J Neurophysiol 114: 1819–1826, 2015.
First published August 5, 2015; doi:10.1152/jn.00407.2014.

Multivariate sensitivity to voice during auditory categorization

Yune Sang Lee,[1] Jonathan E. Peelle,[2] David Kraemer,[2,3] Samuel Lloyd,[1] and Richard Granger[1]
[1]*Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire;* [2]*Department of Otolaryngology, Washington University in St. Louis, St. Louis, Missouri; and* [3]*Department of Education, Dartmouth College, Hanover, New Hampshire*

Submitted 30 May 2014; accepted in final form 31 July 2015

Lee YS, Peelle JE, Kraemer D, Lloyd S, Granger R. Multivariate sensitivity to voice during auditory categorization. *J Neurophysiol* 114: 1819–1826, 2015. First published August 5, 2015; doi:10.1152/jn.00407.2014.—Past neuroimaging studies have documented discrete regions of human temporal cortex that are more strongly activated by conspecific voice sounds than by nonvoice sounds. However, the mechanisms underlying this voice sensitivity remain unclear. In the present functional MRI study, we took a novel approach to examining voice sensitivity, in which we applied a signal detection paradigm to the assessment of multivariate pattern classification among several living and nonliving categories of auditory stimuli. Within this framework, voice sensitivity can be interpreted as a distinct neural representation of brain activity that correctly distinguishes human vocalizations from other auditory object categories. Across a series of auditory categorization tests, we found that bilateral superior and middle temporal cortex consistently exhibited robust sensitivity to human vocal sounds. Although the strongest categorization was in distinguishing human voice from other categories, subsets of these regions were also able to distinguish reliably between nonhuman categories, suggesting a general role in auditory object categorization. Our findings complement the current evidence of cortical sensitivity to human vocal sounds by revealing that the greatest sensitivity during categorization tasks is devoted to distinguishing voice from nonvoice categories within human temporal cortex.

auditory; categorization; human; voice; living; animate; category specific; temporal voice area; conspecific; multivariate pattern-based analysis

CONSPECIFIC VOCALIZATION IS an auditory signal that pervades the lives of vocal species. The everyday task of successfully distinguishing conspecific vocal signals from myriad other environmental sounds can be a matter of life or death and is crucial for social communication. It is therefore reasonable to hypothesize that the auditory systems of vocal species may be more efficiently tuned to the acoustic properties of conspecific sounds, allowing for rapid categorization. Comparative studies have found support for this hypothesis, providing evidence of larger neural responses to conspecific vocalizations than to other sounds in animals (Andics et al. 2014; Faragó et al. 2014; Perrodin et al. 2011; Petkov et al. 2009; Taglialatela et al. 2009). For example, an electrophysiological study in nonhuman primates identified neural populations that exhibited greater firing rates in response to conspecific monkey calls than to other sounds within the anterior portion of right superior temporal plane (Petkov et al. 2008). Likewise, in humans, a number of functional neuroimaging studies have highlighted cortical foci within bilateral superior temporal sulcus (STS) and gyrus (STG) that show increased activation in response to human vocal sounds (Belin and Zatorre 2003; Belin et al. 2000, 2002; Fecteau et al. 2004; Kriegstein and Giraud 2004; Latinus and Belin 2011b; von Kriegstein et al. 2003). These temporal foci are sometimes referred to as temporal voice areas (TVAs), and concordantly, we use the term "voice" to refer exclusively to human vocal sounds.

Although the mid-to-anterior portion of STS/STG has been proposed to be a functional module for voice, the exact operational roles of these voice-sensitive regions remain to be determined. One possibility is that the TVAs play a role in discriminating human from nonhuman vocal sounds by generating characteristically distinct neural representations devoted to voice during auditory object categorization. The present functional MRI (fMRI) study aimed to test this hypothesis using multivariate pattern-based analysis (MVPA) (Ethofer et al. 2009; Formisano et al. 2008; Giordano et al. 2013; Haxby et al. 2001; Leech and Saygin 2011; Staeren et al. 2009) and a signal detection theory. With this analysis approach, we can examine how distinct the neural representations pertaining to various auditory object categories are by means of a classifier algorithm's accuracy or sensitivity measures, such as d-prime or positive predictive value (PPV). In the present study, we relied on both accuracy and PPV throughout the series of classification tests. We note that this type of classification sensitivity is different from a strong response in the univariate sense: whereas a strong univariate response to a particular acoustic feature may occur in the absence of distinguishing the feature from others, high sensitivity in the signal detection sense must occur only when the neural representation of an object is reliably distinguished from that of every other object.

Based on this premise, greater sensitivity for voice than for nonvoice in a particular brain area indicates that the region produced a distinct categorical representation for human vocal sounds. We expect to observe a strong sensitivity for human vocal sounds as a mark of the perceptual sophistication granted to conspecific vocalizations in STS/STG bilaterally.

MATERIALS AND METHODS

Subjects. Thirteen right-handed volunteers (aged 22–34 yr; mean age 26.6; five women) participated in this study. One male subject's data were discarded due to poor performance (69%; <3 SD) in the auditory memory task during scanning, leaving a total of 12 subjects. None of the subjects had hearing difficulties or neurological disorders based on self-report. Written consent was obtained from all subjects, as approved by the Human Subjects Institutional Review Board of Dartmouth College.

Stimuli. Stimuli consisted of four types of vocalizations (humans, birds, dogs, and horses) and four types of mechanical noises (cars,

Address for correspondence and present address: Y. S. Lee, Dept. of Neurology, School of Medicine, Center for Cognitive Neuroscience, Univ. of Pennsylvania, Philadelphia, PA 19104 (e-mail: yslee@mail.med.upenn.edu).

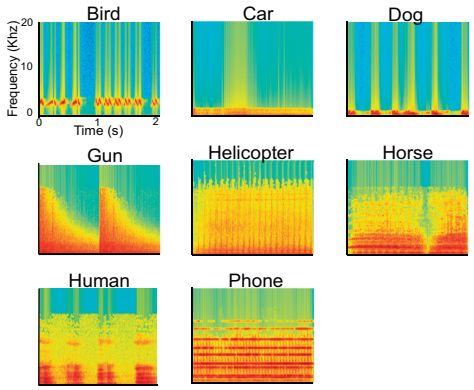

Fig. 1. Spectrograms of 8 auditory object categories. A spectrogram is shown from 1 of 3 exemplars per each category. Stimuli intensity is depicted by color. The *x*-axis indicates time in seconds, and the *y*-axis indicates frequency in kilohertz.

phones, guns, and helicopters). Example spectrograms are shown in Fig. 1. These sounds were recorded in 16-bit stereo at a sampling rate of 44.1 kHz. Each category contained three different exemplars, resulting in a total of 24 sounds. The human stimuli were made by concatenating semantically empty sounds, such as "um" and "eh," produced by three adult men unknown to the subjects, over a 2-s period (e.g., um-um-um). All other stimuli were obtained from a commercial sound-effect library (The Hollywood Edge, Richmond Hill, Ontario, Canada). There were no significant differences across categories in the duration of stimuli (mean = 1.94 s) or their root mean-squared power. We applied 20 ms linear ramps to the onset and offset of each stimulus using sound-editing software (Sound Forge 9.0; Sony, Tokyo, Japan) to avoid acoustic transients. Stimuli were delivered binaurally using a high-fidelity, MRI-compatible headphone (OPTIME 1; MR CONFON, Magdeburg, Germany) in the scanner. During the testing run (before the first run), the sound volume was adjusted to a comfortable level for each subject but was also ensured to be loud enough relative to the scanner noise. To attenuate the further influence of the background noise and increase the signal-to-noise ratio, subjects wore earplugs. All subjects reported that they were able to recognize all of the stimuli clearly.

Acoustic analysis. We performed acoustic analyses using Praat software (version 5.4.08; http://www.praat.org) (Boersma 2001). Several spectral and temporal aspects of acoustic features were measured, including intensity, harmonic-to-noise ratio (HNR), fundamental frequency (f0), center of gravity, and the variability (i.e., SD) of spectrum and intensity (Fig. 2). All analyses were done for the postprocessed event window that did not include silent periods in between or at the end of the stimuli (e.g., the interval between um sounds in the human speech). Although each analysis was based on standard parameters provided by Praat, we also manually set a specific parameter when the algorithm failed completely (e.g., unspecified value) or yielded unreasonable values. For example, Praat sometimes yielded pitch-tracking errors under the standard setting, which affected calculation of the f0. By trial and error, we chose minimum and maximum pitch range until these issues were resolved.

MRI scanning. MRI scanning was conducted on an Achieva 3.0T whole-body scanner (Philips Medical Systems, Best, The Netherlands), with an eight-channel head coil at the Dartmouth College Brain Imaging Center.

Parameters of the continuous echo-planar imaging (EPI) are as follows: repetition time (TR) = 2,000 ms; echo time (TE) = 35 ms; field of view = 240 × 240 mm; 30 axial slices; voxel size = 3 × 3 × 3 mm; interslice interval = 0.5 mm. Each subject completed six functional EPI runs (240 trials/run), and the order of runs was pseudorandomized across subjects. A high-resolution magnetization-prepared rapid acquisition gradient echo structural scan (TR = 9.8 ms; TE = 4.6 ms; 160 sagittal slices; voxel size = 1 × 1 × 1 mm) was acquired at the end of the scan.

Experimental procedures. Stimuli were presented in a slow event-related design (8 s of fixed interstimulus interval) over a total of six

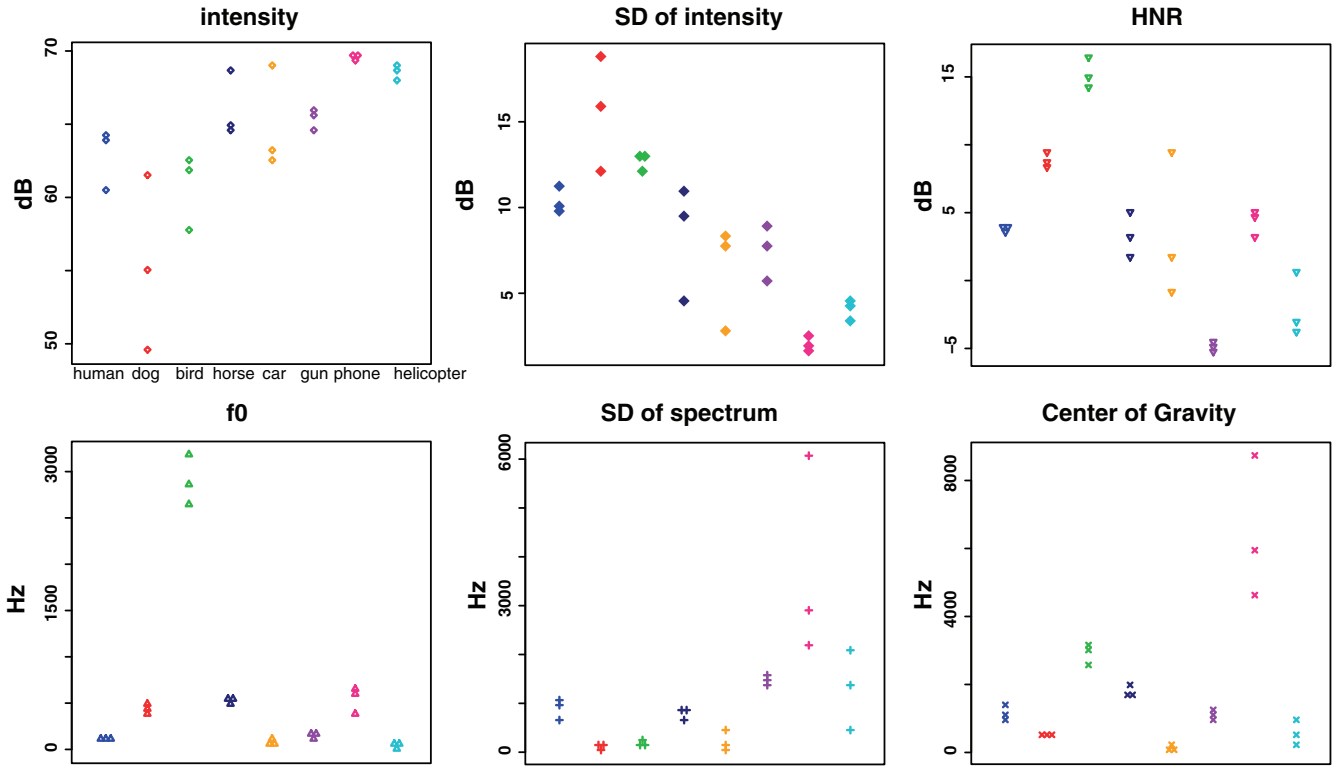

Fig. 2. Acoustic profiles for each of the auditory stimuli. The *x*-axis indicates different categories, and the *y*-axis indicates either decibel (*top*) or frequency (*bottom;* in hertz). HNR, harmonic-to-noise ratio; f0, fundamental frequency.

Table 1. *The P values of post hoc pairwise comparisons between human and nonhuman categories across all acoustic analyses (see Fig. 2)*

Category	Intensity	SD of Intensity	HNR	f0	SD of Spectrum Center of Gravity	
Bird	0.97	0.86	**0.01**	**0.01**	0.96	0.17
Car	0.98	0.3	0.99	0.99	0.97	0.74
Dog	0.07	0.1	0.16	**0.03**	0.93	0.97
Gun	0.95	0.67	**0.01**	0.99	0.98	1
Helicopter	0.3	**0.03**	0.09	0.99	0.99	0.98
Horse	0.85	0.91	0.99	**0.01**	1	0.97
Phone	0.14	**0.01**	0.99	**0.01**	**0.01**	**0.01**

HNR, harmonic-to-noise ratio; f0, fundamental frequency. $P < 0.01$ (e.g., $P = 0.0001$) was equally marked as $P = 0.01$. Boldface indicates statistical significance.

runs. During each run, subjects performed six sessions of an auditory memory task. In each session of the task, subjects first heard eight auditory stimuli, randomly selected from the 24 different exemplar sounds, while maintaining central visual fixation. Concurrent with the onset of the ninth auditory stimulus (i.e., a target sound), the visual fixation cross was changed to the task question: "Was this sound previously presented during the task session?" One-half of the time, the last stimulus was acoustically identical to one of the eight sounds that had already been presented within that run (i.e., dog2 for dog2), and one-half of the time, it was a different sound that was outside of the eight categories (e.g., camera, duck, etc.). Subjects answered the presented question via a button press. The next iteration of the task began after an 8-s resting period. The mean accuracy for the auditory memory task during scanning was 92.0% (SE = 1.4%), indicating that all 12 subjects were attending to each sound closely.

After the scanning, we played back every stimulus to the subjects and asked them to type in the name of the sound that immediately occurred to their mind to ensure that all stimuli were recognizable at the basic categorization level. The mean accuracy on this task was 94.1% (SE = 2.8%).

fMRI data analysis. All fMRI data were preprocessed using the SPM8 software package (Wellcome Trust Centre for Neuroimaging, London, UK) and MATLAB 2009a (MathWorks, Natick, MA). All images for each condition were realigned to the first EPI for correction of movement artifacts and spatially normalized into Montreal Neurological Institute (MNI) standard stereotactic space by directly registering all of the functional data to the MNI template EPI with affine registration and preserved original voxel size ($3 \times 3 \times 3$ mm).

We then extracted the time courses of all voxels that were temporally high-pass filtered with a 128-s cutoff to remove slow signal drifts and mean centered across entire runs. These preprocessed fMRI intensities were used as input vectors for classification. Intensities were obtained by relying on the general linear model framework as follows: first, we created a regressor of each category per each run by convolving the onset of each stimulus with the canonical hemodynamic response function (HRF). Then, the mean value across all time points of the regressor was calculated. A time point was assigned as belonging to the given stimulus class if the value of the regressor at that time point was greater than this overall mean value. This typically rendered time points corresponding to the range around the peak of the HRF (poststimulus TRs, 4, 6, and 8 s). The extracted data were then submitted to the Gaussian Naïve Bayes linear classifier using the MATLAB statistics toolbox (Raizada and Lee 2013). With the use of a whole-brain searchlight analysis (Kriegeskorte et al. 2006), a series of classification tests was performed at every local searchlight sphere (consisting of a center voxel and its neighboring voxels within a three-voxel radius; up to 123 voxels/each sphere; on average, 111 voxels/sphere). These tests included binary classifications (e.g., human vs. every other nonhuman category; living vs. nonliving categories) and eight-way classification. For the binary and eight-way classification tests, accuracy (hit + correct rejection/2) and PPV (hit/hit + false-alarm rate) were computed, respectively. Across all classification tests, a leave-one-run-out scheme was used, whereby data from five scanning runs served as a training set and a remaining one run as a testing set, resulting in sixfold crossvalidation. For the purpose of assessing accuracy in the binary tests, the individual searchlight map was submitted to the second-level analysis after the chance-level accuracy (0.5) was subtracted from the value in each voxel. Likewise, PPV was evaluated in a pairwise manner between human and each of nonhuman auditory objects at the group level by submitting the difference (e.g., $PPV_{human} - PPV_{dog}$; $PPV_{human} - PPV_{phone}$; etc.) to the second-level analysis. All group-level inferences were drawn using an uncorrected threshold at $P < 0.001$ (voxel-wise) in combination with cluster-size correction at $P < 0.05$, based on random field theory (Worsley et al. 1996). For visualizing group results, the *t*-maps generated from the group analysis were projected onto statistical parametric mapping surface renderings, Connectome Workbench's inflated surface map (Van Essen et al. 2012), or MRIcron (Rorden and Brett 2000).

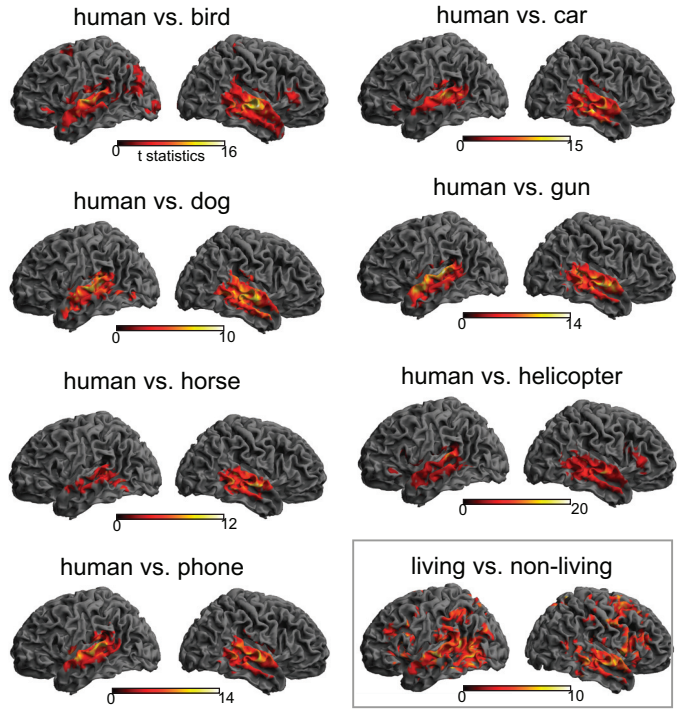

Fig. 3. Cortical rendering views of binary classification results. Seven binary classification tests between human vocalizations and each of the 7 other nonhuman categories consistently revealed significant voxels bilaterally in the superior temporal lobes. An additional classification test between living and nonliving categories (*bottom right*) yielded significant voxels throughout the brain, including temporal, parietal, and frontal cortex. The color bar indicates the *t*-statistics for a group-level result, in which accuracy was compared with a chance-level performance (50%) for each binary classification test.

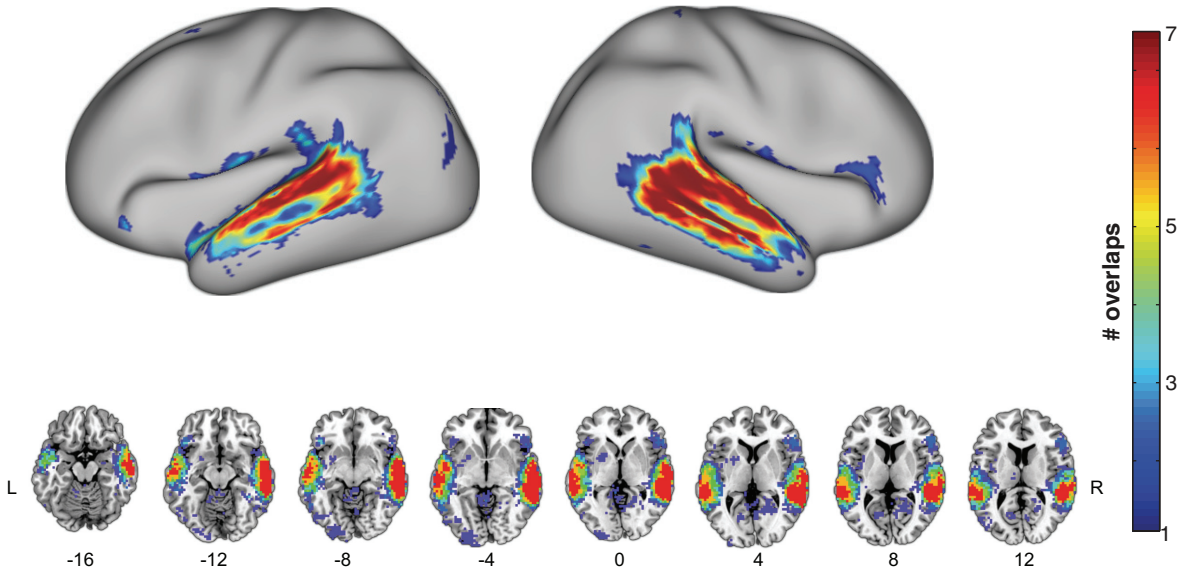

Fig. 4. Overlap of 7 binary classification maps between human and each of nonhuman categories. *Top*: surface-rendering view; *bottom*: multislice view. The color bar indicates the number of overlaps amongst the binary classification map (see Fig. 3) in which red depicts voxels that emerged across all 7 comparisons.

RESULTS

Acoustic analysis data. We performed a one-way ANOVA using R (v 3.1.3) on each of the acoustic measures (Fig. 2), which revealed a significant main effect of category throughout all analyses [intensity: $F(7,16) = 7.87$, $P < 0.05$; SD of intensity: $F(7,16) = 13.77$, $P < 0.05$; HNR: $F(7,16) = 22.61$, $P < 0.05$; f0: $F(7,16) = 229.91$, $P < 0.05$; SD of spectrum: $F(7,16) = 6.65$, $P < 0.05$; center of gravity: $F(7,16) = 21.76$, $P < 0.05$]. To see if there was a difference between human and each of the nonhuman categories, we performed a post hoc paired *t*-test using Tukey's honest significant difference. Across all pairwise comparisons in each acoustic measure, there were differences between human and some, but not all, of the nonhuman categories (see Table 1 for more detailed statistics).

fMRI data. We first performed a series of binary classification tests for comparing the human category with the seven nonhuman categories. All analyses consistently yielded STS/STG bilaterally, although the exact size and location of clusters varied from one comparison to another (Fig. 3 and Table 2). We then overlaid those seven maps to identify the temporal regions that showed voice sensitivity across all binary comparisons. As can be seen in Fig. 4, voxels along the anterior-to-mid posterior parts of superior and middle temporal lobe bilaterally were found to distinguish human from all seven nonhuman categories (i.e., maximum overlap). When the same analysis pipeline was applied to each of the seven nonhuman categories (e.g., bird vs. six nonbird categories after excluding human), only circumscribed portions of bilateral superior temporal lobe were found for sensitivity to each of the nonhuman categories (Fig. 5). Whereas none of the nonhuman categories showed as robust sensitivity as the human category did in the temporal lobe, the bird category appeared to be relatively well distinguished from the six nonbird categories. Overall, the middle portion of the right STG yielded the strongest sensitivity to all nonhuman categories, as evidenced by the maximal overlap across the seven binary classification maps for nonhuman categories (Fig. 5).

Additionally, we performed a binary classification test to explore cortical areas that can distinguish between living and nonliving auditory objects (i.e., at the superordinate level). To this end, the data of four animal vocalizations and four mechanical noises were collapsed into "living" and "nonliving" categories, respectively. The analysis revealed widespread cortical regions, including temporal, frontal, and parietal cortex bilaterally, with the most robust categorization activity observed in right STG/middle temporal gyrus (MTG; see Fig. 3 and Table 3).

Next, we performed a more difficult classification test: discriminating the human category from all seven other nonhuman categories. In this eight-way classification test, the chance-level performance of discriminating one particular category from the seven other categories is 12.5%. We calculated PPV based on hit and false-alarm rate for each of those eight categories and compared the PPV of human with that of each of seven other categories. Consistent with the binary classification result above, these comparisons also yielded the superior and middle temporal lobe bilaterally (Fig. 6). We overlaid all seven PPV comparison maps to identify regions showing the most robust voice sensitivity. This yielded a much more restricted portion of temporal cortex, comprising three distinct

Table 2. *Cortical regions exhibiting greatest sensitivity toward the human voice in binary and 8-way classification tests*

Region Name	MNI Coordinates			Cluster
	x	*y*	*z*	
Binary classification				
Right middle temporal gyrus	63	−15	−21	610
Left superior temporal gyrus	−57	3	−14	228
8-Way classification				
Right middle temporal gyrus	60	−6	−21	13
Right middle temporal gyrus	57	−24	−14	9
Right middle temporal gyrus	57	−39	−7	69

Clusters with >5 contiguous voxels are included. MNI, Montreal Neurological Institute.

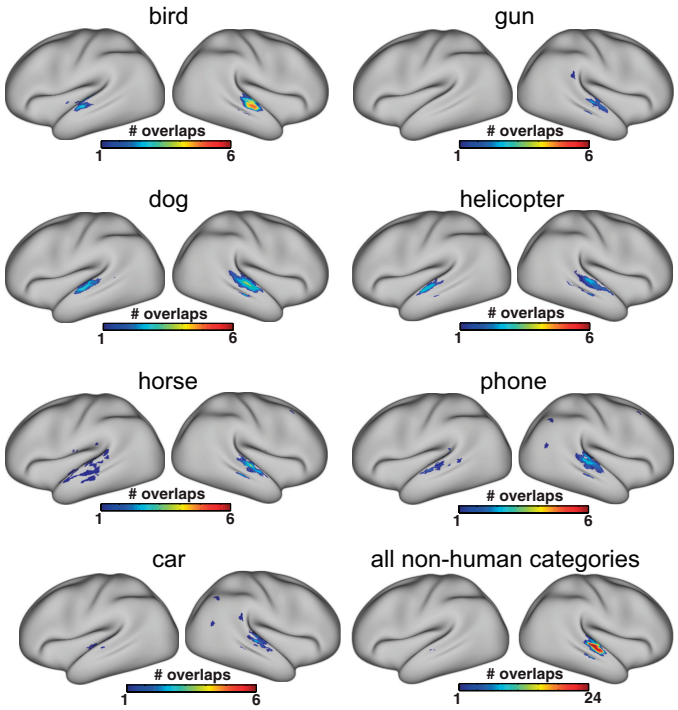

Fig. 5. Overlaps of binary classification for target vs. nontarget, performed on each of the 7 nonhuman categories. For binary classifications on nonhuman categories, human was excluded from the analysis. The color bar indicates the number of overlaps amongst the binary classification map, in which red depicts voxels that emerged across all comparisons.

clusters. As can be seen in Fig. 7, all clusters emerged in right temporal lobe (Table 2). We then plotted sensitivities to all eight categories within each of those clusters to visualize the difference in sensitivity among all categories. The plot for each region clearly shows that human was separated from nonhuman categories with its greater sensitivity. Nonetheless, sensitivity to all nonhuman categories was significantly higher than the chance level, indicating that these areas were capable of categorizing all auditory objects.

DISCUSSION

In the present fMRI study, we used MVPA searchlights in combination with signal detection theory to compare categorization performance between voice and seven other nonvoice categories of auditory objects. A series of binary comparisons revealed that a large expanse of bilateral STG/STS exhibited greater sensitivity (i.e., degree of distinction) for human than for each of the nonhuman categories. A more difficult eight-way classification yielded three robust clusters that distinguished human from all nonhuman categories in restricted portions of bilateral STG/STS. Importantly, these voice clusters were still reliably able to categorize nonhuman categories with a high degree of sensitivity, suggesting that these regions were not exclusively dedicated to the human category. Together, our findings extend the current evidence for voice sensitivity in the temporal cortex by demonstrating superior categorization ability for human voice, rather than simply an increased overall response.

Voice sensitivity: inferences drawn from MVPA and signal detection theory. Since the initial characterization of voice-sensitive regions in the human brain (Belin et al. 2000), numerous studies have replicated the seminal findings and provided more insights into the nature of voice sensitivity [for a detailed review, see Belin et al. (2011)]. For example, Belin and Zatorre (2003) found that right anterior STS/STG showed reduced activity when the same speaker's voice was repeated, even when that voice was pronouncing different syllables. von Kriegstein et al. (2003) reported that recognizing a speaker's identity activated right anterior STS/STG. This region was again found in their follow-up study involving a voice-recognition task (Kriegstein and Giraud 2004). The same study, however, reported that right posterior STS/STG showed the strongest activity during an unfamiliar voice-recognition task. More recently, that right posterior STS/STG was found to be tuned to particular human vocal-track parameters during a speaker-recognition test (von Kriegstein et al. 2010).

There is little doubt that some regions along the STS/STG show increased activity in response to voice and that different voice-sensitive regions participated in different aspects of voice processing (Andics et al. 2010, 2013; Latinus and Belin 2011a; Latinus et al. 2013; Mathias and von Kriegstein 2014; von Kriegstein et al. 2007). In the present study, we tested a hypothesis that the human voice is well distinguished from nonhuman categories of auditory objects in terms of characteristically differential patterns of neural activity. This question can be well suited to a multivariate pattern-analysis approach (Giordano et al. 2013). As discussed in INTRODUCTION, the implications of simple increases in activity can be ambiguous, and positive findings using such an analysis would not necessarily imply that the regions in question are set aside for categorization tasks, since activation to the particular acoustics of voice can occur irrespective of the perceptual categorization process. By contrast, with MVPA, we can objectively measure how the neural representation associated with one object category differs from that associated with others at a particular region. The comparison of the accuracy or sensitivity index (PPV) pertaining to each category thus implies which object category is best distinguished from others (Staeren et al. 2009).

Indeed, all classification tests consistently revealed the most robust sensitivity toward voice within the anterior-to-mid portion of STS/STG bilaterally, with more pronounced patterns in the right hemisphere. Because activation of these regions has been already implicated by conventional neuroimaging studies, finding the voice sensitivity in these regions might be warranted. We note, however, that the novel aspect of the present finding does not lie in the anatomical localization of the results

Table 3. *Cortical regions found in living vs. nonliving binary classification*

Region Name	MNI Coordinates			*z*-Value	Cluster
	x	*y*	*z*		
Right middle temporal gyrus	66	−18	−7	5.27	5,851
Right superior temporal gyrus	48	−33	14	5.2	
Right middle temporal gyrus	−63	−24	0	5.14	
Left precuneus	−6	−54	63	4.47	38
Left precuneus	−12	−75	56	4.41	
Left precuneus	−9	−63	66	3.48	
Right inferior parietal gyrus	48	−33	52	4.04	45
Right postcentral gyrus	51	−30	60	3.93	
Right inferior parietal gyrus	60	−36	49	3.57	
Right insula	42	15	−7	3.99	26
Right insula	36	18	−14	3.45	

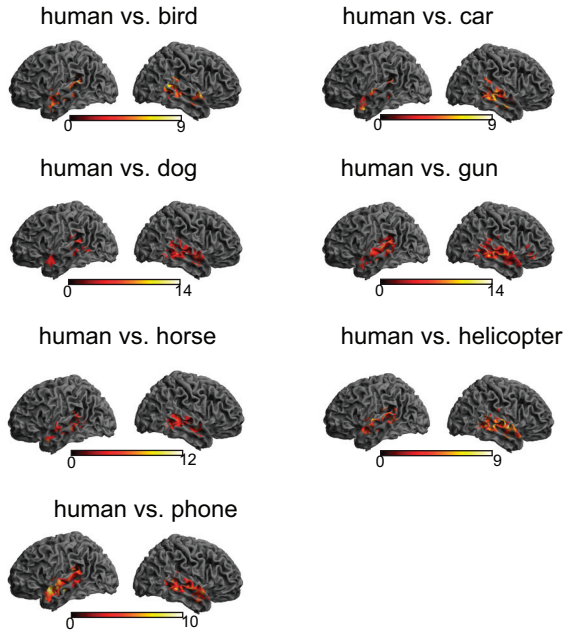

Fig. 6. Cortical rendering views of positive predictive value (PPV) comparisons out of an 8-way classification test. After a PPV map for each category is obtained from the 8-way classification, the sensitivity index of human is compared with that of each of the nonhuman categories at the group level. All comparisons consistently yield greater voice sensitivity in bilateral temporal lobes. The color bar indicates *t*-statistics.

but rather, in the use of a more specific measure of voice sensitivity, indicating the distinct neural characteristics most associated with response to the human voice, as indicated by discrimination from nonhuman categories in performance across all classification tests. To our knowledge, the present study was the first to measure the voice sensitivity using multivariate and signal detection theory, although MVPA was applied to studying other aspects of voice processing, such as emotional (Ethofer et al. 2009) or speaker's gender representation (Ahrens et al. 2014). Among the bilateral temporal foci of voice sensitivity, the more stringent eight-way PPV classification yielded three clusters exclusively within the right middle temporal cortex. This right hemisphere dominance for voice processing is indeed consistent with neuroimaging and patient literature (Latinus and Belin 2011b; Van Lancker et al. 1988). Relatedly, a transcranial magnetic stimulation study demonstrated that stimulating the mid portion of the right STS/STG impaired voice/nonvoice distinction (Bestelmeyer et al. 2011).

The role of the human temporal lobes in auditory object categorization. Another important question that we sought to address was whether those voice-sensitive clusters were exclusively dedicated to the human category. Our results suggest that this is not the case: whereas categorization performance for the seven other nonhuman categories was clearly poorer than that for the human category, the PPVs were still substantially higher than chance-level performance. This indicates that these neural resources are not exclusively devoted to the computation of voice processing but are rather involved in transcribing various time-varying acoustic signals into perceptual units (i.e., a single auditory object) (Leech and Saygin 2011). In the binary classification result, the significant voxels occurred beyond the core auditory cortex (A1) and ran along the mid-to-anterior portions of the STS/STG, which is often termed as the auditory "what" pathway (Bizley and Cohen 2013). Our findings suggest that the human voice is privileged in this pathway. Although such a proclivity for voice sensitivity may be one of the innate features of the human auditory system, a recent study found acquired voice sensitivity in 7-mo-old infants but not in 4 mo olds (Grossmann et al. 2010), speaking to the possibility that voice sensitivity could also reflect auditory expertise (Leech et al. 2009).

Intriguingly, the sensitivity measures for nonhuman categories were comparable with each other, whether those sounds were from living or nonliving objects. This result may speak against the notion that human auditory cortex is more strongly tuned to the living than nonliving auditory categories (Altmann et al. 2007; Doehrmann et al. 2008; Engel et al. 2009; Kraut et al. 2006; Lewis et al. 2005; Murray et al. 2006). For example, past neuroimaging studies have consistently reported that living sounds (e.g., animal vocalizations) activate foci within the superior temporal lobe anterior to Heschl's gyrus, whereas nonliving sounds (e.g., hand-manipulated tools) instead activate higher-order regions beyond the auditory cortex, including left posterior MTG and motor-associated parietal and frontal cortices (Doehrmann et al. 2008; Lewis et al. 2005). In particular, the left STG was exclusively adapted to the animal vocalizations, suggesting that this region is more sensitive to the spectrotemporal properties emanated by living objects (Altmann et al. 2007; Doehrmann et al. 2008). Such neural propensity can be mirrored by a behavioral study showing faster identification on the living than nonliving sounds (Giordano et al. 2010).

Thus whereas it is still plausible that living auditory objects are privileged by the early-to-mid stage of auditory processing, due to their ethological significance, our study points to the evidence that various environmental sounds, whether living or not, appear to be equally well represented in the auditory what pathway, which distinguished them from the conspecific human vocal sounds. Of course, we cannot completely rule out that such distinction is based on differences in low-level acoustic characteristics between living and nonliving categories of auditory objects. However, a recent MVPA study abstracted coherent representation for living and nonliving auditory objects within the STG and planum temporale, which cannot be explained by the low-level acoustic differences

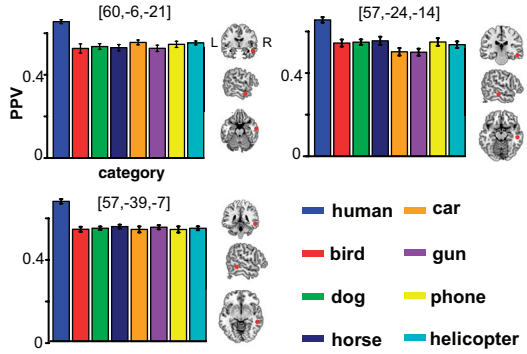

Fig. 7. PPVs of all categories in each of the 3 temporal clusters found by 8-way classification. The bar plot indicates that sensitivity to human category is clearly separated from sensitivity to nonhuman categories in each of the clusters. The red circles depict the local peak of each cluster in the cross-section views.

(Giordano et al. 2013). Furthermore, the fact that such distinction emerged outside of the primary auditory cortex suggests that low-level acoustic properties may not be the entire cause of our finding.

Together, despite an ample body of related work, the questions of how rich acoustic signals contained in myriad environmental sounds are eventually transcribed into distinct auditory objects, as well as how human voice is privileged by the neural process, still deserve further research. Nevertheless, our findings are in accordance with the existing evidence regarding the auditory what processing pathway and the location of voice-sensitive regions. Here, we suggest that these voice-sensitive clusters in mid and anterior STS/STG may serve as generic cerebral processors for auditory object processing, rather than as processors exclusively dedicated to the human voice.

Other considerations. In the present auditory fMRI study, we used a continuous imaging paradigm, meaning that stimuli were presented against a background of acoustic scanner noise. Although sparse imaging (Hall et al. 1999) has been commonly used in auditory fMRI experiments to minimize the impact of scanner noise, a major drawback of the technique lies in the reduced amount of data collected. This is particularly undesirable when using MVPA classification, because insufficient data often result in an overfitting problem (Pereira et al. 2009). Furthermore, although sparse imaging may be well suited to examine tonotopic organization in the primary auditory cortex, there are other regions in which continuous imaging has been shown to produce stronger results (Peelle et al. 2010; Petkov et al. 2009). Accordingly, continuous imaging is often used for investigating higher-order auditory processing, such as music (Janata et al. 2002; Lee et al. 2011), speech (Kriegstein and Giraud 2004; Raizada et al. 2010), and object categorization (Adams and Janata 2002; Altmann et al. 2007; Kriegstein and Giraud 2004). Because our primary aim was to understand better the voice sensitivity, presumably observed beyond the early auditory cortex, we decided to use a continuous imaging paradigm.

Our study differs from some previous voice studies in that we used an overt memory task to ensure subjects were attending to the stimuli. Although we acknowledge that this is an important difference, it is unlikely that our results reflect task-specific effects, as we did not observe any voice sensitivity in the areas in the frontal or parietal cortex, frequently implicated in verbal working memory (Chein and Fiez 2010; Smith and Jonides 1998). In addition, although the current study was designed to engage attention as equally as possible across sound stimuli by using "meaningless" nonspeech vocal sounds, there can be no guarantee that the sounds were semantically empty for all participants. Together, we stress some degree of caution in interpreting our findings, and future studies should address these potential issues in experimental design.

Conclusion. A hitherto unanswered question is whether strong cortical responses to voice reflect improved categorization ability. In the present study, we explicitly tested this possibility using MVPA and signal detection theory and found that voice (i.e., human category) was indeed best distinguished from all other nonhuman categories within the mid-to-anterior portions of the temporal lobes. The present finding may open up new directions for extending the current knowledge of the neural mechanisms of voice processing mediated by temporal cortex.

ACKNOWLEDGMENTS

The authors thank Kwang-Mo Jung and Kathryn West for their help with stimuli selection. The authors also thank Michael Hanke and Sang-Hyeun Park for their helpful advice on conceptualizing multiclass classification using machine-learning techniques. Special thanks from Y. S. Lee go to Yong-Cheol Lee, who provided help on acoustic analysis. Lastly, the authors thank the two anonymous reviewers for their helpful suggestions.

GRANTS

Support for this work was provided by grants from the Office of Naval Research.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

Author contributions: Y.S.L. and R.G. conception and design of research; Y.S.L. and S.L. performed experiments; Y.S.L. and S.L. analyzed data; Y.S.L., J.E.P., and D.K. interpreted results of experiments; Y.S.L., J.E.P., and S.L. prepared figures; Y.S.L. and J.E.P. drafted manuscript; Y.S.L., J.E.P., S.L., and R.G. edited and revised manuscript; Y.S.L., J.E.P., D.K., S.L., and R.G. approved final version of manuscript.

REFERENCES

Adams RB, Janata P. A comparison of neural circuits underlying auditory and visual object categorization. *Neuroimage* 16: 361–377, 2002.

Ahrens M-M, Awwad Shiekh Hasan B, Giordano BL, Belin P. Gender differences in the temporal voice areas. *Front Neurosci* 8: 228, 2014.

Altmann CF, Doehrmann O, Kaiser J. Selectivity for animal vocalizations in the human auditory cortex. *Cereb Cortex* 17: 2601–2608, 2007.

Andics A, Gácsi M, Faragó T, Kis A, Miklósi Á. Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Curr Biol* 24: 574–578, 2014.

Andics A, McQueen JM, Petersson KM. Mean-based neural coding of voices. *Neuroimage* 79: 351–360, 2013.

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z. Neural mechanisms for voice recognition. *Neuroimage* 52: 1528–1540, 2010.

Belin P, Bestelmeyer PE, Latinus M, Watson R. Understanding voice perception. *Br J Psychol* 102: 711–725, 2011.

Belin P, Zatorre RJ. Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14: 2105–2109, 2003.

Belin P, Zatorre RJ, Ahad P. Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 13: 17–26, 2002.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature* 403: 309–312, 2000.

Bestelmeyer PE, Belin P, Grosbras MH. Right temporal TMS impairs voice detection. *Curr Biol* 21: R838–R839, 2011.

Bizley JK, Cohen YE. The what, where and how of auditory-object perception. *Nat Rev Neurosci* 14: 693–707, 2013.

Boersma P. Praat, a system for doing phonetics by computer. *Glot Int* 5: 341–345, 2001.

Chein JM, Fiez JA. Evaluating models of working memory through the effects of concurrent irrelevant information. *J Exp Psychol Gen* 139: 117–137, 2010.

Doehrmann O, Naumer MJ, Volz S, Kaiser J, Altmann CF. Probing category selectivity for environmental sounds in the human auditory brain. *Neuropsychologia* 46: 2776–2786, 2008.

Engel LR, Frum C, Puce A, Walker NA, Lewis JW. Different categories of living and non-living sound-sources activate distinct cortical networks. *Neuroimage* 47: 1778–1791, 2009.

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P. Decoding of emotional information in voice-sensitive cortices. *Curr Biol* 19: 1028–1033, 2009.

Faragó T, Andics A, Devecseri V, Kis A, Gácsi M, Miklósi Á. Humans rely on the same rules to assess emotional valence and intensity in conspecific and dog vocalizations. *Biol Lett* 10: 20130926, 2014.

Fecteau S, Armony JL, Joanette Y, Belin P. Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* 23: 840–848, 2004.

Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322: 970–973, 2008.

Giordano BL, McAdams S, Zatorre RJ, Kriegeskorte N, Belin P. Abstract encoding of auditory objects in cortical activity patterns. *Cereb Cortex* 23: 2025–2037, 2013.

Giordano BL, McDonnell J, McAdams S. Hearing living symbols and nonliving icons: category specificities in the cognitive processing of environmental sounds. *Brain Cogn* 73: 7–19, 2010.

Grossmann T, Oberecker R, Koch SP, Friederici AD. The developmental origins of voice processing in the human brain. *Neuron* 65: 852–858, 2010.

Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW. "Sparse" temporal sampling in auditory fMRI. *Hum Brain Mapp* 7: 213–223, 1999.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293: 2425–2430, 2001.

Janata P, Birk JL, Horn JD, Leman M, Tillmann B, Bharucha JJ. The cortical topography of tonal structures underlying Western music. *Science* 298: 2167–2170, 2002.

Kraut MA, Pitcock JA, Calhoun V, Li J, Freeman T, Hart J. Neuroanatomic organization of sound memory in humans. *J Cogn Neurosci* 18: 1877–1888, 2006.

Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103: 3863–3868, 2006.

Kriegstein KV, Giraud AL. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* 22: 948–955, 2004.

Latinus M, Belin P. Anti-voice adaptation suggests prototype-based coding of voice identity. *Front Psychol* 2: 175, 2011a.

Latinus M, Belin P. Human voice perception. *Curr Biol* 21: R143–R145, 2011b.

Latinus M, McAleer P, Bestelmeyer PE, Belin P. Norm-based coding of voice identity in human auditory cortex. *Curr Biol* 23: 1075–1080, 2013.

Lee YS, Janata P, Frost C, Hanke M, Granger R. Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. *Neuroimage* 57: 293–300, 2011.

Leech R, Holt LL, Devlin JT, Dick F. Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J Neurosci* 29: 5234–5239, 2009.

Leech R, Saygin AP. Distributed processing and cortical specialization for speech and environmental sounds in human temporal cortex. *Brain Lang* 116: 83–90, 2011.

Lewis JW, Brefczynski JA, Phinney RE, Janik JJ, DeYoe EA. Distinct cortical pathways for processing tool versus animal sounds. *J Neurosci* 25: 5148–5158, 2005.

Mathias SR, von Kriegstein K. How do we recognise who is speaking? *Front Biosci (Schol Ed)* 6: 92–109, 2014.

Murray MM, Camen C, Andino SLG, Bovet P, Clarke S. Rapid brain discrimination of sounds of objects. *J Neurosci* 26: 1293–1302, 2006.

Peelle JE, Eason RJ, Schmitter S, Schwarzbauer C, Davis MH. Evaluating an acoustically quiet EPI sequence for use in fMRI studies of speech and auditory processing. *Neuroimage* 52: 1410–1419, 2010.

Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45: S199–S209, 2009.

Perrodin C, Kayser C, Logothetis NK, Petkov CI. Voice cells in the primate temporal lobe. *Curr Biol* 21: 1408–1415, 2011.

Petkov CI, Kayser C, Augath M, Logothetis NK. Optimizing the imaging of the monkey auditory cortex: sparse vs. continuous fMRI. *Magn Reson Imaging* 27: 1065–1073, 2009.

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. A voice region in the monkey brain. *Nat Neurosci* 11: 367–374, 2008.

Raizada RD, Lee YS. Smoothness without smoothing: why Gaussian Naive Bayes is not naive for multi-subject searchlight studies. *PLoS One* 8: e69566, 2013.

Raizada RD, Tsao FM, Liu HM, Kuhl PK. Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cereb Cortex* 20: 1–12, 2010.

Rorden C, Brett M. Stereotaxic display of brain lesions. *Behav Neurol* 12: 191–200, 2000.

Smith EE, Jonides J. Neuroimaging analyses of human working memory. *Proc Natl Acad Sci USA* 95: 12061–12068, 1998.

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. Sound categories are represented as distributed patterns in the human auditory cortex. *Curr Biol* 19: 498–502, 2009.

Taglialatela JP, Russell JL, Schaeffer JA, Hopkins WD. Visualizing vocal perception in the chimpanzee brain. *Cereb Cortex* 19: 1151–1157, 2009.

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, Della Penna S, Feinberg D, Glasser MF, Harel N, Heath AC, Larson-Prior L, Marcus D, Michalareas G, Moeller S, Oostenveld R, Petersen SE, Prior F, Schlaggar BL, Smith SM, Snyder AZ, Xu J, Yacoub E. The human connectome project: a data acquisition perspective. *Neuroimage* 62: 2222–2231, 2012.

Van Lancker DR, Cummings JL, Kreiman J, Dobkin BH. Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex* 24: 195–209, 1988.

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL. Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res* 17: 48–55, 2003.

von Kriegstein K, Smith DR, Patterson RD, Ives DT, Griffiths TD. Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Curr Biol* 17: 1123–1128, 2007.

von Kriegstein K, Smith DR, Patterson RD, Kiebel SJ, Griffiths TD. How the human brain recognizes speech in the context of changing speakers. *J Neurosci* 30: 629–638, 2010.

Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 12: 900–918, 1996.