig. 5C). Data from region rC supported high classification accuracy, comparable to accuracy in any subsequent time window. On the other hand, responses from rR barely differed from chance level in the 0–20 ms window, and the responses from region rM fell in between. Latencies in area CM are shorter than latencies in area A1 in rhesus monkeys (Scott et al. 2011; also Kuśmierek and Rauschecker, unpublished observations); the same was found by Kajikawa et al. (2005) in the marmoset. Latencies in R and RM were shown to be longer than in A1 and MM in macaques (Kuśmierek and Rauschecker 2009; Recanzone et al. 2000a; Scott et al. 2011) as well as in marmosets (Bendor and Wang 2008). The differences between classification capabilities found in the 0–20 ms window likely reflect this posterior-to-anterior progression of neural latencies in the core and medial belt. In the first 20 ms of the stimulus, fewer region rM neurons than region rC neurons began firing and, consequently, contributed to classification, and only a few rR neurons were active at that time.

When classification of natural stimuli was analyzed separately from classification of PT/BPN, the high classification accuracy based on rC responses early after stimulus onset was apparent for the latter but not the former analysis (Fig. 7). This contrasts with the findings in rR; there, from 80 ms after stimulus onset, high classification accuracy was found in all analyses. Moreover, it clearly exceeded classification accuracy in rM1 when all stimuli or natural stimuli only were clustered, but less so in PT/BPN analysis (Figs. 5 and 7). Apparently, although region rC can provide a quick estimation of sound identity, it can do so only if sounds can be distinguished by simple features such as frequency.

**Object categories or combinations of features?** Finding category specificity in neural responses always raises an important question: Does this effect correspond to genuine object categorization, or simply to low-level stimulus features? A direct answer can be given if sharp categorical boundaries are found in responses to gradual continua of low-level features: invariance of responses to large changes of features irrelevant to object identity and specificity of responses when small changes are introduced to highly relevant features. This approach requires more knowledge about feature relevance and on parameters of object invariance than is available for auditory perception in macaques. Therefore, we have chosen a simpler alternative approach (Kiani et al. 2007), that is, we attempted to classify the stimuli based on low-level acoustic features and compared the results with results of classification based on neural data. An important caveat is that a failure of feature-based classification to match neural data-based classification does not prove that the cortical region in question operates on representation of sound objects beyond simple combinations of features. Another explanation might be that crucial low-level features were not entered into the analysis.

We have chosen three direct stimulus features that covered a wide range of qualities: Spectrum center of gravity approximated which cochlear channels were activated by the sound; harmonicity measured whether a stimulus was more periodic or more noisy; and standard deviation of harmonicity revealed amplitude modulation structure. In addition, parameters were derived from dissimilarity of simple frequency-time representation (log-frequency spectrogram) and from dissimilarity of spectrotemporal modulation spectra (Cohen et al. 2007; Singh and Theunissen 2003).

Our results show that classification of PT/BPN frequency in all regions may be driven by low-level features, as the performance of stimulus classification based on low-level features overlapped with classification performance based on neural responses (Fig. 10). This is true also for the high classification success found for these stimuli in region rC early after stimulus onset. Similarly, in the 0–80 ms period and in all regions, classification of natural (MC and ES) stimuli as well as classification of all stimuli, both moderately successful, may be supported by low-level features. A similar picture was found for classification of natural stimuli and all stimuli in the 80–160 ms period in regions rM and rC. What clearly stands out is classification of natural stimuli (and all stimuli) in region rR in the 80–160 ms period: Here, population responses supported much higher classification performance than achieved by analysis of any acoustic features.

In other words, regions rC and rM, and region rR early after stimulus onset, can only support classification as good as can be achieved by utilizing similarities and dissimilarities of low-level acoustic features. On the other hand, later after stimulus onset, region rR provides a basis for classifications beyond simple comparisons of low-level features, putatively by using nonlinear combinations of features. Although more research is needed before definite conclusions can be drawn, this result may suggest that already in region rR the stimulus representation begins to shift from feature based toward object based.

**Processing streams in the auditory cortex.** The function of the dorsal stream of auditory cortical processing compared with the ventral stream remains less sharply defined. Although processing of space (i.e., sound source location) has been emphasized (e.g., Rauschecker and Tian 2000; Romanski et al. 1999), more recent proposals include sensorimotor integration with the purpose of learning and control of auditory production (Rauschecker 2011; Rauschecker and Scott 2009). Specifically, parietal regions are fed by posterior areas of auditory cortex.
with a fast, temporally precise, but relatively rough “primal sketch” of ongoing auditory information (Rauschecker 2011). Properties of area CM, as shown in the present study, fit this description well: Neural data collected from CM support stimulus classification at an earlier time point than data from any other region (Fig. 5C), but high classification accuracy was only achieved for classification based on a simple low-level feature of frequency (Fig. 7). While a common denominator can be found for both functions postulated for the dorsal stream (i.e., spatial processing and audio-motor integration; Rauschecker 2011), emerging properties of areas CM and CL hint at a possible division of labor. The spatial sensitivity appears to be more pronounced in CL (Miller and Recanzone 2009; Tian et al. 2008; Poremba et al. 2004; Remedios et al. 2009; Romanski et al. 2006) even argued that areas corresponding to our region rR are outside the ventral stream, computations may possibly extract more complex combinations of features, better suited for certain types of classification. These computations might include applying previously learned abstract rules governing classifications that cannot be easily accomplished based on low-level feature similarity. For example, a sound of a vacuum pump (an ES) and a pant-threat MC could be classified together by regions such as rM because of similarity of low-level features, e.g., frequency content, bandwidth, or amplitude modulations. On the other hand, regions of the ventral stream might rely on specific combinations of features to put these two stimuli into separate categories of environmental sounds and conspecific vocalizations, and, within these categories, group them with stimuli characterized by differing low-level features, such as the sound of a monkey-pole latch, or a high-frequency scream.

Further, rM and rR should not be considered as hierarchically connected stages of the ventral stream. Although response latencies might suggest a hierarchical relationship between rR and rM (see, e.g., Kušmírek and Rauschecker 2009; Pingst and O’Connor 1981; Recanzone et al. 2000a; Scott et al. 2011), it has been shown that these areas are rather organized in parallel: Responses in R (the core component of our region rR) cannot depend on integrity of the core component of rM, that is, A1 (Rauschecker et al. 1997). On the basis of anatomical connections in the marmoset, de la Mothe et al. (2006) even argued that areas corresponding to our region rR
belong to a different processing subsystem than those correspond- ing to region rM and, possibly, rC.

In the present study, we found better discrimination between stimulus classes (including ES) in rR versus rM in neural populations, whereas previously, using a linear discriminator, we demonstrated worse discrimination within the ES stimulus class in rR than in rM in individual units (Kuśmierenk and Rauschecker 2009). How can these results be reconciled? As we mentioned above, discrimination between classes may often require applying rules based on complex interactions of low-level features, whereas analyzing low-level features may lead to confusions. Within classes, however, in some cases low-level features may support accurate classification. Within the ES class such features may be derived from temporal structure, as analysis of temporal relationships is important for identification of certain environmental sounds (Gyggi et al. 2004; Warren and Verbrugge 1984). Regions rM and rC, with their fast and temporally precise responses, may represent such temporal information at the individual unit level better than rR.

An intriguing finding in the present study was that classification based on region rR data initially was only as good as classification based on data from other regions, or as classification based on acoustic parameters. But ~60–80 ms after stimulus onset, rR-based PCC began to clearly exceed those derived from other regions, or from acoustic measurements. This delay might have been due to time needed by rR neurons to integrate incoming information and perform computations needed to discriminate stimuli at a high accuracy level, beyond what was possible based on simple analysis of acoustic features. We have shown previously that (for temporally structured stimuli) the best linear discriminator bin width, which can be interpreted as a measure of temporal integration scale, is on the order of 40–50 ms in area R and 20–30 ms in A1 (Kuśmierenk and Rauschecker 2009). From synchronization cutoff frequencies, Scott et al. (2011) estimated temporal integration windows to be ~100 ms in R and 20–30 ms in A1.

Instead of (or in addition to) the computation in local circuits, arrival of feedback or “top-down” information from areas further downstream could have contributed to the late (60–80 ms past stimulus onset) surge in classification accuracy. Based on latency estimates from Kikuchi et al. (2010), candidates for the source of such feedback information include area RT, reported to respond with mean latencies of 40–70 ms, while area RTp would have to be excluded (70- to 110-ms latencies). Romanski and Hwang (unpublished observations), on the other hand, found mean latencies exceeding 100 ms in the VLPFC, but a subset of cells responded to sounds with latencies as fast as 50 ms; thus these neurons could theoretically be the source of increased classification accuracy found in region rR in the present study. It has to be noted as well that relying on latencies measured across different studies is somewhat risky. For example, Kikuchi et al. (2010) reported mean latencies in area A1 to be on the order of 40–60 ms, far longer than found by us [median 17–20 ms; Kuśmierenk and Rauschecker (2009)] or by others [Scott et al. (2011): median 20 ms; Recanzone et al. (2000a): mean 32.4 ms].

The relatively slow emergence of between-class classification accuracy in rR may appear to contradict the earlier appearance of global (between class) than fine (within class) information in visual areas of monkey inferior temporal cortex (Sugase et al. 1999). The crucial difference between this study and ours lies in the distinguishability of stimulus classes. Sugase et al. (1999) compared three classes: monkey faces, human faces, and plain geometric shapes. These three classes, arguably, could be quite easily distinguished based on relatively simple low-level features. However, our MC and ES classes overlapped much in terms of simple acoustic features (Fig. 10; Kuśmierenk and Rauschecker 2009). In this way, the MC versus ES distinction may be actually more similar to Sugase et al.’s (1999) fine discrimination of facial expression, as it requires detailed analysis of combinations of features. Discrimination of low-frequency versus high-frequency PT/BPN may be a closer equivalent of Sugase et al.’s (1999) global discrimination, because it can be accomplished by using simple low-level features (Fig. 10). Indeed, frequency discrimination was performed very quickly in region rC (Fig. 7). However, this parallel should be taken with caution, because region rC belongs to the auditory dorsal processing stream, whereas Sugase et al. (1999) studied the visual ventral stream only.

**Methodological notes on linear discriminator analysis.** The results of linear discriminator analysis reported by Recanzone (2008), as well as by Russ et al. (2008), and presented again by Recanzone (2011), were somewhat perplexing. In both studies, linear discriminator performance was found to reach very high values (80–90%) at a very short bin width of 2 ms.

Such a result implies that neurons typically produced highly replicable (with an accuracy of 2 ms or better) and clearly different spiking patterns to all stimuli and that they were driven by almost all these stimuli. Neuronal firing illustrated by raster plots presented by Recanzone (2008) and Russ et al. (2008) do not seem to show these characteristics. Furthermore, replicability of spiking patterns could be based on one of two mechanisms: a strict abstract temporal code or following spectrotemporal features in the ongoing stimulus with 2-ms accuracy. The former is unlikely, at least in auditory cortex (Kuśmierenk and Rauschecker 2009); the latter would suggest the neurons’ capability to lock to stimulus modulations on the order of 500 Hz, which has not been demonstrated in the auditory cortex (Bendor and Wang 2008; Malone et al. 2007; Oshurkova et al. 2008; Scott et al. 2011). Furthermore, various authors used the linear discriminator technique or comparable classification methods to study cortical responses to sounds, and the best bin width was never as small as 2 ms. The values ranged from 5 to 50 ms in the auditory cortex of monkeys and ferrets (Kuśmierenk and Rauschecker 2009; Malone et al. 2007; Schnupp et al. 2006; see also below) and was ~60 ms in the prefrontal cortex (Averbeck and Romanski 2006). Interestingly, Russ et al. (2008) also used another measure in addition to the linear discriminator, i.e., mutual information, and this measure’s performance appeared to peak at ~20-ms bin width.

It appears, therefore, that the 80–90% performance of the linear discriminator at a 2-ms bin width in prefrontal or auditory cortex (Recanzone 2008; Russ et al. 2008) resulted from an additional factor not present in the other studies. It is difficult to determine what this factor could be without knowing all technical minutiae of the studies, in particular the exact implementation of the linear discriminator algorithm. What may be significant, however, is that we have identified a step in the algorithm described by Russ et al. (2008) and Recanzone (2008) that, if slightly modified from the published form, produces results resembling theirs. The modification
is, in our opinion, of a kind that could happen as the result of a simple oversight while streamlining the code and optimizing performance.

This step was described by Recanzone (2008) as follows: “A stimulus PSTH was then constructed using all 12 trials for the other 7 stimuli and the remaining 11 trials for that particular stimulus.” We have found that if this procedure was exactly followed (“original discriminator”), the performance of the linear discriminator based on our data peaked at a bin width of 10–50 ms, with the best average performance dependent on stimulus type but remaining below 65% (Fig. 13; Kuśmierk and Rauschecker 2009). However, if the “same” stimulus PSTH was constructed with all trials for that particular stimulus (not only the remaining trials), the performance of such a “modified” discriminator peaked at >80% accuracy at the shortest bin width of 2 ms (Fig. 13). The effect was similar whether data from experiment 1 or 2 were used; thus it was little influenced by differences in stimulus sets, presentation manner, spike acquisition system, or cortical areas. We have a reason to consider results of the “modified” discriminator inaccurate: When the algorithm was applied to randomly generated time stamps, the performance reached almost 100% performance at narrow bin widths, whereas results of the “original” algorithm remained at chance (Fig. 14). MATLAB scripts showing this effect are provided as Supplemental Material.1

This analysis demonstrates how a departure from the published linear discriminator algorithm can result in altered linear discriminator performance. It is possible that the unique results shown by Russ et al. (2008) and Recanzone (2008) might have stemmed from a similar departure from the published algorithm, or from another factor of comparable consequences.

It is noteworthy also that the method utilized by Russ et al. (2008) has been criticized by other authors on another methodological basis, that is, overfitting due to insufficient trials-to-parameters ratio (Romanski and Averbeck 2009). To overcome this possible pitfall, a different cross-validation method has been suggested to minimize the overfitting problem. While this is definitely a valid criticism, our analysis suggests that yet another factor was involved: Despite the risk of overfitting, in our hands the method of Russ et al. (2008) yields reasonable best bin sizes of 20–50 ms, if the “original” algorithm is followed.

In the context of the present report, we propose that the results of Recanzone (2008) should be reinterpreted as not inconsistent with increased stimulus selectivity in area R compared with more posterior areas.

Summary. We have shown that specialization for sound-identity processing in the ventral stream can be found at its earliest stages, at the level of areas R and RM. The processing appears to develop in two stages: Within 20–60 ms after stimulus onset, stimulus clustering quality based on R

Fig. 13. Performance of the linear discriminator for ES and MC plotted against analysis bin width. In experiment 1 (regions rR and rM1 pooled), there were 10 stimuli in each class, hence the chance level was 0.1. In experiment 2 (regions rC and rM2 pooled), there were 7 stimuli in each class, for a chance level of 0.143. Left: “original” discriminator according to descriptions in Recanzone (2008), Russ et al. (2008), and Kuśmierk and Rauschecker (2009). Right: “modified” discriminator, as described in DISCUSSION.

Fig. 14. Performance of the “original” and “modified” linear discriminator for random time stamps, plotted against analysis bin size. For 100 “neurons,” time stamps were generated as 10–50 values in the 0–1,000 ms range, for 10 “stimuli” and 12 “trials” per “stimulus.” Both the number of time stamps per “trial” and the value of the time stamps were randomly and uniformly distributed.

1 Supplemental Material for this article is available online at the Journal website.
responses cannot be distinguished from responses of more posterior early areas (A1, MM, CM), as well as clustering based on acoustic measures. Later on, it surpasses both.

A methodological achievement of the study is the demonstration of how the use of population responses analyzed in short temporal windows yields substantially more information than is available with more conventional methods.

ACKNOWLEDGMENTS

We express our gratitude to Dr. Hans Engler and Dr. Mark Chevillet for data analysis suggestions, to Michael Lawson and Carrie Silver for assistance with animal training and care, and to Dr. John VanMeter for help with MRI scanning.

GRANTS

This work was supported by National Institute of Neurological Disorders and Stroke Grant R01-NS-052494 and by a PIRE grant from the National Science Foundation (OISE-0730255) to J. P. Rauschecker.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: P.K. and J.P.R. conception and design of research; P.K. performed experiments; P.K. and M.O. analyzed data; P.K., M.O., and J.P.R. interpreted results of experiments; P.K. drafted manuscript; P.K., M.O., and J.P.R. edited and revised manuscript; P.K., M.O., and J.P.R. approved final version of manuscript.

REFERENCES


