

Voice cells in the primate temporal lobe

Catherine Perrodin¹, Christoph Kayser¹, Nikos K. Logothetis^{1,2} & Christopher I. Petkov^{1,3}

1. Dept. Physiology of Cognitive Processes, Max-Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

2. Division of Imaging Science and Biomedical Engineering, University of Manchester, Manchester, M13 9PT, United Kingdom.

3. Institute of Neuroscience, Newcastle University Medical School, Henry Wellcome Building, Newcastle upon Tyne, NE2 4HH, United Kingdom.

Corresponding author: Christopher Petkov

E-mail: chris.petkov@ncl.ac.uk

The authors declare no competing financial interests.

Running title: Voice cells in the temporal lobe.

Please e-mail me (catherine.perrodin@tuebingen.mpg.de) for final copy (Perrodin, C., Kayser, C., Logothetis, N.K. & Petkov, C.I. Voice Cells in the Primate Temporal Lobe. *Current Biology* **21**, 1408-1415 (2011)).

Summary

Communication signals are important for social interactions and survival and are thought to receive specialized processing in the visual and auditory systems. Whereas the neural processing of faces by face clusters and face cells has been repeatedly studied [1-5], less is known about the neural representation of voice content. Recent functional magnetic-resonance imaging (fMRI) studies have localized voice-preferring regions in the primate temporal lobe [6, 7], but the fMRI hemodynamic response cannot directly assess neurophysiological properties. We investigated the responses of neurons in an fMRI-identified voice cluster in awake monkeys, and here we provide the first systematic evidence for voice cells. “Voice cells” were identified, in analogy to “face cells”, as neurons responding at least 2-fold stronger to conspecific voices than to “nonvoice” sounds or heterospecific voices. Importantly, whereas face clusters are thought to contain high proportions of face cells [4] responding broadly to many faces [1, 2, 4, 5, 8-10], we found that voice clusters contain moderate proportions of voice cells. Furthermore, individual voice cells exhibit high stimulus selectivity. The results reveal the neurophysiological bases for fMRI-defined voice clusters in the primate brain and highlight potential differences in how the auditory and visual systems generate selective representations of communication signals.

Highlights

- FMRI-guided electrophysiology identifies voice-sensitive cells in primates.
- Voice cells encode multiple aspects of sounds, such as voice category.
- Voice cells show a sparse coding strategy that has not been seen for face cells.
- Primate brain adopts divergent processing strategies for voices and faces.

Results

Vocalizations are acoustically complex and richly informative communication signals. Many social animals are sensitive to voice characteristics, implying that their brains can extract voice content from the other acoustic features of communication sounds (e.g., vocalization referential meaning, caller affective state, etc.). For instance, in one setting an animal might distinguish the voice of a conspecific from another class of sounds (“nonvoice”), while in another setting it might be important to distinguish different voices. As a first step toward advancing our understanding of the neuronal processing of voice content, recent fMRI studies in humans and monkeys have provided evidence for brain regions that strongly respond to voice-related content (e.g., [6, 7]). However, the fMRI signal does not allow a direct assessment of neuronal properties [11], thus, the neurophysiological underpinnings of fMRI voice-preferring clusters remained unexplored. Further, it was not clear whether neuronal strategies for generating selective representations of voices and faces might differ in, respectively, the auditory and visual systems [12].

Visual studies of face-preferring cells have reported that single neurons in the monkey inferior temporal lobe, 1) exhibit strong responses to categories of faces (relative to other categories of objects), 2) appear to cluster in large proportions, and, 3) are broadly responsive to different faces within the category of face stimuli, see [1, 2, 4, 5, 8-10]. We used fMRI-guided electrophysiology in two awake rhesus macaques (*Macaca mulatta*) to record from neurons in an fMRI voice-preferring cluster [7]. We observed evidence for single “voice cells”, that 1) exhibited strong preferential responses to a category of stimuli consisting of many conspecific voices relative to two other categories of acoustically matched natural sounds, 2) appear to cluster in moderate proportions, and 3) have highly stimulus-selective responses. These results highlight interesting potential divergences in how auditory or visual communication signals are represented in the primate brain.

To study neuronal voice-related processing, we used three carefully controlled categories of complex natural sounds for stimulation: 1) macaque calls from 12 different callers (MVocs), 2) other animal calls from 12 different callers (AVocs), and 3) 12 natural/environmental sounds (NSnds). Motivated from the study of face-preferring cells that have evaluated neuronal responses to “face” vs. “nonface” stimulus categories [1-5, 8-10], a key goal of our study was the comparison of neuronal responses to the voice (MVocs) vs. the nonvoice (NSnds) stimulus categories, including how these auditory results might compare to those for visual face cells. The

AVocs category was included to provide additional information about whether the distinction of conspecific voices (MVocs) vs. heterospecific voices (AVocs) might be important, as previous fMRI results on voice preference have suggested [7, 13]. These 36 stimuli were sampled from a larger set of sounds using the following criteria: We required that each of the vocalizations in the MVocs and AVocs categories were produced by different callers (i.e., many voices), and that the sound categories did not significantly differ in at least two key acoustical features, i.e., the overall frequency spectrum and modulations in the temporal envelope (for details see Fig. S1A-B and the Supplemental Experimental Procedures). In practice, matching the acoustics across the stimulus categories necessitated that the sampling of MVocs and AVocs stimulus sets included acoustically distinct types of commonly produced calls (as opposed to sampling only one or a few call types). Nonetheless, because we constrained our MVocs to consist of multiple call type exemplars produced by different individuals, we could separately analyze the impact of “call-type” and “voice” factors on the neuronal responses, see below.

The targeted voice-preferring fMRI cluster resides in hierarchically high-level auditory cortex on the supratemporal plane (STP), anterior to tonotopically organized auditory core and belt fields [6, 7] (Fig. 1B; S1C-D). Electrophysiological recording sites were localized using the stereotactic coordinates of the fMRI maps, targeting the anterior cluster in the right hemisphere with a strong fMRI-derived response to MVocs [7]. Access to the target region was confirmed by either using the Brain Sight© neurosurgical targeting system (Fig. 1A, S1F) or electrophysiological mapping of the posterior tonotopically organized auditory cortex (Fig. S1D). Recording locations were also confirmed at the completion of the experiments with postmortem structural MRI and histology (Fig. S1G).

Within the target region, neuronal activity was sampled from an area of $\sim 66\text{mm}^2$ in each monkey (M1 and M2) centered on the coordinates of the fMRI-identified cluster (Fig 1B; S1C). We recorded from 328 sites with auditory-responsive local-field potential (LFP) activity. We also obtained auditory-responsive spiking activity consisting of 186 multi-units (multi-unit activity: MUA, which combines multiple unit and single-unit responses from individual recording sites; 87 from M1 and 99 from M2), of which 85 were classified as well-isolated single units (SUA; see Supplemental Experimental Procedures).

Initially, we identified a population of neurons with a preferred (best) category response for MVocs (Fig. 2; S2). We observed that a significant proportion of auditory-responsive units (SUA/MUA) responded maximally to the MVocs, rather than to the other (AVocs or NSnds) sound categories (see Fig. 2A; MUA: 45% with a maximal response to the MVocs, 84/186 units; χ^2 test comparing to a uniform distribution of 33%, $p = 0.0013$; SUA: 46% with a maximal response to the MVocs, 39/85 single units, $p = 0.036$). The temporal response profiles of individual units in response to the spectro-temporally complex natural sounds showed considerable variety (see example units in Fig. 1C, S1E). However, the MVocs preference was apparent in both the population spiking response (averaged over all auditory responsive MUA, see Fig. 2B; S2A,F) and in the local-field potentials (LFP, see Fig. 2C and S2C): The preference for MVocs at the population level emerges at a latency of ~ 50 ms post-stimulus onset in the spiking response (Fig. 2B, S2E) and ~ 75 ms in the LFPs (Fig. 2C, S2E), and the MVocs preference is seen to persist throughout the stimulus presentation period (also see the cumulative response functions in Fig. S2B,D,E). These results demonstrate a significant, temporally sustained neuronal response preference for MVocs in the anterior voice cluster that is consistent with the fMRI-derived response to MVocs.

Next, we identified voice cells, in analogy to studies of face cells in the visual system, as single-units that respond at least twice stronger to voices than to other sounds [1, 4, 8, 9]. We quantified such a response preference for MVocs vs. the other categories of sounds using Voice-Selectivity Index (VSI) values $\geq 1/3$ (see Fig. 3, Experimental Procedures and [4]). This identified voice cells within the anterior fMRI voice-preferring cluster in 25% of the auditory responsive SUA (Fig. 3A) which is a considerable proportion similar to the proportion of face cells reported in the earlier visual studies [1-3, 5, 9]. This proportion of voice-preferring units was robust and did not strongly depend on the choice of a particular response window size (Fig. S3). Because a recent visual study has found very high proportions of face cells ($> 90\%$) by oversampling neighboring sites at face clusters [4], in our auditory dataset we evaluated a focal set of sites with the highest density of MVocs preferring units (Fig. 3B). In this case, the proportion of voice-preferring units increased to a still moderate 55% (11/20 responsive MUA, Fig. 3B).

Since our MVocs category consisted of several call types produced by multiple individuals, it was important to determine whether the presumed voice cells ($n = 21$ single units) were well sensitive to the different

voice-related aspects in the MVocs category, and not only the call type (although the neurons could in principle be sensitive to both voice and call type aspects, i.e., see [7, 14-16]). Figure 4 shows the response selectivity of these cells, which is also identified by voice and call-type characteristics in Fig. 4B. The results reveal highly selective responses for the MVocs stimuli, with each neuron responding to only a few of the presented vocalizations (see Fig. 4A and the relatively few black boxes seen on the x-axis in Fig. 4B). With this analysis, no clear selectivity of individual neurons associated with specific call types (e.g., grunts, etc.) is apparent (Fig. 4B). We further analyzed the responses of these neurons using a 2-way ANOVA with a call-type factor and voice as a nested factor. This revealed 6 units (29% of the voice cell SUA subsample; 7% of all SUAs) being significantly sensitive ($p < 0.01$) to the voice factor, and 6 units being significantly sensitive ($p < 0.01$) to the call-type factor. These results confirm that a considerable subset of the identified voice cells were significantly sensitive to the voice-related aspects of the MVocs stimuli and not just to the call-type aspects (Fig. 4B).

Lastly, we quantified the selectivity of the voice cells ($n = 21$) to allow comparison to what is known on the encoding properties of face cells. We quantified the response selectivity to individual stimuli within the MVocs category, which showed that the voice cells selectively responded to an average of 21% of the MVocs stimuli (i.e., 2.47 / 12 of the stimuli elicited responses greater than the half-maximum response; see Fig. 4A). This contrasts with the general impression of face cells in the temporal lobe, which are known to be much less selective for individual faces [8, 9] (Fig. 4C); see also the broad responsiveness to faces reported in [1, 2, 4, 5]. We also quantified the sparseness of the neural encoding by the identified voice cells, because face cells seem to represent faces in a fairly stimulus-nonselective (dense coding) fashion [8, 10]; sparse coding is defined as being along a continuum between local codes, where neuronal responses are extremely selective, and dense codes, where neurons respond to most stimuli [17]. To do this, we computed the sparseness index [17] of voice cells which ranges from 1 (sparse-coding strategy: strong response to a few select stimuli) to 0 (responses to most of the stimuli). The mean sparseness index for the voice cells is 0.78 (i.e., more sparse), whereas for face cells [10] the index is much lower (0.42; i.e., less sparse), see Fig. 4D. Taken together these comparisons suggest that, 1) auditory voice-preferring cells are more selective for individual voices than face-preferring cells are known to be selective for individual faces (Fig. 4C), and, 2) voice cells rely more on a sparse-coding strategy (Fig. 4D).

Discussion

We studied neurons in an fMRI voice-preferring cluster located in the right anterior STP. The results on the, 1) proportion of voice-sensitive neurons, 2) their stimulus selectivity, and 3) coding strategies, provide insights into the neurophysiological bases of the fMRI signal at voice-preferring clusters in the primate brain. Also, relative to what is known about visual face cells, our data on auditory voice cells provide new insights into organizational principles underlying the brain specialization for communication signals.

Evidence for voice cells and their auditory response characteristics. The combined results reveal that, 1) the anterior STP contains a considerable proportion of single neurons with a two-fold stronger preference for the MVocs stimulus category than to the other two acoustically controlled, natural sound categories, 2) these neurons exhibit highly selective responses and a sparse-coding strategy, and 3) a considerable fraction of the identified voice cells were significantly sensitive to the voice-related aspects of our MVocs category of stimuli, including in some cases the acoustical aspects related to call type (based on the results of the 2-factor ANOVA using call-type and voice factors).

The primary observation of the combined results is that single neurons identified as voice cells appear able to encode multiple aspects of the sound categories. This is an important step towards understanding their neuronal mechanisms and potential role in the processing of voice content as a basis for voice recognition. A voice-category representation itself can be an important signal for recognizing whether a sound was a conspecific voice (i.e., species voice) rather than some other natural sound. The high selectivity for specific stimuli is a process that, in addition, could be used to distinguish individual voices. Our previous fMRI results could not specify whether these aspects were encoded by single neurons or separately by intermingled populations of neurons [7]. The electrophysiological results obtained here reveal that a significant proportion of single neurons in the anterior fMRI voice-preferring cluster seem to both encode the MVocs category membership, and, at the same time, exhibit highly selective responses to the different voices in that category.

It is interesting that neurons in the anterior fMRI voice cluster can be sensitive to voice and call-type aspects of our MVocs stimuli. This observation resonates with human fMRI results that have noted an overlap of voice and speech processing networks in the human temporal lobe [15] and that voice regions are sensitive to speech as well as voice content [16]. Interestingly, a human selective-attention experiment has been conducted

with fMRI where the participants detected voice or speech content in the same stimulus set [14]. The authors observed that, relative to several temporal lobe regions that were either task non-specific or modulated by attention to the speech content, when attention was focused on voice content the right anterior voice-preferring region was involved. That study underscores the preferential activity for voice in the right anterior temporal lobe and its role in voice recognition [14]. This right anterior voice region in humans appears to be a functional homolog to the monkey region from which neurons were recorded here [7, 14, 18, 19].

We observed high stimulus selectivity to the MVocs sounds. We also confirmed that a subset of the neurons identified as potential voice cells were significantly sensitive to the voice-related aspects, and not only the call type aspects. Since we matched two key acoustical features across the three sound categories, to exclude these as trivial explanations for any observed category preferences, our study was based on a category of conspecific MVocs consisting of a fairly well balanced set of several commonly produced call types that were produced by different conspecific individuals (i.e., many voices). Thus, the question of how voice cells encode the voice identity of *specific* individuals cannot be directly gleaned from our data. However, we have previously obtained monkey fMRI evidence that the anterior voice cluster is preferentially responsive to both voice category and voice identity [7]. There we tested for voice-identity sensitivity using an fMRI adaptation paradigm and a stimulus set consisting of 2 exemplars of 2 call types (coos and grunts), each produced by the same 3 monkey individuals. The results revealed greater sensitivity to voice-identity (holding the call type constant but varying the callers) than to call-type (holding the caller constant but varying the call type) [7]. Notably, although the voice sensitivity was greater than the call type sensitivity, both voice and call type sensitivity were significant, which seems to relate to the sensitivity seen here for the identified voice cells. Yet, the fMRI results also reveal that the exact proportion of voice and call-type sensitive cells will depend on the stimulus set and experimental paradigm used, which are important to consider for pursuing neuronal voice-identity coding.

The inclusion of a category of heterospecific voices (AVocs) in our stimulus set was motivated by previous fMRI studies of voice-preferring regions [6, 7, 13]. It is interesting that conspecific voices (MVocs) were also preferentially represented over the heterospecific voices (AVocs), allowing us to comparably treat the AVocs and NSnds categories in our analyses. The comparison of MVocs vs. AVocs suggests, as have the fMRI reports on voice processing in monkeys [7] and humans [13], that the voices of different species are not all

equally represented. By comparison, in many visual studies on face processing in monkeys, human and monkey faces are interchangeably used for stimulation. Interestingly, recent analyses of human fMRI activity and monkey inferotemporal cortex (IT) neuronal responses to faces [20] suggest that the human brain segregates (larger dissimilarity in response patterns) human faces than does the monkey brain, but that the monkey brain appears not to significantly segregate primate faces better than the human brain. Thereby, the species being studied, their prior experience and the species of the voice or face stimuli being used will require careful comparison in studies of voice and face processing.

Certain aspects of our auditory results can generally be compared with results from auditory electrophysiological studies in animals, even though it is unclear how prior results relate to the processing of voice content since the previous work has obtained responses to vocalizations, with an interest in understanding how call type acoustics are encoded, and/or the studies have used unspecified numbers of callers [21-27]. There appears to be weak or absent neuronal preferences for vocalizations in the initial processing stages of the auditory cortex [21-23], but the selectivity for species-specific vocalizations increases in the auditory hierarchy outside of primary auditory cortex [22-24, 26]. For instance, the neuronal selectivity noted for the voice region here is higher than the selectivity for vocalizations reported at several stages of the auditory cortical processing hierarchy [23, 24], including an auditory region in the insula [28] and the superior-temporal gyrus [27]. However, auditory responsive neurons in the monkey prefrontal cortex that were stimulated with conspecific vocalizations [29] appear to be highly stimulus-selective like the voice-region neurons. These impressions are consistent with the position of the voice-preferring region in the auditory processing hierarchy, as a region in the ventral processing pathway [30] anterior to the auditory core, belt and parabelt fields (see Fig. 1B, S2C and [7, 26, 31, 32]).

Do there appear to be differences in voice and face cell processing properties? We analyzed several neurophysiological properties of the identified voice cells and compared these to the available visual studies on face cells. Below, we separately consider the comparisons of voice vs. face cell, 1) proportions [1-5, 9], 2) selectivity [8, 9], and 3) sparse coding strategies [10], before concluding whether the available data suggest more similarities or differences. Many visual studies of cells preferring face to nonface stimulus categories were included for comparison. Because the available data have sampled from various parts of the mid to anterior

temporal lobe, it is important to consider the areas from where the voice and face cells were sampled: the studied voice region is located in anatomically delineated regions Ts1/Ts2 in the STP (the 4th or 5th auditory cortical processing stage, anterior to the auditory core (1°), belt (2°) and parabelt (3°) [7, 22, 31]). The voice regions may be auditory analogs to face regions in the broadly defined visual IT cortex [18, 33-35]. The face cell data are from subregions of IT [1, 9], including the fundus [1, 5, 10], and lower [2-4, 8, 9] and upper [3, 5, 8] banks of the superior-temporal sulcus (STS).

Although proportions of cell types might reflect under- or over-sampling, these measurements have often been reported for face cells. The proportion of voice cells (i.e., those with a preference for voice [MVocs] vs. nonvoice [NSnds] or other [AVocs] sound categories), was significant (25%). However, this is relatively moderate in comparison to a recently reported very high proportion (>90%) of face-preferring cells at an fMRI-identified visual face cluster [4]. That visual study [4] questioned whether mislocalization might have resulted in the ~15-30% face-cell proportions reported in earlier studies [1-3, 5, 9]. Yet, our voice cell proportions (even when we analytically oversampled a focal cluster of MVocs preferring sites) tend to be closer to or within the range of face cell proportions in the earlier visual studies [1-3, 5, 9], which supports the notion that voice and face cell proportions are similar in the primate brain. Given the considerable variability in the numbers of face cell proportions reported in the visual literature, it is currently unclear based solely on cell proportions whether voice and face cell representations are comparable or not.

Nonetheless, if we look beyond neuronal proportions there appear to be differences in the response properties of voice- and face-preferring cells. In particular, we observed that voice cells responded only to about a fifth (21%) of the MVocs, a high level of selectivity consistent with the results of another study in the anterior, hierarchically higher-level regions of auditory cortex [26]. In contrast, face cells seem to respond to ~39-62% of face stimuli [8, 9], suggesting that they are less selective for specific faces (Fig. 4C). Also, voice cells exhibited a sparse coding strategy for voices (Fig. 4D). Again, by contrast, face cells are known to adopt a more distributed representation of faces (see Fig. 4D; also see [10] and the Supplemental Experimental Procedures).

In summary, the available data on voice- and face-cell selectivity and coding strategies suggest that there are more differences than similarities in the processing characteristics of voice and face cells. It is possible that these differences may become less apparent, for instance, once it is better known how dynamic facial

expressions are processed by neurons, since face cells have been over-abundantly studied with static faces whereas natural sounds are dynamic spectro-temporally varying sounds. As might be relevant for future comparisons of neuronal response dynamics, our temporally resolved analyses of the population MVocs preference show that this preference is present with a short latency after stimulus presentation and persists for the duration of the stimulus period (but see the variety in single neuron response dynamics).

If voice/face cell differences persist this would be interesting from an evolutionary perspective since it has been often suggested (e.g., [36]) that the auditory system could have specialized in different ways than the visual, or at least the organizational properties of auditory neurons have been difficult to delineate [12, 37] in relation to those for visual neurons [38]. Canonical facial features, for example two eyes, a nose and a mouth, have been broadly conserved in vertebrates, whereas vocal production varies considerably [36]. In particular, animals gain adaptive advantages by producing vocalizations that are acoustically distinct from those of other species, that acoustically circumvent environmental noise, and that contain different levels of voice information (voiced/unvoiced calls); not to mention environmental influences on sound acoustics. One could hypothesize that the observed sparse code for auditory voice cells is efficient for encoding elements from a more variable category of dynamic sounds, while the less sparse coding of face cells is efficient for discerning subtle differences between facial features [10] within a relatively more stable category of visual objects such as faces.

Conclusions. Our results identify voice cells using analogous analyses as were used to reveal face cells—which have been the subject of numerous studies. This investigation of the neurophysiological properties of voice cells reveals important initial impressions on the functional characteristics of these auditory cells and clarifies the neurophysiological bases of the fMRI voice-related activity response. This study builds on the links that are being established between how the brains of humans and other animals process communication signals such as voices and faces, and the results extend an animal model system for understanding the processing of vocal communication at the neuronal level. We also note a more stimulus-selective (e.g., sparse) representation by the identified voice cells in the auditory system than that reported for face cells in the visual regions of the ventral temporal lobe. At this juncture, our results indicate that neuronal specialization for voice and face information appears to rely on different processing strategies. Cross-sensory comparisons such as ours can now be extended to address how neurons in the other sensory systems of various animal species might selectively

encode communication signals. Our combined results highlight the selectivity and processing strategies of neurons in the primate brain for representing auditory aspects of communication signals.

Experimental Procedures

Full methodological details are provided in the Supplemental Experimental Procedures and are summarized here. Two adult male rhesus macaques (*Macaca mulatta*) participated in these experiments. The macaques were part of a group-housed colony. All procedures were approved by the local authorities (Regierungspräsidium Tübingen, Germany) and were in full compliance with the guidelines of the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals.

Stimuli. To balance the acoustical features of the experimental sound categories while maintaining their ethological relevance, three categories of 12 complex natural sounds were sub-sampled from a larger set of vocalizations and natural/environmental sounds that we have previously used, see Experiment 1 in [7]. The categories of sounds consisted of the following: (1) macaque vocalizations from 12 different callers (MVocs); (2) other animal vocalizations from 12 different callers (AVocs); and, (3) 12 natural/environmental sounds (NSnds). See Supplemental Experimental Procedures, section *Acoustical stimuli* for further details. Each vocalization in the MVocs and AVocs categories was produced by a different individual, thereby consisting of a category of many voices (as in the categories of many faces used to study face processing [1-5, 8, 39-42]). Moreover, these categories were composed of a mixture of commonly produced call types, to balance the impact of any particular form of referential information in the vocalizations [7]. The intensity of all of the sounds was normalized in RMS level and was calibrated at the position of the head to be presented at an average intensity of 65 dB SPL within a sound-attenuating chamber (Illtec).

Functional MRI. The two macaques had previously participated in fMRI experiments to localize their voice-preferring regions, including the anterior voice clusters, see [7] and Supplemental Experimental Procedures. Briefly, monkey 1 (M1) was scanned awake in a 7-Tesla MRI scanner (Bruker Medical), and monkey 2 (M2) was scanned anesthetized in a 4.7T scanner. To better compare with the electrophysiological data analyses (see *preferred category* analyses below and in Fig. 1B, 2A, S1C, S2C) the fMRI activity cluster that prefers MVocs was analyzed using the MVocs > max [activity response of other sound categories] criterion. The stereotactic coordinates of the voice cluster centers were used to guide the electrophysiological recordings.

Electrophysiological recordings. Standard extracellular electrophysiological recordings were performed using epoxy-coated tungsten microelectrodes (FHC Inc.). During recordings, the animals were awake and

passively listening to the sounds in a darkened and sound-attenuating booth (Illtec). The electrophysiological recording chamber was positioned using the preoperatively obtained stereotaxic coordinates of the individual fMRI maps of the animals, allowing access to the auditory regions on the STP (see Fig. 1A,B and S1C, F). The precise angle of the recording electrodes and depth to reach the center of the fMRI cluster were obtained by using the BrainSight neurosurgical targeting system which combines MRI- and fMRI-based markers (Rogue Research, Inc.) or tonotopic mapping of neighboring auditory cortical fields, see Supplemental Experimental Procedures. The experimental sounds were presented individually in randomized order, using a rapid stimulus presentation procedure (similar to [4]) with a randomly varying inter-stimulus interval ranging from 100 to 175ms. We obtained responses to at least 20 repetitions of each stimulus.

The data were analyzed in Matlab (Mathworks). The recorded broadband signal was separated into spiking activity by high-pass filtering the raw signal at 500 Hz. Spiking activity was subsequently sorted offline (Plexon). The low-frequency signal from 4-150Hz yielded the local-field potential. The filter was set at 4Hz for us to optimize the recording of the higher frequency spiking activity, which can be affected by large slow-wave oscillations. Because of this we do not know if we might have overlooked a potential contribution to the LFP response preference for MVocs in very slow oscillations below 4 Hz.

A significant response to sensory stimulation (auditory-responsive activity) was determined by comparing the response amplitude of the average response to the response variability during the baseline period. Arithmetically this involved normalizing the average response to standard-deviation units (SD) with respect to baseline (i.e., z-scores), and a response was regarded as significant if the z-score exceeded 2.5 SDs during a continuous period of at least 25ms (50ms for LFP responses) during stimulus presentation. A unit or recording site was considered auditory responsive if its response breached this threshold for any of the 36 experimental auditory stimuli.

For each response type, the mean of the baseline response was subtracted to compensate for fluctuations in spontaneous activity. Response amplitudes were defined by first computing the mean response for each category (MVocs, AVocs, NSnds) across trials and different sounds. For the category response, the peak of the category average response was calculated and the response amplitude was defined as the average response in a 200ms window centered on the peak of the category average response. The preferred category for each unit was

defined as the one eliciting the largest (maximal) response amplitude. Voice-preferring cells were classified according to the face-preferring criterion used in visual studies [1, 4, 8, 9]. In our case, the response to MVocs is defined as being at least twice larger than the response to the other categories. Formally this was based on the approach used in [4], as follows. We defined a voice selectivity index as $VSI = \frac{\text{mean}(MVocs) - \text{mean}(others)}{\text{mean}(MVocs) + \text{mean}(others)}$ using the average response amplitudes to the different sound categories. A single unit was defined as a voice cell if its VSI was larger than or equal to 1/3, also see [4]. Finally, we computed a standard sparseness index [17] of the form $s = \frac{1-a}{1-\frac{1}{n}}$ where $a = \left(\frac{\left(\sum_{i=1}^n \frac{r_i}{n} \right)^2}{\sum_{i=1}^n \frac{r_i^2}{n}} \right)$, r_i is the trial-averaged, baseline-corrected response amplitude to the i th stimulus of the MVocs sound category and n is the total number of stimuli in that category (here, $n = 12$ voices in the MVocs category). The index s is a scaled version of the index a , which was used to estimate sparseness for visual face cells [10]. To directly compare to these results we converted a to s (see Fig. 4D and the Supplemental Experimental Procedures).

All auditory responsive units distributed across the sampled area (approx. 66mm²) on the anterior STP of both monkeys were used in the analyses. Since this was a broad recording region and to allow better comparison to a recent visual study on face cells where the authors recorded from the center of a face cluster [4], we defined in each monkey a focal cluster of the same dimensions as in the visual study, i.e., three adjacent grid holes (spacing of 0.75 mm) that contained the highest density of MVocs-preferring units (see Fig. 3B).

Acknowledgements

We thank M. Augath for technical assistance with MRI scanning, H. Evrard for assistance with histology, C. Stamm for veterinary care for the animals, and M. Munk for encouragement and support. D. Blaurock, M. Munk, C. Poirier and K. Whittingstall provided useful comments on prior versions of the manuscript. Financial support provided by the Max-Planck Society (NKL, CK, CIP) and the Wellcome Trust (CIP).

References

1. Perrett, D.I., Rolls, E.T., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res* 47, 329-342.
2. Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* 4, 2051-2062.
3. Baylis, G.C., Rolls, E.T., and Leonard, C.M. (1987). Functional subdivisions of the temporal lobe neocortex. *J Neurosci* 7, 330-342.
4. Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 670-674.
5. Bruce, C., Desimone, R., and Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol* 46, 369-384.
6. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309-312.
7. Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N.K. (2008). A voice region in the monkey brain. *Nat Neurosci* 11, 367-374.
8. Baylis, G.C., Rolls, E.T., and Leonard, C.M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res* 342, 91-102.
9. Hasselmo, M.E., Rolls, E.T., and Baylis, G.C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav Brain Res* 32, 203-218.
10. Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol* 73, 713-726.
11. Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature* 453, 869-878.
12. King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? *Nat Neurosci* 12, 698-701.

13. Fecteau, S., Armony, J.L., Joanette, Y., and Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* 23, 840-848.
14. von Kriegstein, K., Eger, E., Kleinschmidt, A., and Giraud, A.L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res* 17, 48-55.
15. Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322, 970-973.
16. Belin, P., Zatorre, R.J., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 13, 17-26.
17. Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273-1276.
18. Petkov, C.I., Logothetis, N.K., and Obleser, J. (2009). Where are the human speech and voice regions, and do other animals have anything like them? *Neuroscientist* 15, 419-429.
19. Belin, P., and Zatorre, R.J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14, 2105-2109.
20. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126-1141.
21. Wollberg, Z., and Newman, J.D. (1972). Auditory cortex of squirrel monkey: response patterns of single cells to species-specific vocalizations. *Science* 175, 212-214.
22. Rauschecker, J.P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111-114.
23. Recanzone, G.H. (2008). Representation of con-specific vocalizations in the core and belt areas of the auditory cortex in the alert macaque monkey. *J Neurosci* 28, 13184-13193.
24. Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J.P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science* 292, 290-293.
25. Wang, X., and Kadia, S.C. (2001). Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J Neurophysiol* 86, 2616-2620.

26. Kikuchi, Y., Horwitz, B., and Mishkin, M. Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J Neurosci* 30, 13021-13030.
27. Russ, B.E., Ackelson, A.L., Baker, A.E., and Cohen, Y.E. (2008). Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *J Neurophysiol* 99, 87-95.
28. Remedios, R., Logothetis, N.K., and Kayser, C. (2009). An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. *J Neurosci* 29, 1034-1045.
29. Romanski, L.M., Averbeck, B.B., and Diltz, M. (2005). Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *J Neurophysiol* 93, 734-747.
30. Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc Natl Acad Sci U S A* 97, 11800-11806.
31. Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc Natl Acad Sci U S A* 97, 11793-11799.
32. Rauschecker, J.P., and Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *J Neurophysiol* 91, 2578-2589.
33. Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci U S A* 104, 20600-20605.
34. Campanella, S., and Belin, P. (2007). Integrating face and voice in person perception. *Trends Cogn Sci* 11, 535-543.
35. Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 8, 129-135.
36. Miller, C.T., and Cohen, Y.E. (2010). Vocalizations as Auditory Objects: Behavior and Neurophysiology. In *Primate Neuroethology*, A.A. Ghazanfar and M.L. Platt, eds. (Oxford University Press), pp. 237-255.
37. Griffiths, T.D., Warren, J.D., Scott, S.K., Nelken, I., and King, A.J. (2004). Cortical processing of complex sound: a way forward? *Trends Neurosci* 27, 181-185.
38. Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195, 215-243.

39. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* *17*, 4302-4311.
40. Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* *272*, 1665-1668.
41. Sergent, J., Ohta, S., and MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain* *115 Pt 1*, 15-36.
42. Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B. (2003). Faces and objects in macaque cerebral cortex. *Nat Neurosci* *6*, 989-995.
43. Petkov, C.I., Kayser, C., Augath, M., and Logothetis, N.K. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biol* *4*, e215.

Figures

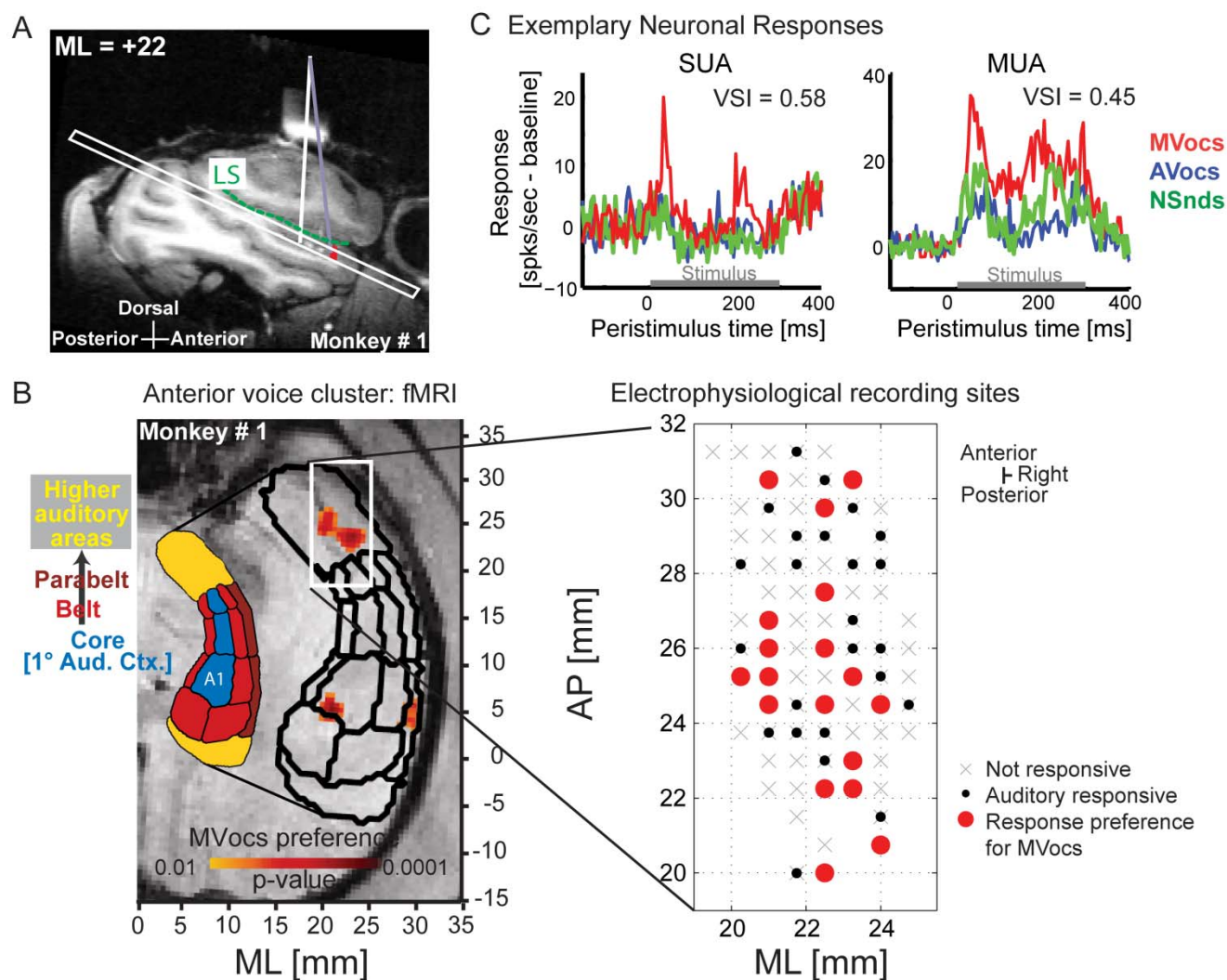


Figure 1: Targeting the anterior monkey fMRI voice cluster for electrophysiological recordings.

(A) Sagittal structural MRI of the liquid-filled recording chamber (white bar above brain, with vertical white line projecting to the supra-temporal plane (STP) below the lateral sulcus, LS). Brain Sight© and stereotactic coordinates guided electrode placement to the anterior fMRI voxels (red) with a strong preference for MVocs.

(B) Axial slice from (A), including the separately localized auditory fields (black outlines), see: [7, 43]. Antero-posterior (AP) and medio-lateral (ML) coordinates are shown for the fMRI (left) and the electrophysiological recording sites (right). The stereotactic coordinates used the Frankfurt-zero standard, where the origin is defined as the midpoint of the interaural line and the infraorbital plane.

(C) Exemplary ‘voice cell’ (SUA) and multi-unit activity (MUA) exhibiting preferential responses to MVocs, including voice-selectivity index (VSI) values (see Experimental Procedures and Fig. 3).

See also Fig. S1E in the Supplemental Information.

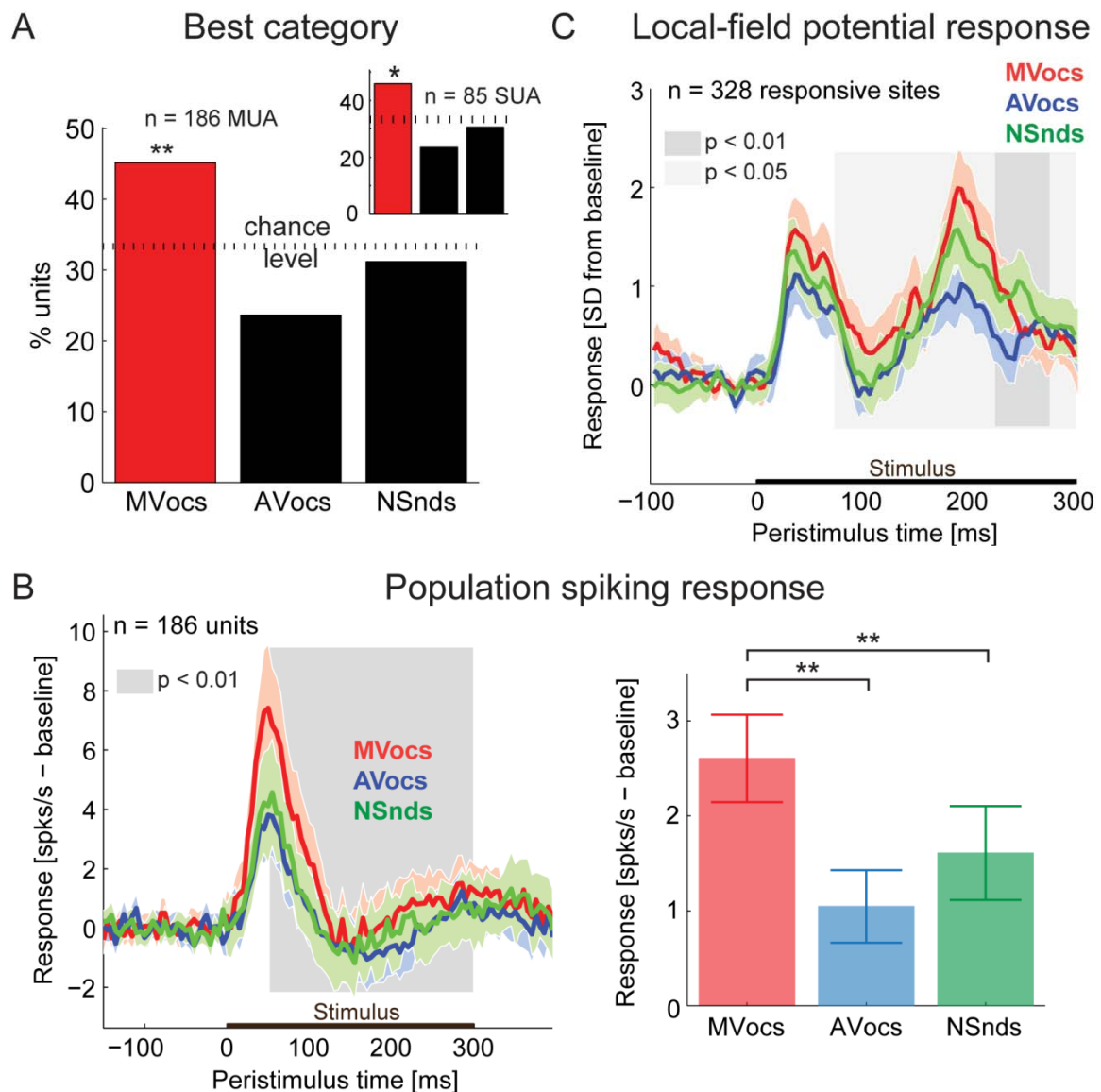


Figure 2: Neuronal preferred sound categories in the anterior fMRI voice cluster.

(A) Proportion of auditory responsive units (MUA) as a function of the sound category eliciting the maximal response (SUA results in inset), χ^2 -test; *: p<0.05; **: p<0.01.

(B) Average population spiking response to the three sound categories. The color shading indicates the 95% confidence interval for each response. The grey shaded area indicates the time interval in the cumulative spiking response (see Fig. S2B in the Supplemental Information) during which the population preference for MVocs vs. mean[AVocs, NSnds] is significant (paired-sample t -test, see Fig. S2E). The bar plot (right) shows the mean \pm SEM average response amplitudes for each category (paired-sample t -test; *: $p < 0.05$; **: $p < 0.01$).

(C) Average local-field potential responses (SD from baseline) over all auditory responsive sites (mean and 95% confidence interval). Shaded area indicates the time points with significant MVocs preference (paired-sample t -test, MVocs vs. mean[AVocs, NSnds]. See Fig. S2E) in the cumulative LFP response (see Fig. S2D in the Supplemental Information).

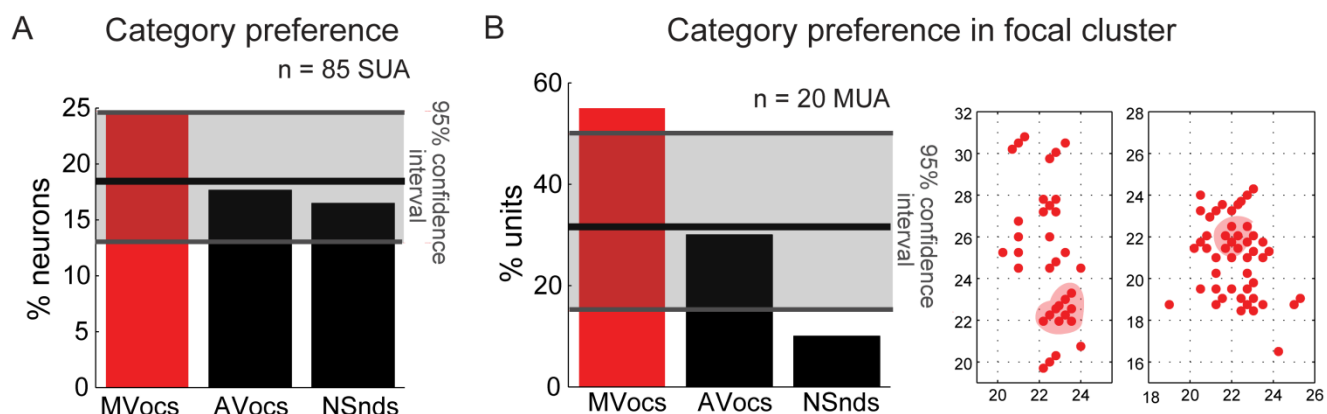


Figure 3: Neuronal sound-category preferences using the Voice-Selectivity Index.

(A) Significant proportions of ‘voice cells’ were observed (defined as SUA with a strong preference for MVocs, see red bar, i.e., a voice-selectivity index (VSI) value greater than or equal to 1/3). Cells with a preference for another sound category (shown in black bars) did not exceed chance levels. The horizontal black line indicates the chance level and the shaded grey area indicates the two-tailed, 95% confidence interval, estimated using a bootstrap procedure consisting of shuffled category labels for every unit ($n = 1000$ iterations).

(B) Restricting the analysis to a focal cluster of sites (3 neighboring grid holes with the largest density of MVocs-preferring units, see right panel) resulted in 55% of the MUA meeting the VSI-based criterion as in (A). See also Fig. S3 in the Supplemental Information.

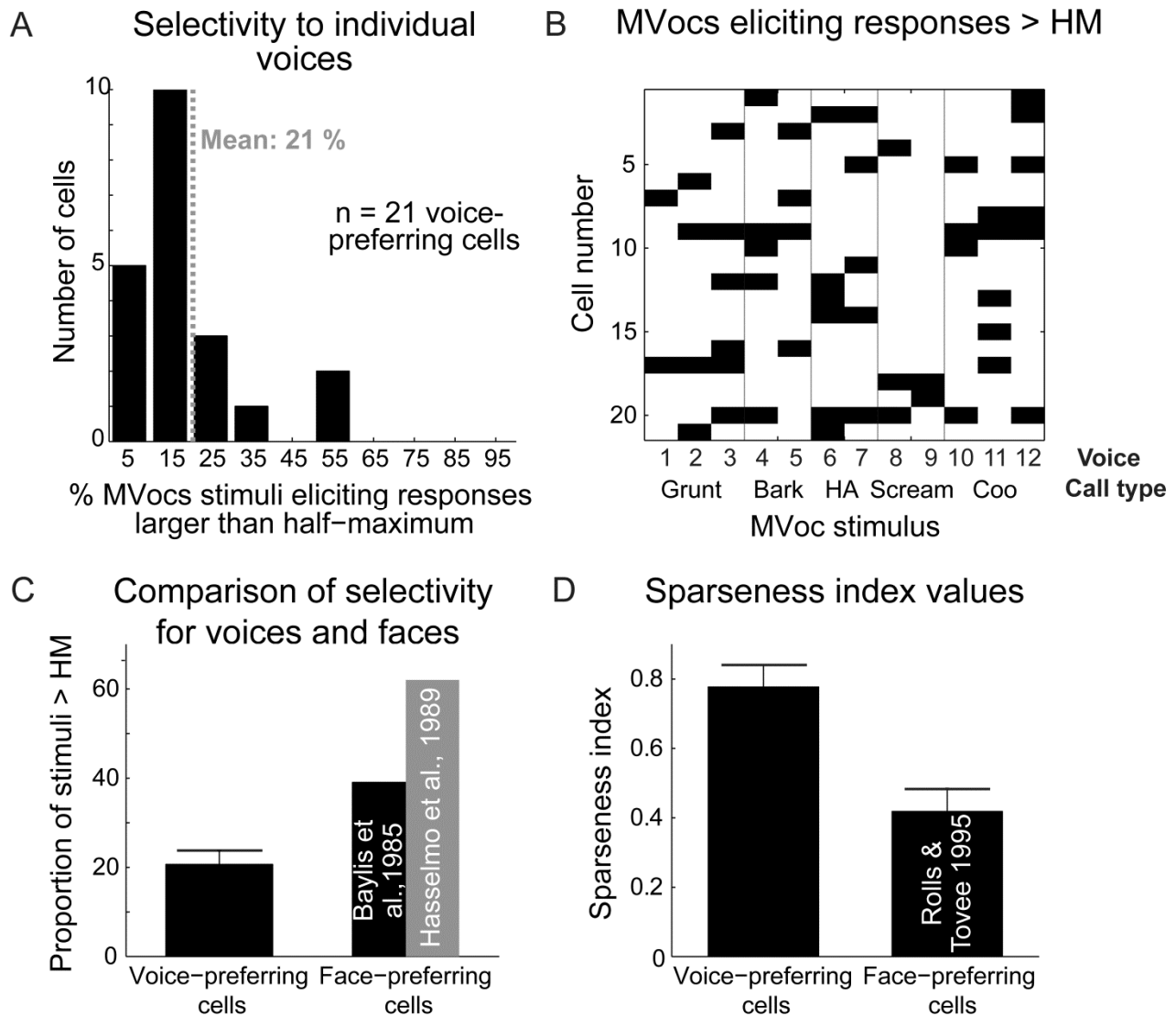


Figure 4: Selectivity for MVocs stimuli and ‘voice vs face cell comparisons.

(A) Distribution of response selectivity values across MVocs-preferring cells (SUA). Selectivity was computed as the percentage of the 12 stimuli from the MVocs sound category eliciting responses larger than half of the maximum ($> \text{HM}$) response for each cell.

(B) Distribution of effective calls for the population of identified voice cells ($n = 21$). A black square indicates that the particular MVoc stimulus elicited a response larger than half of the maximum response for a particular voice cell. See text for comparisons of voice vs. call-type responses.

(C) Comparison of the average selectivity values for voice-preferring cells with values reported for face-preferring cells in the visual system [8, 9]. Shown is mean \pm SEM.

(D) Comparison of sparseness index values for voice cells to values reported for face cells [10] (see Supplementary Experimental Procedures for details). Shown is mean \pm SEM.