NeuroImage 62 (2012) 1376–1389

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Processing of vocalizations in humans and monkeys: A comparative fMRI study

Olivier Joly [a,*,1], Christophe Pallier [b,1], Franck Ramus [c], Daniel Pressnitzer [d], Wim Vanduffel [a,e,f], Guy A. Orban [a]

[a] Lab Neuro-en Psychofysiologie, K.U. Leuven, Medical School, Herestraat 49, B-3000, Leuven, Belgium
[b] INSERM-CEA Cognitive Neuroimaging Unit, Neurospin center, Bat 145, F-91191 Gif-sur-Yvette, France
[c] Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS, EHESS, ENS, 29 rue d'Ulm, 75005 Paris, France
[d] Département d'études cognitives, ENS, 29 rue d'Ulm, 75230 Paris cedex 05, France
[e] Athinoula A. Martinos Center for Biomedical Imaging, 13th Street, Charlestown, MA 02129, USA
[f] Harvard Medical School, Department of Radiology, Boston, MA 02114, USA

ARTICLE INFO

Article history:
Accepted 25 May 2012
Available online 31 May 2012

Keywords:
Audition
Macaque
Human
Monkey calls
Speech
Functional brain imaging

ABSTRACT

Humans and many other animals use acoustical signals to mediate social interactions with conspecifics. The evolution of sound-based communication is still poorly understood and its neural correlates have only recently begun to be investigated. In the present study, we applied functional MRI to humans and macaque monkeys listening to identical stimuli in order to compare the cortical networks involved in the processing of vocalizations. At the first stages of auditory processing, both species showed similar fMRI activity maps within and around the lateral sulcus (the Sylvian fissure in humans). Monkeys showed remarkably similar responses to monkey calls and to human vocal sounds (speech or otherwise), mainly in the lateral sulcus and the adjacent superior temporal gyrus (STG). In contrast, a preference for human vocalizations and especially for speech was observed in the human STG and superior temporal sulcus (STS). The STS and Broca's region were especially responsive to intelligible utterances. The evolution of the language faculty in humans appears to have recruited most of the STS. It may be that in monkeys, a much simpler repertoire of vocalizations requires less involvement of this temporal territory.

Introduction

The evolutionary origins of human language remain largely mysterious. To mediate social interactions, many non-human species use vocalizations which might constitute precursors of human speech. These vocalizations can convey meaning (e.g. alarm calls in vervet monkeys; Seyfarth et al., 1980), as well as the identity and the emotional state of the speaker (Banse and Scherer, 1996; Ghazanfar et al., 2007). An important question is to what extent the processing of vocalizations in humans relies on mechanisms shared with our close relatives, non-human primates. Did the human auditory system become very different from that of other primates because of special demands of speech perception (Liberman and Mattingly, 1985)? The data available do not provide a clear-cut answer. Cross-species studies have shown that non-human animals can be as sensitive as humans to differences between human speech sounds (Brown and Sinnott, 2006). Like humans, monkeys spontaneously perceive changes in formant frequencies (Fitch and Fritz, 2006) and recognize other individuals by their voices (Ghazanfar et al., 2007). From such behavioral evidence, Hauser et al. (2002) argued that, with respect to speech perception, "The available data suggest a much stronger continuity between animals and humans than previously believed".

Monkeys' and humans' responses to vocalizations have already been explored in several single-unit recording and brain imaging experiments, although never within a single comparative study. Studies on monkeys have reported neurons selective for monkey calls in different regions such as the anterior lateral belt of the auditory cortex (Rauschecker et al., 1995), the ventrolateral prefrontal cortex (Romanski, 2004) and the insular cortex (Remedios et al., 2009; for a review see Romanski and Averbeck, 2009). Positron Emission Tomography (PET) in monkeys (Gil-da-Costa et al., 2004, 2006) reported that conspecific vocalizations elicited greater activity than non-biological sounds in higher-order visual areas of the temporal lobe (TEO, TE and superior temporal sulcus, STS), as well as in the temporo-parietal area (Tpt) and the ventral premotor cortex, considered by these authors as homologous to Wernicke and Broca's areas, respectively. In contrast, a recent monkey fMRI study (Petkov et al., 2008) found no strong preference for species-specific vocalizations in these areas but instead reported a "voice" region in the anterior superior temporal plane. This region showed response suppression when several vocalizations from the same individuals were played, similar to a

* Corresponding author at: Institute of Neuroscience, Newcastle University Medical School, Newcastle upon Tyne, NE2 4HH, UK. Fax: +44 191 222 5227.
E-mail addresses: olivier.j.joly@gmail.com (O. Joly), christophe@pallier.org (C. Pallier), franck.ramus@ens.fr (F. Ramus), daniel.pressnitzer@ens.fr (D. Pressnitzer), wim.vanduffel@med.kuleuven.be (W. Vanduffel), guy.orban@med.kuleuven.be (G.A. Orban).
[1] The two first authors contributed equally.

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.05.070

previous finding in humans that reported adaptation to the speaker's voice in the right anterior superior temporal gyrus (STG, Belin and Zatorre, 2003).

In humans, listening to speech or to vocal non-speech sounds, compared to environmental or musical sounds (Belin et al., 2000) or to other animals' vocalizations (Fecteau et al., 2004) activates several bilateral regions along the STS. When compared to acoustic controls, responses to speech in passive listening conditions can be bilateral (Hickok and Poeppel, 2004) or left-dominant (Narain et al., 2003), while responses to non-linguistic vocal sounds (laughs, cries, moans, sighs) involve mainly the right anterior STS (Belin et al., 2002; Meyer et al., 2005).

Here, we compared the neural substrates involved in processing vocalizations in rhesus monkeys and humans, using whole-brain functional magnetic resonance imaging (fMRI). Rhesus monkeys and humans were scanned while listening to monkey calls, human speech, human emotional non-speech vocalizations, bird songs (for humans only) and acoustic controls matched in spectral content (see Fig. 1). For humans there was an additional distinction in that speech stimuli could either be intelligible (mother language) or not (foreign language). The goal of the study was to compare cortical activations in humans and monkeys tested under experimental conditions, as similar as possible, in order to determine to what extent monkey cortical activations associated to processing of vocalizations resemble that of humans. There were two main questions of interest: first, would we observe species-specific responses, that is, regions responding more strongly to monkey calls than to human vocalizations in monkeys, and vice versa in humans? Second, would the pattern of areas activated by monkey calls in monkeys be similar to that involved in low-level speech processing (unintelligible speech), high-level speech processing (intelligible speech), or emotional vocalizations in humans?

In a previous report (Joly et al., 2012), the data acquired with monkeys were studied with a focus on detailed analyses of the acoustic properties of the signal associated with the activations observed. In the present paper, we perform a systematic comparison of activations obtained using very similar protocols in monkeys and in humans.

Materials and methods

Subjects

Monkeys

Three adult rhesus monkeys (*Macaca mulatta*), one female (M13) and two males (M14, M18), 5 to 6 years of age and weighing between 4 and 5 kg, participated in the experiment. These animals were born in captivity and had social experience limited to interaction with conspecifics in group housing and with humans during experiments. Before the scanning sessions, monkeys were trained daily to perform a visual fixation task with the head rigidly fixed to a primate chair. The fixation task was used to equalize attention across conditions and minimize body movement during scanning. The monkeys had little or no prior exposure to the French language, as this is not the primary language spoken in the laboratory. However, they were exposed daily to human voices in the animal facilities from both the radio and communication between monkey handlers. Details concerning head-post surgery and behavioral procedures are described in Vanduffel et al. (2001). Animal care and experimental procedures met the National and European guidelines and were approved by the local ethical committee.

Humans

Twenty right-handed native French speakers (9 men; 11 women; average age = 23.7 years (range, 20 to 28 years) with no history of neurological or psychiatric disease, participated in the experiment. None understood Arabic, the language of the stimuli in the unintelligible-speech condition. All participants gave written informed consent and were paid for their participation.

Stimuli

Five classes of sounds were used to construct the stimuli used in the experiment (Fig. 1): monkey vocalizations, human emotional (non-linguistic) vocalizations, intelligible speech, non-intelligible speech and bird songs. One hundred monkey vocalizations (Mvoc), uttered by several individuals of both sexes and drawn from the Rhesus Monkey Repertoire recorded in Cayo Santiago, Puerto Rico over a period of several years were provided by Marc Hauser. We selected recordings of five types of social calls which were described as having either positive (coos, girneys and harmonic arches) or negative valence (screams and shrill barks) (Gouzoules et al., 1984; Hauser and Marler, 1993). Human vocalizations, including intelligible speech (French), unintelligible speech (Arabic) and emotional sounds (Hemo), were recorded from eight speakers (4 females and 4 males) while others were extracted from movie soundtracks. In an attempt to match the typical brevity of the monkey calls, speech utterances were very short sentences ("It is raining", "It is not possible"…) or interjections ("Hi!", "Excuse me!") averaging 1 s in duration. The human emotional (Hemo)

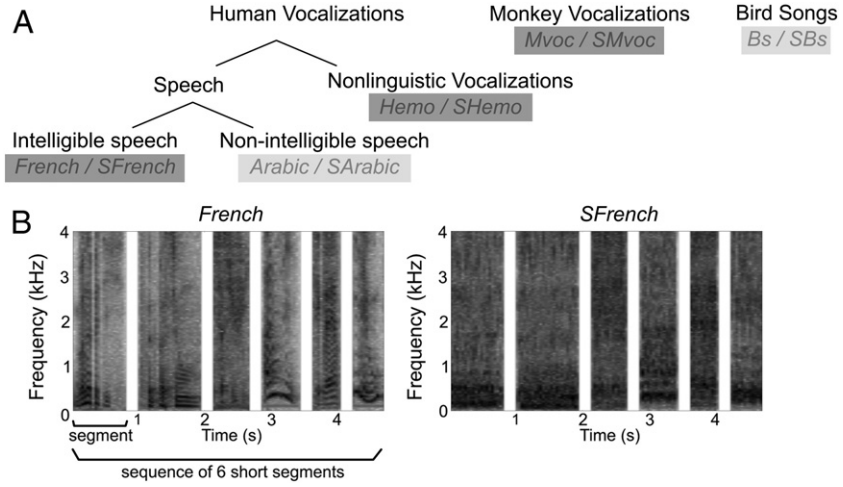

Fig. 1. Different categories of vocalizations. (A) The five sound categories are: intelligible speech, unintelligible speech, non-linguistic vocalizations (emotional sounds), monkey vocalizations and bird songs. Light gray represents conditions presented to human subjects only. For each category, the abbreviated condition label is indicated, the prefix *S* stands for the scrambled control. (B) Spectrograms of an example stimulus from the French category and the corresponding scrambled control (SFrench). A stimulus is defined as a sequence of short segments.

segments, had either a positive (e.g. laughter, contentment) or a negative valence (e.g. cries, shouts). They were uttered by the same speakers who recorded the French stimuli, and did not contain any identifiable phonetic element. Finally, bird songs (Bs) were extracted from high-quality field recordings of a variety of species, with the constraint that only one bird was singing at any time, and the duration of the segments matched that of the other sound categories.

The experimental stimuli were created by concatenating series of sounds, blocked by category (moreover, for the monkey calls (Mvoc) and human emotional sounds (Hemo), the recordings used in a given stimulus were blocked by valence). Successive sounds were separated by silent intervals of 200–250 ms. Because the durations of the silent gaps between MRI volume acquisition differed in experiments with monkeys or humans (see Magnetic resonance imaging section), the series presented to the monkeys were shorter (mean total duration of 2133 ± 224 ms) than those presented to the human participants (mean total duration of 5092 ± 263 ms). Human participants also heard Arabic and bird songs (see Fig. 1).

The intent of such strings of sounds was to maximize the amount of stimulation, but it must be noted that individual rhesus monkeys do not produce such sequences in the wild. Similarly, the concatenated segments of human speech were also unnatural in the context of a conversation (e.g. "I'll take a taxi", "Hello", "It's not possible!" spoken by different speakers).

Scrambled controls

Because the sounds from the different categories had different acoustic characteristics (see Joly et al., 2012 for spectrograms and various measurements), we added "scrambled" control stimuli for each of the 5 sound categories (SMVoc, SHemo, SFrench, SArabic, SBs). Scrambled sounds were made by processing all intact sequences through a gammatone filterbank (Patterson et al., 1995) with 64 channels. As in Patterson et al. (1995), the filterbank was chosen to mimic human frequency selectivity. The equivalent rectangular bandwidth (ERB) of each channel was thus set to ERB = 24.7 (1 + 4.37 F), with F being the center frequency in kHz. This choice was motivated by the observation that macaque monkeys have a peripheral frequency selectivity that appears to be comparable to that of humans (Ruggero and Temchin, 2005; Serafin et al., 1982). In each channel, the signal was windowed with overlapping Hanning windows of 25 ms duration. The windows were then shuffled randomly within a channel, with the additional constraint that a window could be displaced by no more than ±500 ms from its original temporal position. The scrambled signals were finally obtained by putting all frequency channels back together. This method produces an exact match of spectral excitation patterns between original and scrambled signals, while rendering speech totally unintelligible. The resulting scrambled controls sounded like flowing water. Example stimuli are available online (brainsenses.x10hosting.com/joly/monkeyvshuman.html).

Magnetic resonance imaging

During scanning, both human and monkey subjects had to gaze at a small fixation point presented at the center of the projection screen. Each functional time series consisted of gradient echo echo-planar whole brain images (GE-EPI) in a sparse acquisition scheme. The MR acquisition parameters in monkeys (M) and in humans (H) were as follows: time to repeat (TR in s) = 5 (M) and 10 (H); acquisition time (in s) = 2.2 (M) and 2.4 (H), time to echo (TE, in ms) = 27 (M) and 60 (H); slice thickness (in mm) = 2 (M) and 4 (H); matrix size = 64 × 64 (M and H) and spatial resolution (mm) = 2 × 2 × 2 (M) and 3 × 3 × 4 (H). Sounds were delivered binaurally using MR-compatible headphones and were presented centered in the silent gap (2.8 s in monkeys, 7.6 s in humans) between the acquisitions of two functional volumes.

Monkeys

The monkeys sat in a sphinx position in a plastic chair within the horizontal 1.5-T whole body MRI system (Sonata, Siemens medical solutions, Erlangen, Germany). Functional MR images were acquired using a receive-only surface coil positioned over the head. Sounds were presented through electrodynamic headphones (Baumgart et al., 1998) integrated into the ear cups to attenuate scanner noise (MR Confon GmbH, Magdeburg, Germany). The position of one eye was monitored at 120 Hz using a pupil–corneal reflection tracking system (Iscan, MA, USA). Monkeys received a juice reward for maintaining fixation within a small window centered on the fixation target. Before each scanning session, monocrystalline iron oxide nanoparticle contrast agent (MION, Sinerem) was injected into the saphenous vein (4–10 mg/kg) to increase the contrast-to-noise ratio (Leite et al., 2002; Vanduffel et al., 2001). Each fMRI acquisition (run) consisted of 112 volumes. A run was divided in 14 blocks of 8 volumes (40 s/block). During a block, one of the 7 conditions (6 sound conditions: French, Hemo, Mvoc, SFrench, SMvoc, SHemo, and silent baseline) was presented 8 times. The block orders were randomized across runs. Hence, within a run (~10 min), each sound condition was played in two blocks of 8 presentations (a total of 96 sound conditions/run). A total of 96 runs (10752 volumes) were acquired across all scanning sessions. Based on the quality of fixation displayed by the monkey, a subset of 84 runs (28 runs per individual, total of 9408 volumes) entered the group analysis. This dataset represented a total of 392 (28 × 14) blocks and 2688 (28 × 96) sound condition presentations (448 for each sound condition) per individual. A T1-weighted MRI (M12, matrix size: 178 × 256 × 256, 0.35 × 0.35 × 0.35 mm voxels) was used as an anatomical template. This dataset is identical to that of the study of Joly et al. (2012).

Humans

The human participants were asked to listen attentively to the stimuli delivered through piezoelectric headphones. They wore earplugs to shield them from the scanner noise, and sound intensity was adjusted individually to the most comfortable level. Scanning was performed on a 3 T whole-body MRI system (Bruker, Germany) using a standard head coil. A T1-weighted anatomical scan (FOV 256 × 192 × 153.6 mm; resolution of 1.3 × 1.2 × 1.2 mm) was acquired for each participant.

A run (35 volumes) consisted of the presentation of 25 intact conditions (5 each for French, Arabic, Hemo, Mvoc, Bs), 5 scrambled conditions (each belonging to one of the SFrench, SArabic, SHemo, SMvoc, SBs conditions) and 5 silent baselines. Ten runs of 6 min were administered to each participant, resulting in the presentation of 50 stimuli from each intact condition, and 10 stimuli from each of the scrambled controls. In total, 7000 volumes were acquired from the human group.

Data analysis

For both species, fMRI data were analyzed using SPM software (version SPM5, Wellcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). Spatial preprocessing consisted of motion correction (realignment), spatial normalization to a species-specific anatomical template and spatial smoothing with an isotropic Gaussian kernel. The motion parameter estimates were included in the SPM statistical models to regress movements out of the MR signal. Then, SPM maps were projected onto a flattened cortical surface of either the monkey template anatomy or the human PALS-B12 (Van Essen, 2005) using caret software (Van Essen et al., 2001, brainmap.wustl.edu). Some analysis parameters were optimized within each species and are described below.

Monkeys

For monkeys, the functional images were first rigidly registered to the template anatomy (M12, in stereotaxic space) and further

warped to the template using a non-rigid matching technique (BrainMatcher software, INRIA) to compensate for echo planar image distortion and inter-individual anatomical differences. The images were resampled to 1 mm isotropic and smoothed with a kernel of 1.5 mm full-width-at-half-maximum (FWHM). Fixed-effect group analysis was performed with equal numbers of 28 runs per monkey. Using MION as an exogenous contrast agent, MR signal change reflects changes in cerebral blood volume (CBV) and differs from the blood oxygenation-level dependent (BOLD) effect measured in human fMRI. Therefore, our CBV-weighted fMRI analysis was performed using a Mion hemodynamic response function (HRF), as defined earlier (Vanduffel et al., 2001), instead of the BOLD HRF available in SPM. The M12 template anatomy (after skull stripping) was registered to the population-average MRI-based template for rhesus monkey, later referred to as 112RM-SL (McLaren et al., 2009) which is also aligned to the MRI volume from a histological atlas (Saleem and Logothetis, 2006). This registration was performed using the non-rigid, symmetric diffeomorphism approach (SyN) implemented in the ANTS (version 0.5) package (Avants et al., 2008). Activity profiles were extracted with Marsbar SPM toolbox (marsbar.sourceforge.net) and the mean percent of signal change (standard error of the mean across runs) is displayed for each condition relative to the "silent" baseline.

Humans

In humans, BOLD-weighted functional images were spatially normalized to the MNI 152 template, and smoothed with a kernel of 5 mm (FWHM). Note that the size of this kernel, commonly applied to human fMRI data, is about 3 times greater than that applied to the monkey functional images and that this ratio is similar to the ratio of brain sizes in the two species. For each subject, a Finite Impulse Response model of order 1 (1 stick function per stimulus) was used with one regressor for each of the ten conditions: five intact categories (French, Hemo, Mvoc, Arabic and Bs) and five scrambled controls (SFrench, SHemo, SMVoc, SArabic and Sbs). The silent condition was not modeled and served as an implicit baseline. For the random-effect group analysis, the individual contrasts representing each condition (versus the implicit, silence baseline) were smoothed at 8 mm to overcome the individual anatomical differences and entered into a within-subject analysis of variance model, allowing us to define contrasts comparing the conditions (Henson, and Penny, 2003). Activity profiles were derived from contrast estimates and represent the mean, across subjects, of the percent signal change relative to the "silent" baseline.

Results

The results are presented following a hierarchical approach. We start by describing, in each species, the areas activated by the scrambled, acoustic control, stimuli (1st stage), then we proceed to examine the areas responding more strongly to the intact primate vocalizations than to their scrambled controls (2nd stage), and finally we directly compare the responses evoked by the different types of vocalizations (3rd stage).

Responses to scrambled, acoustic controls

We first examined the responses to the scrambled control stimuli, which did not contain high-level, species-specific information, nor intelligible content. As shown on Fig. 2, which displays the areas collectively activated by scrambled stimuli (SMvoc + SFrench + SHemo) compared to silent baseline, these sounds activated large networks in both species, extending much beyond primary auditory areas. In monkeys, the contrast revealed significant voxels in a large portion of the lower bank of the lateral sulcus (LaS), in the superior temporal gyrus (STG), the anterior part of the intraparietal lobule (IPL), the left lateral bank of the intraparietal sulcus (IPS) and in the premotor cortex (F5c) bilaterally. In humans, activations were observed in Heschl gyri, the planum temporale, the planum polare, the STG and the upper bank of the STS, and in small regions of the supramarginal, inferior frontal and precentral gyri bilaterally. The pattern of activated regions was similar to the global network obtained by contrasting all the stimuli (intact + scrambled) with silence (see Joly et al., 2012 for monkey) but with reduced extents in the STS and frontal regions of humans.

The scrambled stimuli differed acoustically, mostly in spectral content. To locate regions sensitive these differences, we used an F-test contrasting the SMvoc, SFrench and Shemo categories. This contrast identified small regions in both species (outlined in light-blue on the flat maps of Figs. 2A and B; threshold: $p<0.05$ FWE-corrected). In monkeys, the maxima of these activations were located bilaterally in the ventral banks of the LaS near the local maxima for the main auditory activation, most likely in area A1. In humans, the activation was located at the most lateral tip of the Heschl's gyrus at coordinates (−51; −18; 6) and (57; −9; 3) in the left and right hemispheres, respectively. Human primary auditory cortical areas A1 and R have been localized to this region (Formisano et al., 2003; Langers and van Dijk, 2011). The activity profiles at the local maxima within the blue outlines in Fig. 2 are shown in Figs. 3A and B for monkeys and humans, respectively (note that measurements differ between species: BOLD effect in humans and CBV (negative) weighted signal in monkeys and error bars represent standard-errors over entire runs in monkeys or over subjects in humans. Therefore, the absolute values and their associated standard errors should be considered only within each species, and effects should be evaluated across conditions and not across species). In each plot, the profiles across scrambled conditions are similar to the profiles across intact stimuli (black lines in Fig. 3). In these regions, the responses to intact stimuli are likely driven by the spectral content that remains in the scrambled stimuli. Hence, this first stage of auditory processing reveals a high degree of similarity across species concerning both the extent of the auditory-related activations and the localization of frequency-dependent effects within or close to primary auditory cortices. Also, in both species, the species-specific vocalizations evoked slightly more activity at this level than the other sounds (Fig. 3). Importantly, both the relative extent of the activation and the significance levels reached (Fig. 2) indicate that sensitivities of the human and monkey analyses are reasonably comparable, despite differences in acquisition parameters.

Responses to intact versus scrambled vocalizations

To identify regions sensitive to vocalizations, we first contrasted all intact versus all scrambled stimuli [(French + Hemo + Mvoc) − (SFrench + SHemo + Smvoc)]. This contrast is displayed using the yellow − red scale on Fig. 4, overlapped on the global activation network excited by all sound conditions (French + Hemo + Mvoc + SFrench + SHemo + Smvoc) versus silence (shown in green). The monkey results, shown in Fig. 4A, are identical to those reported in Fig. 5A of Joly et al. (2012). This analysis revealed activations restricted to the most lateral part of the ventral bank of the LaS, namely in the lateral belt, the STG (parabelt auditory regions) and in the left orbito-frontal cortex. In humans (Fig. 4B) the same contrast revealed activations in the STG bilaterally (with local maxima at −60; −15; −3 and 60; −18; −3) extending into the STS (mostly in the upper bank). Again, the relative extent of the activation in this contrast, as well as the significance levels reached, clearly suggest that the analyses of monkey and human data are comparable in sensitivity, despite a number of differences in the methods used.

The next step was to compare the intact versus scrambled stimuli separately for each type of vocalization, both in monkeys and humans. These contrasts are displayed on Figs. 5–7. Within a category, when the contrast *intact* versus *scrambled* did not reach the threshold $p<0.05$ FWE-corrected either in monkeys or humans, the

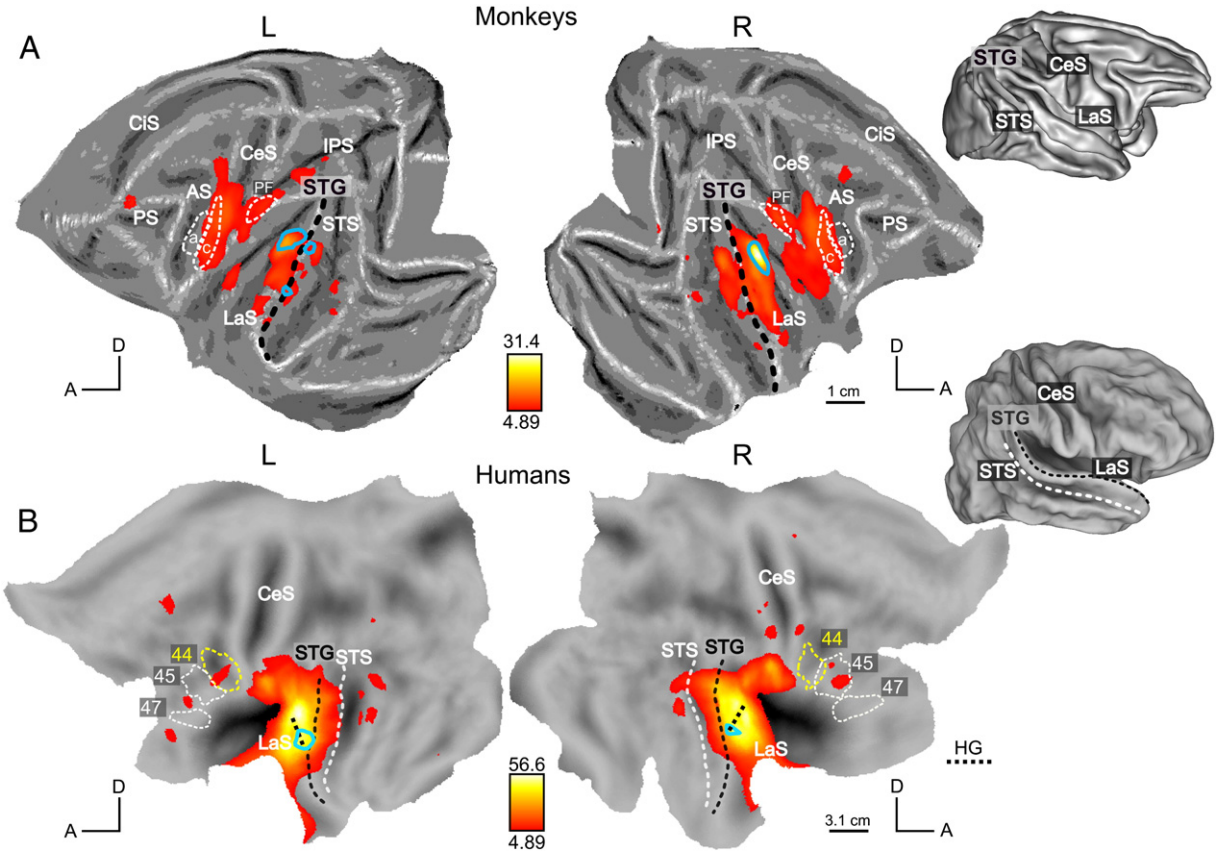

Fig. 2. Activation elicited by scrambled acoustic controls. To facilitate the comparison between species, activations are displayed on flattened representations of the two hemispheres in both species; lateral views of the 3D brains indicate anatomical landmarks (dashed lines) on the flat maps. (A and B) Activation by scrambled controls (SFrench, SHemo, SMvoc vs. silence) in monkeys (A) and humans (B). SPM T-maps ($p<0.05$ FWE corrected) are shown on the left and right flattened hemispheres of template anatomy M12 (A) and human PALS Caret surface (B). White and yellow closed dashed outlines represent anatomical regions: in monkeys, F5a and F5c in the ventral premotor cortex (Belmalih et al., 2009) and area PF (area 7b) in the anterior part of the inferior parietal lobule. In humans, Brodmann' areas 44 and 45 were functionally defined by Amunts et al. (1999, 2004), and area BA47 was extracted from the PALS atlas using Caret. Abbreviations: CeS, central sulcus; LaS, lateral sulcus; AS, arcuate sulcus; PS, principal sulcus; IPS, intraparietal sulcus; STS, superior temporal sulcus. The light-blue outlines represent a significant effect for the difference among scrambled controls, F-maps thresholded at $p<0.05$ FWE corrected.

map is shown at a lower threshold ($p<0.001$, uncorrected for multiple comparisons) in both species.

The contrast between monkey vocalizations and their corresponding scrambled stimuli (Mvoc–SMvoc) is shown on Fig. 5. It did not reach $p<0.05$ FWE-corrected in monkeys, but the most significant voxels were located in the LaS and in the STG corresponding to the lateral belt and the parabelt region, respectively. In humans, this contrast revealed activations in the STG, lying mainly dorsal to the black dashed line that represents the crown of the STG (Fig. 5B).

The contrast between human emotional vocalizations and their scrambled controls (Hemo–Shemo) is shown on Fig. 6. It did not reach $p<0.05$ (FWE-corrected) in humans, but its maximum was found in the STG and were located mainly ventrally with respect to the black dashed line (Fig. 6A). In monkeys (Fig. 6B), this contrast resulted in activations in the LaS, the STG and also the left orbito-frontal regions (as in Fig. 4A).

Figs. 7A and B show the responses to intact French stimuli versus their scrambled counterparts (French–SFrench). In monkeys, this contrast yielded more reliable activations (Fig. 7A) than the previous contrasts (Mvoc–SMvoc; Fig. 5A) or (Hemo–SHemo; Fig. 6A). The most significant voxels were again found in the LaS and in the STG but we also observed activation in the orbito-frontal region, similar to the activation detected by the main effect Intact–Scrambled (Fig. 4A), and similar to the orbito-frontal activation for Hemo–SHemo (Fig. 6A, at uncorrected level). In humans, the contrast [French–SFrench] elicited widespread activations in the STG/STS (Fig. 7B) and the contrast [Arabic–SArabic], controlling for intelligibility, had a relatively similar effect (Fig. 7C). Both contrasts were significant in regions mainly ventral to the black dashed line of the STG and these activation sites were therefore more ventral than those in the contrasts [Mvoc–SMvoc] and [Hemo–Shemo] (Figs. 5B and 6B).

Thus the human STG and upper bank STS may be functionally equivalent to the monkey lateral belt and parabelt, yet this human region is much more specialized for speech signals than is its monkey counterpart. Indeed compared to their scrambled controls, French and Arabic activate the human STG much more strongly than Hemo or Mvoc compared to their controls. In the monkey, the conspecific vocalization compared to its scrambled control evoked less activation of the belt and parabelt than either French or Hemo, relative to their scrambled counterparts.

Comparisons across vocalizations

This last stage aims at testing for species-specificity in both species and to compare the responses to human vocalizations in humans with the responses to monkey vocalizations in monkeys. Therefore, we searched for category-specific effects by comparing the different categories of stimuli to each other. In a first approach, we computed the interaction contrasts (e.g. (French − SFrench) − (Mvoc − SMvoc)) to take into account potential acoustic differences between categories. However, we found that these interaction contrasts lacked statistical sensitivity due to the uncertainty in the estimation of the effects of scrambled stimuli as illustrated by the error bars in Fig. 3B. Therefore, we decided to compare the intact stimuli of the different categories directly, but,

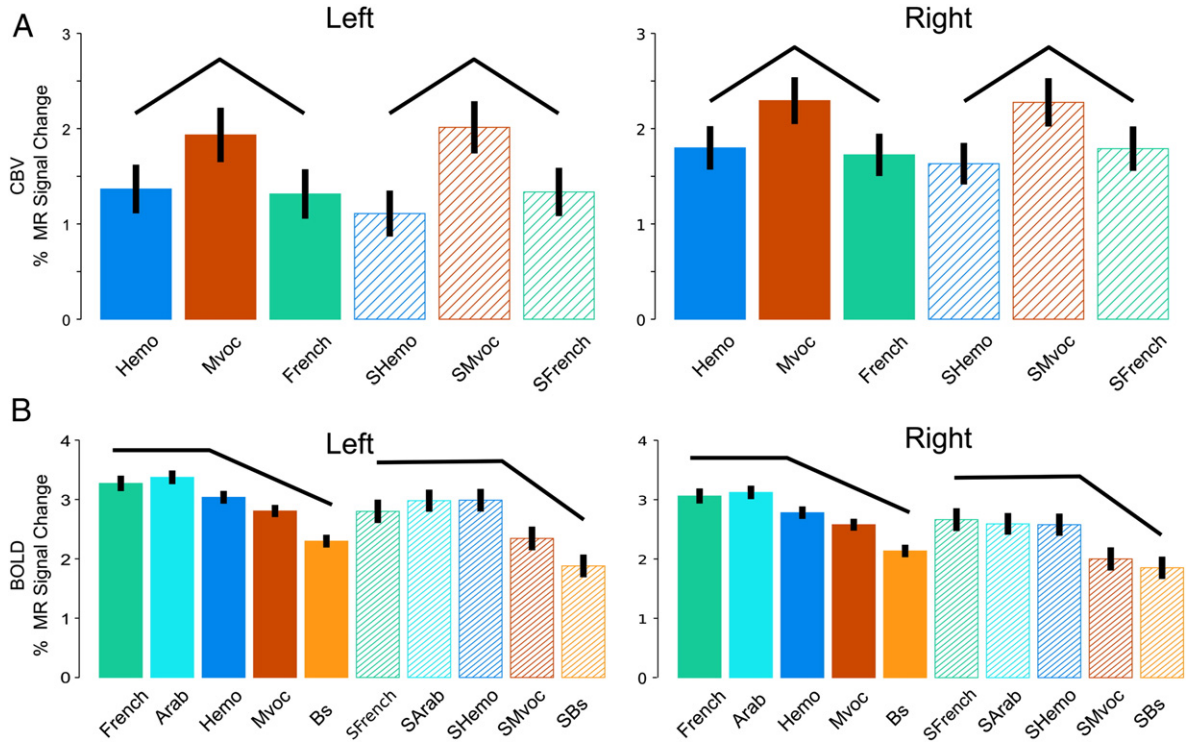

Fig. 3. Activity profiles in regions showing differences across scrambled stimuli. Activity profiles plotting MR signal change for the various conditions in monkeys' (A) and humans' (B) local maxima within the blue outlines in Fig. 2A (−17, 0, 22 and 17, 0, 22) and 2B (−51, −18, 6 and 57, −9, 3). Black lines indicate similar profiles in intact and scrambled conditions. Vertical bars represent SE across runs (A) and subjects (B).

crucially, excluding the regions showing differences between the corresponding scrambled stimuli (thresholded at $p<0.05$ voxel-based, uncorrected).

In monkeys, none of the direct contrasts (contrasting French, monkey vocalizations, and human emotional vocalizations) reached statistical significance even at a low threshold of 0.001 voxel-based uncorrected. In humans, these direct contrasts yielded significant activations displayed in Fig. 8 (red–blue color map for positive–negative t-scores). The subtraction French speech vs monkey calls addresses our first question, that of species specific responses. Fig. 8A shows the regions responding more to French speech than to monkey vocalizations, which included the gyral surface of the STG, bilaterally, the STS (mainly in the upper bank located between white and black dashed lines), inferior frontal areas, bilaterally, as well as a left precentral region. In an attempt to disentangle the effects of species and speech as such, we tested two further contrasts: emotional stimuli compared to monkey calls targeting chiefly the factor species, and French compared to emotional utterances to isolate, although not perfectly, the factor speech. Human emotional vocal utterances, also provoked stronger responses than monkey vocalizations, mostly in the right STG/STS (Fig. 8B). The comparison between French utterances and human emotional sounds, contrasting intelligible speech versus non-speech, showed stronger activations in the bilateral STG/STS, in the left inferior frontal and in precentral regions (Fig. 8C).

Finally, we considered the stimuli heard only by the human participants, that is, non-native, unintelligible speech (Arabic) and bird songs. Fig. 9 shows the relevant contrasts. Responses to Arabic were contrasted to Hemo, Mvoc and French (intelligible speech). The presentation of utterances from a non-intelligible foreign language to the humans provides a relevant counterpoint to the French stimuli in monkeys (to whom French is also a non-intelligible foreign language). The comparison between Arabic and monkey calls (Fig. 9A) showed activations in the STG/STS in a manner similar to the contrast French-Mvoc (Fig. 8A), but does not activate the frontal regions, except for a small left precentral site. Similarly, the contrast Arabic vs. human emotional sounds, which is a fairer contrast to compare to French vs emotional utterances in monkeys, also yielded activations restricted to the temporal region (Fig. 9B), centered on the STG and extending into the STS on the left. The comparison between French (intelligible speech) and Arabic (unintelligible speech for our participants) which targets the highest levels of language processing (parsing and understanding) is shown in Fig. 9C. French stimuli elicited stronger responses than Arabic in the STS (centered on the fundus of the STS marked by white dashed line), bilaterally, (maxima in the anterior STS at −54, −3, 18 and 51, 9, −24), in the inferior frontal gyrus, bilaterally (maxima at −51, 27, −3 and 48, 30, −3) and in the left precentral gyrus (−45, 0, 54).

Since monkey vocalizations are meaningless for humans and may just appear as an animal sound, we contrasted it to another animal sound, bird song. Monkey calls elicited, in humans, activations that were stronger than those produced by listening to bird songs within the posterior part of the LaS bilaterally (no figure shown).

Discussion

In this study, we have compared the processing of vocalizations in human and non-human primates at three levels of processing: the first level targeted early auditory processing, the second level targeted higher-order auditory processing, namely complex spectro-temporal processing of sounds, and finally the third level, compared the different categories of vocalizations. At the first level, we observed widespread activations in response to vocalizations or to their scrambled controls, both in monkeys and humans, in the temporal cortex but also in frontal and parietal areas (Fig. 2, and green map in Fig. 4). At the second level, we observed similar maps in both species, showing preferences for intact sounds over scrambled controls in non-primary auditory cortices (red map in Fig. 4). At the last stage, we obtained a hierarchy of preferences in humans: for human

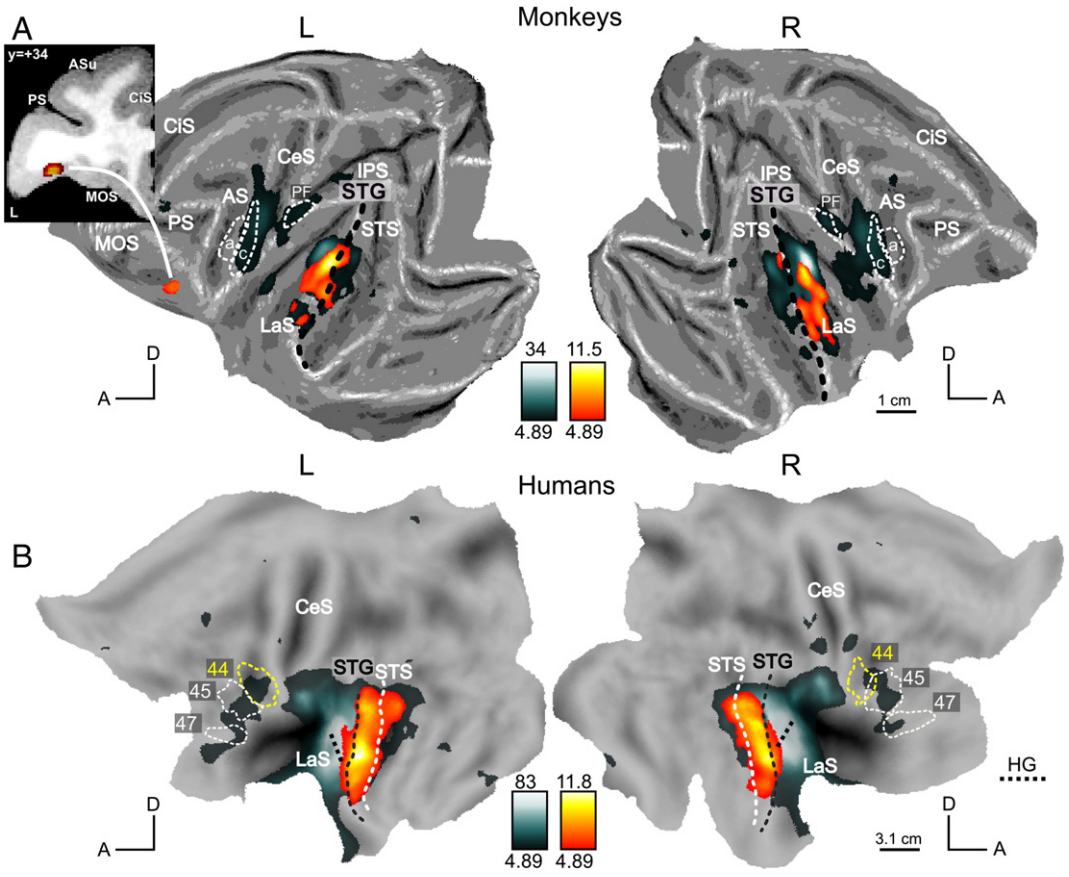

Fig. 4. Processing of primate vocalizations. (A and B) In red, SPM T-maps for the contrast [(French + Hemo + Mvoc) − (SFrench + SHemo + SMvoc)] at $p<0.05$ FWE corrected, masked inclusively with the positive effect of intact vocalizations ((French + Hemo + Mvoc), $p<0.05$ uncorrected) in monkeys (A) and humans (B). In green, auditory activation for [French + Hemo + Mvoc + SFrench + SHemo + Smvoc] vs silence. The monkey orbito-frontal activation is shown on a coronal slice ($y = +34$) overlaid onto the anatomical template (after registration to 112RM-SL space). Abbreviations: MOS, medial orbital sulcus; ASu, arcuate sulcus upper limb; CiS, cingulate sulcus.

vocalizations over monkey calls in the STG, for speech sounds over non-linguistic sounds in the same region, and a preference for intelligible speech compared to unintelligible speech sounds, more ventrally, within the STS. In monkeys, however, no STG or STS regions appeared to be specialized for the processing of monkey calls, providing a negative answer to our first experimental question ("would we observe species-specific response in each species?"). Species-specific high-level regions were detected in humans but not in monkeys. With respect to the second question ("At what level monkeys calls are processed"), the regions processing monkey vocalizations in monkeys (parabelt extending into STG) matched the areas activated by unintelligible speech and emotional utterances in humans (the STG extending into the upper bank of STS).

Early processing in humans and monkeys

Even the meaningless, scrambled stimuli, compared to silence, activated more cortex than the expected auditory regions in the temporal lobes, evoking additional activations in the parietal and frontal areas of both species. This is not entirely unexpected, as auditory single-cell responses have been recorded in monkey ventral premotor (Kohler et al., 2002), parietal (Grunewald et al., 1999; Mazzoni et al., 1996) and insular regions (Remedios et al., 2009). Furthermore previous imaging and metabolic studies also reported activations by auditory stimuli outside the auditory system (Joly et al., 2012; Poremba et al., 2003)

Only small regions close to the primary auditory cortex were sensitive to differences in the various scrambled conditions (Figs. 2 and 3). It is likely that these regions simply encode spectral information, the main feature differentiating the various categories of scrambled stimuli. At the early auditory processing level, monkey calls in monkeys and human vocalizations (speech or not) in humans dominated the responses (see Fig. 3). Since these regions responded similarly to scrambled controls, it indicates a stronger response when the spectral content of the stimuli matches the spectral content of the vocalizations produced by conspecifics.

Higher-order auditory processing of primate vocalizations

In both species, the combined human and monkey vocalizations compared to their scrambled controls elicited bilateral activations in the STG, with maxima lateral to the peak of the main auditory activation (Fig. 4). In monkeys, more focused comparisons contrasting monkey calls, French and emotional utterances to their corresponding scrambled stimuli elicited patterns of activation that were relatively similar to one another (Figs. 5A, 6A, 7A). Poremba et al. (2004) also recorded responses in monkey STG to monkey and human vocalizations and to scrambled monkey calls, but in this PET study only a single condition was tested in a given session and conditions were not directly compared. In their study, the authors reported a left–right asymmetric response in the temporal pole for the processing of conspecifics. They described a very similar profile along the STG across conditions (human and monkey vocalizations) with a peak at the level of the primary auditory cortex. Hence, this last finding is compatible with our finding in that human and monkey vocalizations can activate similarly the monkey STG. On the other hand, in humans, unlike in monkeys, the speech conditions (French and Arabic) differed more sharply from their scrambled counterparts than did the non-speech conditions (MVoc and Hemo) (Figs. 5B, 6B, 7B and C). In both species, the strongest

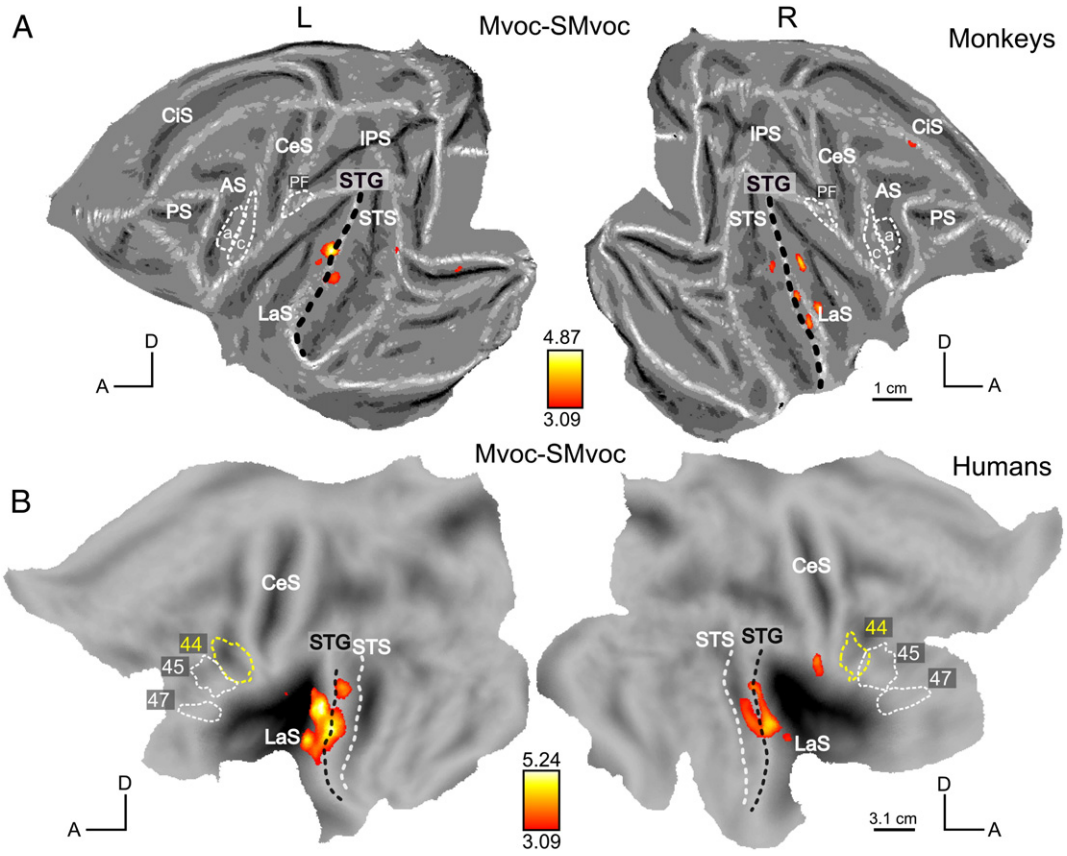

Fig. 5. Scrambling effect for monkey vocalizations. SPM T-maps ($p<0.001$ uncorrected) for the contrast Mvoc–SMvoc in monkeys (A) and in humans (B).

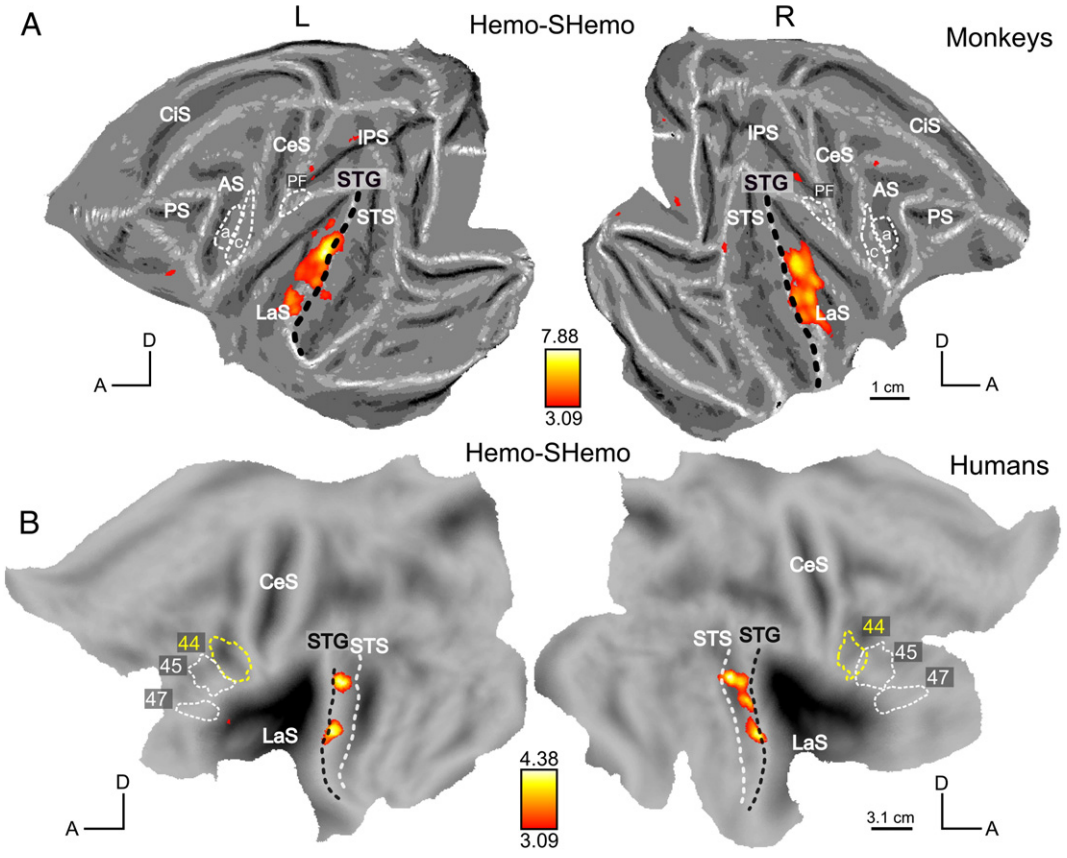

Fig. 6. Scrambling effect for human emotional sounds. SPM T-maps ($p<0.001$ uncorrected) for the contrast Hemo–SHemo in monkeys (A) and in humans (B).

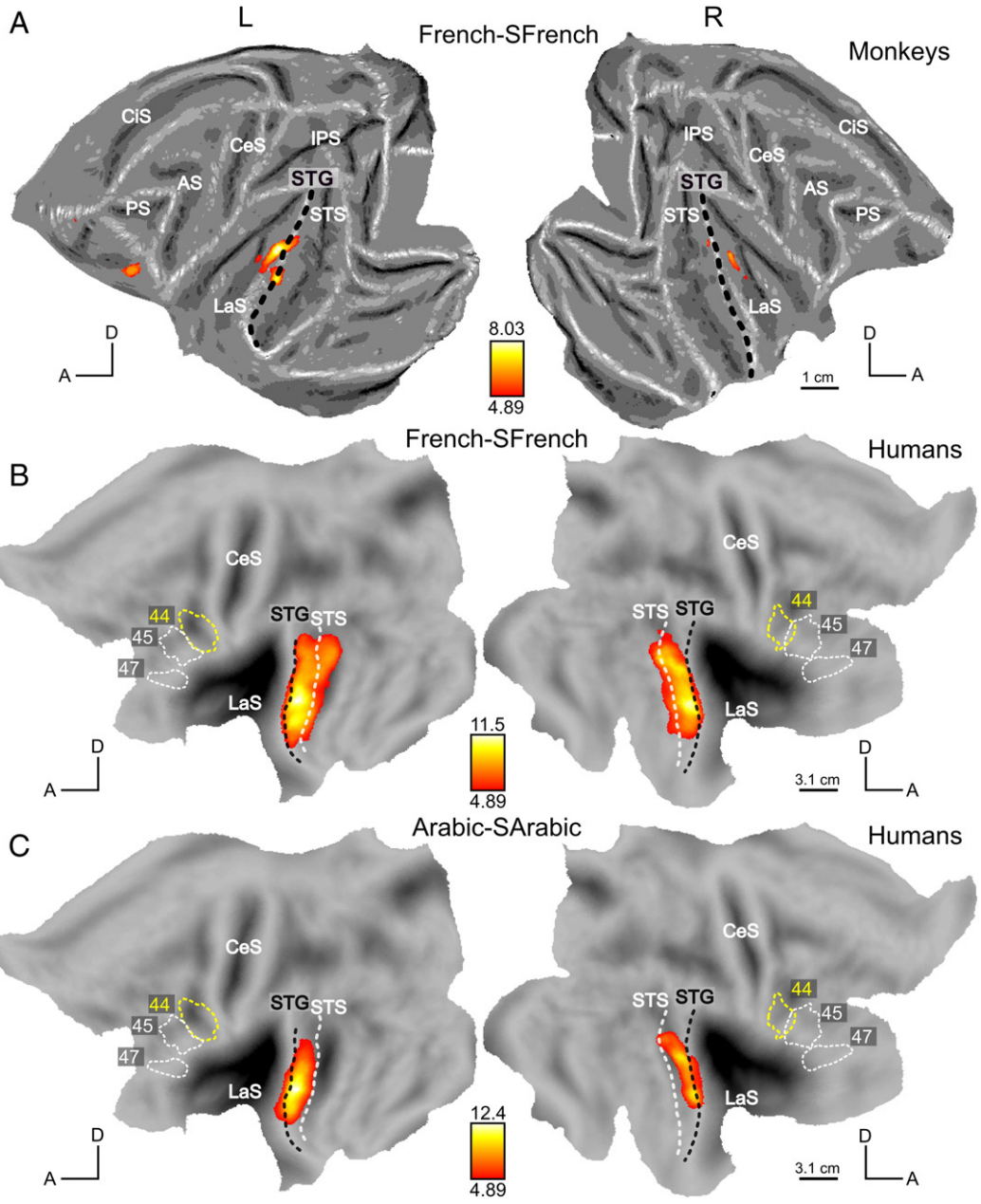

Fig. 7. Scrambling effect for human speech. SPM T-maps ($p<0.05$ FWE corrected) for the contrast French–SFrench in monkeys (A) and in humans (B). SPM T-map for Arabic–SArabic in humans (C).

activation was observed by contrasting French to its scrambled control. While in humans this matches the preference for human speech at the early auditory level, it does not match the relative dominance of conspecific calls at the early level in monkeys.

In monkeys, we observed little difference between monkey calls and their scrambled counterparts. This negative result is unlikely to result from a lack of power, insofar as MR signals obtained in the monkey, because of the use of a contrast agent, are about the same order of magnitude as BOLD signals in humans (Denys et al., 2004). It must be noted that the effect of scrambling on the monkey calls was shown to be slightly less disruptive than its effect on the French stimuli (Joly et al., 2012). This could explain the rather strong response to scrambled monkey calls. However, Fig. 3 shows that the effect of scrambling is greater in humans than in monkeys, independently of the category. In contrast, parts of human STG showed strong responses to human speech relative to their scrambled counterparts, confirming earlier reports by Fecteau et al. (2004) and Leaver and Rauschecker (2010).

In monkeys, the clusters of voxels responding to vocalizations more than to their scrambled controls were located in the auditory lateral belt and parabelt (Kaas and Hackett, 2000). Concerning humans, two rather different architectonic schematics have been proposed for the superior temporal region. On one hand, Sweet et al. (2005) reported an auditory core region lying within the posteromedial two-thirds of Heschl's gyrus, a lateral belt located predominantly in the anterior and posterior banks of Heschl's sulcus, and a parabelt region localized largely to the planum temporale. On the other hand, Fullerton and Pandya (2007) have suggested that the core area includes not only Heschl's gyrus but also areas rostral and caudal to it, and that the lateral belt is located lateral to the core line and extends over the lateral edge of the STG. Here, human activations for intact vocalizations compared to their scrambled control stimuli peaked close to the lateral edge of the STG, while very little was found in or around Heschl's sulcus. If one assumes that the effect observed in the monkeys' lateral belt is “equivalent” to the effect

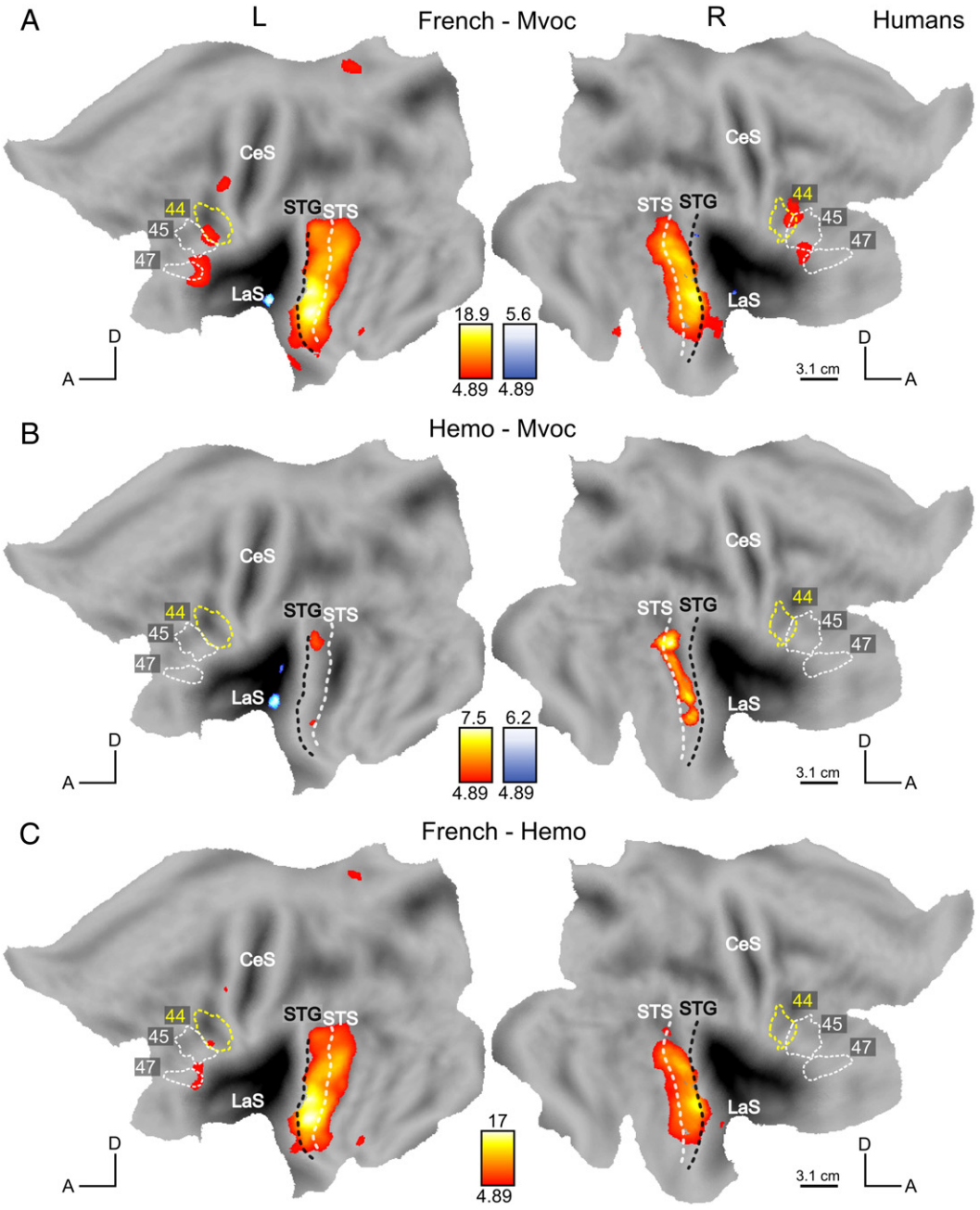

Fig. 8. Comparison between sound categories in humans. SPM T-maps for the contrast French–Mvoc (A), Hemo–Mvoc (B) and French–Hemo (C) at $p<0.05$ FWE-corrected, masked exclusively with *effect of scrambling* ($p<0.05$ uncorrected) in a red (positive)–blue (negative) color map.

observed in humans, our data are in agreement with the description of Fullerton and Pandya (2007).

Even if the present results suggest that the lateral parabelt in monkeys corresponds in humans to the STG plus the dorsal bank of STS, these regions are differently engaged in the two species by the three types of stimuli we tested. In the monkey, the emotional utterances compared to their scrambled counterparts tended to evoke more response than monkey calls compared to their scrambled controls, while the opposite was observed in humans. In monkeys, the larger response to emotional utterances, relative to their scrambled counterparts, compared to monkey calls, reflected the greater spatio-temporal complexity of the emotional utterances (Joly et al., 2012). In humans the opposite tendency, however may reflect differences in variability of the scrambled conditions, as emotional utterances had similar effects as monkey calls when compared to conditions other than scrambled controls, e.g. French (Fig. 8). Despite the stronger relative response to emotional utterances in monkeys, it proved not to be the case that the processing of these utterances in monkeys was more similar to that of monkey calls than that of human speech. Hence, we found little grounds to support the view that calls have a predominantly emotional content in monkeys.

A previous monkey fMRI study (Petkov et al., 2008) has reported voice-preferring regions in the lateral sulcus (including A1 and anterior regions). The authors have shown fMRI adaptation to the caller in their most anterior cluster, but the design did not allow them to compare monkey calls and human vocalizations in order to assess species-specificity. In this anterior voice region, the first evidence for "voice" cells was reported recently (Perrodin et al., 2011) but only a modest proportion of such voice cells were found. This latest finding could partly explain our difficulty in detecting these regions using fMRI.

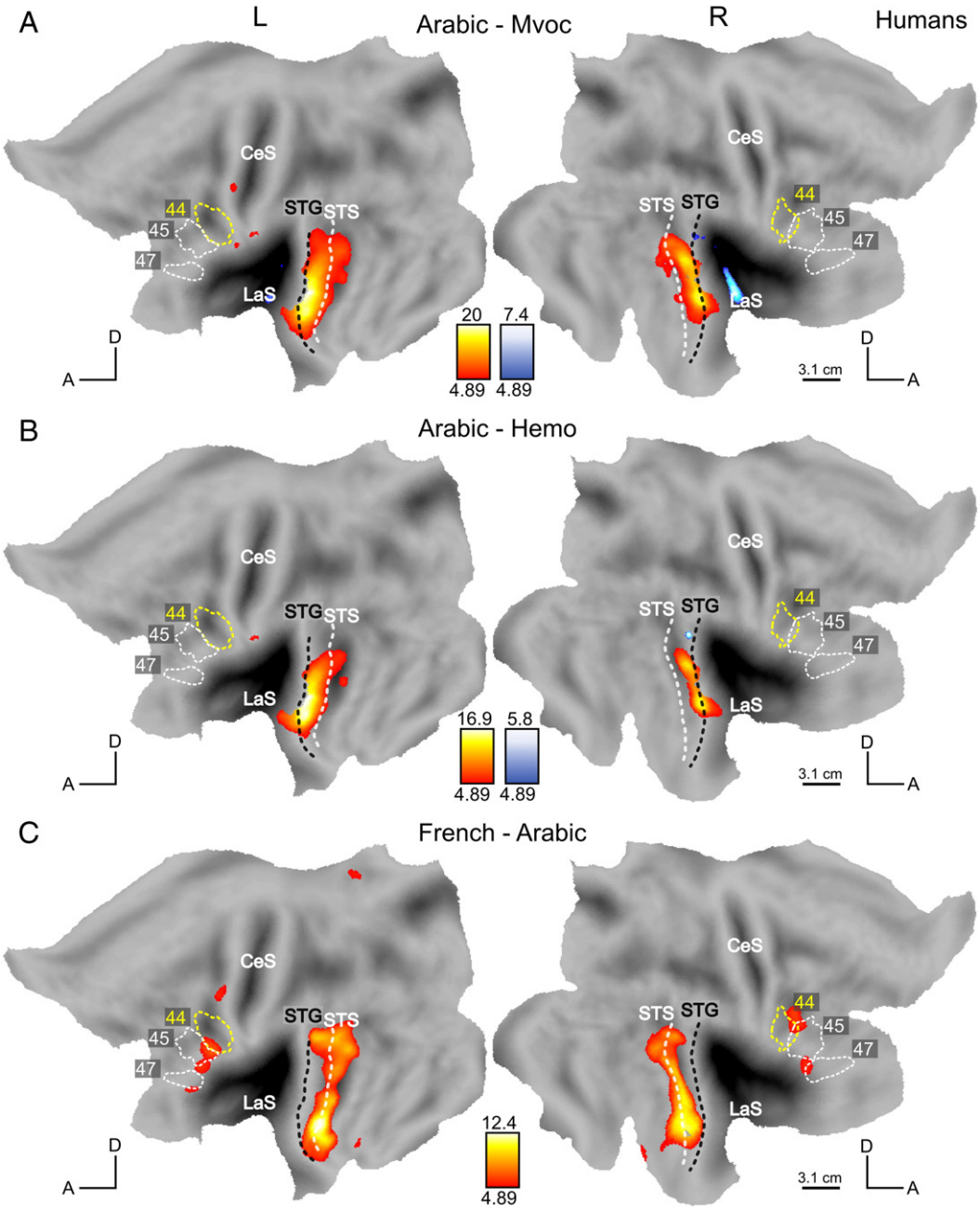

Fig. 9. Processing of speech and intelligibility in humans. SPM T-map ($p<0.05$ FWE corrected) for contrast Arabic–Mvoc (A), Arabic–Hemo (B) and French–Arabic (C) masked exclusively with *effect of scrambling* ($p<0.05$ uncorrected) and shown in a red (positive)–blue (negative) color map.

Activations by vocalizations in the frontal cortex

In monkeys, an additional region sensitive to intact vocalizations was detected in the orbito-frontal cortex (OFC, area 12o). This is consistent with the observation of Romanski et al. (1999) and Romanski and Averbeck (2009) that the middle and anterior lateral belt areas, where maximum activations were found for intact stimuli, project directly to area 12o. The activation in OFC is driven mainly by the contrast French–SFrench (Fig. 7A) which is also the strongest contrast in the lateral belt. Romanski and Goldman-Rakic (2002) recorded neuronal responses to monkey and human vocalizations in this region, but did not mention any difference between human and monkey vocalization, nor between left or right hemisphere in this brief report. Neurons in this region are selective for monkey calls and on average respond to 2 to 5 different calls. This selectivity reflects the acoustic features of the calls rather than their functional meaning (Romanski et al., 2005). In a subsequent report Sugihara et al. (2006) reported a convergence of auditory vocalizations and facial gestures onto single neurons of this region. Finally, Tsao et al. (2008a) reported face patches bilaterally in this region. Whether this orbito-frontal auditory activation underlies the processing of vocalization value or contributes to decision-making (Wallis, 2007), or plays a more general role in inter-subject communication (Sugihara et al., 2006), remains to be investigated. Although single-cell responses to monkey calls have previously been reported (Romanski and Goldman-Rakic, 2002) in other ventrolateral prefrontal areas (12vl/47 and 45), we did not observe a preference for primate vocalizations relative to their scrambled control in these areas.

We did not observe any similar activation in human OFC. However, ventral prefrontal cortex in humans was activated by the scrambled

vocalizations, and also when French was compared to monkey vocalizations or to human emotional utterances. The absence of OFC activation reflects a general property of human prefrontal cortex which suppresses direct sensory inputs. This has been well-documented for static visual stimuli such as objects (Denys et al., 2004), and also for dynamic visual stimuli such as visual actions (Jastorff et al., 2010; Nelissen et al., 2005). The present study extends this observation to the auditory modality.

The differential role of the STS in humans and monkeys

Perhaps the most striking result of our study comparing humans with monkeys is the absence of any specific response to monkey calls in the STS of monkeys. This region, crucially involved during language processing in humans, did not respond significantly to monkey vocalizations in the monkeys themselves. This result is in agreement with the study from Petkov et al. (2008) reporting no response to vocalization in the STS. The dorsal bank of the STS, the temporal parieto-occipital (TPO) area, (also called superior temporal polysensory area, STP), is known to display robust responses to faces and multimodal stimuli (Ghazanfar et al., 2008). Hence in monkeys, multimodal stimulation might be necessary to activate the upper bank of the STS. In humans, by way of contrast, complex auditory stimuli conveying speech are sufficient to activate most of the STS.

In humans, intelligible speech, compared to unintelligible speech, yielded activations extending all along the STS bilaterally; and areas in the left inferior frontal gyrus and left precentral gyrus were also recruited. These results are typical of studies comparing a known language versus an unknown one (Mazoyer et al., 1993; Pallier et al., 2003; Papathanassiou et al., 2000), and also fit with studies comparing intelligible versus unintelligible acoustic controls (e.g. Davis and Johnsrude, 2003). The left-dominance often reported in language studies is less obvious in the present investigation, especially for the contrast comparing human speech to its scrambled counterpart, which in the monkey yielded the clearest left dominance (Joly et al., 2012). This may be due to the fact that our participants listened to relatively simple, short verbal utterances and did not have to perform any explicit task (Hickok and Poeppel, 2007). Interestingly, the left, but not the right, precentral cortex was activated by intelligible speech. Nearby sites on the left have in fact been shown to be activated by speech perception and production (−54, −4, 48, Meister et al., 2007) and have been implicated in the integration of speech and action (−52, −6, 49, Willems et al., 2007).

Limitations of this study

When comparing activations in the two species, we must keep in mind the differences in scanning procedures and analysis, in that different voxel sizes, MR signals, smoothing factors, and numbers of subjects were tested. Although, the SNR is lower in monkeys than in humans (smaller voxels in monkeys), the statistical power of the analysis is relatively similar, as demonstrated by comparable t-scores in SPMs visualizing the processing at the first two levels (Figs. 2 and 4). For both species, the stimuli were generated by concatenating several short segments, but only two segments were concatenated in the stimuli presented to monkeys, while sequences of about five segments were presented to humans. Care was taken, however, to concatenate segments of the same valence for monkey calls and emotional utterances. It is not obvious how such a difference in the number of segments thus concatenated might explain a difference at one level of processing and not at the other. It could be argued that monkeys might be relatively impaired in their recognition of the calls when these are concatenated than are humans when speech segments from different speakers are concatenated, but there is little empirical evidence to support such a notion. In the same vein, it could be that monkeys, more so than humans, were influenced by the fact that none of the callers were familiar, but again this is conjecture. However, the use of unfamiliar callers may explain the lack of activation in the anterior voice region described earlier (Petkov et al., 2008), as this region might be optimally stimulated when the various voices are familiar. Furthermore, it may be that for monkeys the scanner, although familiar, remains an unnatural context in which all complex sounds have a similar meaning. The human voices may be more familiar or expected because of the interactions with researchers before and after the scanning, and thereby have more contextual saliency than the monkey calls. Although we cannot exclude this possibility, is seems not very attractive as the same reasoning should apply to complex visual stimuli, such as bodies. Yet in a recent comparative fMRI study we showed that monkey body patches respond slightly more to monkey than to human bodies (Jastorff et al., 2012). Of course it remains possible that contextual saliency has more effect on auditory than visual stimuli.

Hence, it appears that the major difference observed in the categorical level for the processing of vocalizations reflects a genuine species difference. The available evidence suggests that, in monkeys, calls are simply processed by the higher-order auditory system and dispatched, without much further specific processing, to the voice region and orbito-frontal cortex. In that respect, the auditory system may differ from the visual system, in that regions specifically involved with visual categories such as faces have been documented in both humans and monkeys (Tsao et al., 2008b). However, most of the studies searching for face-processing regions did not investigate the species-specificity of these regions, which may be a closer equivalent to the categorical level assessed in our study.

Finally, one must be cautious when interpreting the lack of species-specific auditory processing in monkeys within the context of a passive fixation task, in that the discrimination of the categories was not assessed in a behavioral task. The effects of an auditory task on the activities of single units along the central auditory pathway has been documented (Otazu et al., 2009; Ryan et al., 1984) and, even though behavior might have a moderate effect on the core auditory cortex in macaques (Scott et al., 2007), it remains difficult to predict the effect of passive and active listening on the responses in belt, parabelt and other cortical regions. Furthermore, despite the use of an identical passive fixation task, attention or memory may nonetheless have been involved quite differently during passive fixation in monkeys and humans. Yet our data showed that, in the monkeys, both conspecific calls and human vocal sounds (whether speech or not), evoked similar MR responses, mostly in the STG. In contrast, a clear preference for human vocalizations was observed in human STG and in the STS. The STS was especially responsive to intelligible utterances. The evolution of the language faculty in humans thus seems to have recruited most of the STS for processing language, thereby displacing cortex involved in biological motion and in visual action processing into the posterior MTG/STS and posterior OTS (Jastorff et al., 2012). It may be the case in monkeys that a much simpler repertoire of vocalizations did not require any significant involvement of this superior temporal region.

Acknowledgments

The help of C. Giffard, P. Kayenbergh, G. Meulemans, M. Depaep, C. Fransen, A. Coeman, M. Hauser, A.-D. Devauchelle and E. Dupoux is kindly acknowledged. The authors are indebted to M. D. Hauser for the recordings of monkey calls, to G. Luppino for help with the definition of the ROIs and to S. Raiguel for comments on earlier versions of the manuscript. The study was supported by a Marie Curie Early Stage Research Training Fellowship (MEST-CT-2004-007825 to OJ), Neurocom (NEST 012738 to GAO), EF 05/14, and FWO G 151.04. Sinerem was kindly provided by Guerbet (Roissy, France). The experiment was part of a general research program on functional neuroimaging of the brain sponsored by the Atomic Energy Commission and approved by the regional ethical committee (Comité de Protection des Personnes, Hopital de Bicetre).

References

Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H.B., Zilles, K., 1999. Broca's region revisited: cytoarchitecture and intersubject variability. J. Comp. Neurol. 412, 319–341.

Amunts, K., Weiss, P.H., Mohlberg, H., Pieperhoff, P., Eickhoff, S., Gurd, J.M., Marshall, J.C., Shah, N.J., Fink, G.R., Zilles, K., 2004. Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotaxic space—the roles of Brodmann areas 44 and 45. Neuroimage 22, 42–56.

Avants, B., Epstein, C., Grossman, M., Gee, J., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41 (Special Issue on The Third International Workshop on Biomedical Image Registration—WBIR 2006).

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70, 614–636.

Baumgart, F., Kaulisch, T., Tempelmann, C., Gaschler-Markefski, B., Tegeler, C., Schindler, F., Stiller, D., Scheich, H., 1998. Electrodynamic headphones and woofers for application in magnetic resonance imaging scanners. Med. Phys. 25, 2068–2070.

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

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

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

Belmalih, A., Borra, E., Contini, M., Gerbella, M., Rozzi, S., Luppino, G., 2009. Multimodal architectonic subdivision of the rostral part (area f5) of the macaque ventral premotor cortex. J. Comp. Neurol. 512, 183–217.

Brown, C.H., Sinnott, J.M., 2006. Cross-species Comparisons of Vocal Perception. New Jersey.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.

Denys, K., Vanduffel, W., Fize, D., Nelissen, K., Sawamura, H., Georgieva, S., Vogels, R., Essen, D.V., Orban, G.A., 2004. Visual activation in prefrontal cortex is stronger in monkeys than in humans. J. Cogn. Neurosci. 16, 1505–1516.

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

Fitch, W.T., Fritz, J.B., 2006. Rhesus macaques spontaneously perceive formants in conspecific vocalizations. J. Acoust. Soc. Am. 120, 2132–2141.

Formisano, E., Kim, D.S., Di Salle, F., van de Moortele, P.F., Ugurbil, K., Goebel, R., 2003. Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40, 859–869.

Fullerton, B.C., Pandya, D.N., 2007. Architectonic analysis of the auditory-related areas of the superior temporal region in human brain. J. Comp. Neurol. 504, 470–498.

Ghazanfar, A.A., Turesson, H.K., Maier, J.X., van Dinther, R., Patterson, R.D., Logothetis, N.K., 2007. Vocal-tract resonances as indexical cues in rhesus monkeys. Curr. Biol. 17, 425–430.

Ghazanfar, A.A., Chandrasekaran, C., Logothetis, N.K., 2008. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J. Neurosci. 28, 4457–4469.

Gil-da-Costa, R., Braun, A., Lopes, M., Hauser, M.D., Carson, R.E., Herscovitch, P., Martin, A., 2004. Toward an evolutionary perspective on conceptual representation: species-specific calls activate visual and affective processing systems in the macaque. Proc. Natl. Acad. Sci. U.S.A. 101, 17516–17521.

Gil-da-Costa, R., Martin, A., Lopes, M.A., Muñoz, M., Fritz, J.B., Braun, A.R., 2006. Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. Nat. Neurosci. 9, 1064–1070.

Gouzoules, S., Gouzoules, H., Marler, P., 1984. Rhesus monkey (*Macaca mulatta*) screams: representational signalling in the recruitment of agonistic aid. Anim. Behav. 32, 182–193.

Grunewald, A., Linden, J.F., Andersen, R.A., 1999. Responses to auditory stimuli in macaque lateral intraparietal area. I. Effects of training. J. Neurophysiol. 82, 330–342.

Hauser, M., Marler, P., 1993. Food-associated calls in rhesus macaques (*Macaca õmulatta*): I. Socioecological factors. Behav. Ecol. 4, 194–205.

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve? Science 298, 1569–1579.

Henson, R., Penny, W., 2003. ANOVAs and SPM. Technical Report Wellcome Department of Imaging Neuroscience.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., Orban, G.A., 2010. Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. J. Neurophysiol. 104, 128–140.

Jastorff, J., Popivanov, I.D., Vogels, R., Vanduffel, W., Orban, G.A., 2012. Integration of shape and motion cues in biological motion processing in the monkey sts. Neuroimage 60 (2), 911–921 (Apr 2, Epub 2012 Jan 10).

Joly, O., Ramus, F., Pressnitzer, D., Vanduffel, W., Orban, G.A., 2012. Interhemispheric differences in auditory processing revealed by fMRI in awake rhesus monkeys. Cereb. Cortex 22 (4), 838–853 (Apr, Epub 2011 Jun 27).

Kaas, J.H., Hackett, T.A., 2000. Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. U.S.A. 97, 11793–11799.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V., Rizzolatti, G., 2002. Hearing sounds, understanding actions: action representation in mirror neurons. Science 297, 846–848.

Langers, D.R.M., van Dijk, P., 2011. Mapping the tonotopic organization in human auditory cortex with minimally salient acoustic stimulation. Cereb. Cortex (Oct 6, Epub ahead of print).

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. 30, 7604–7612.

Leite, F.P., Tsao, D., Vanduffel, W., Fize, D., Sasaki, Y., Wald, L.L., Dale, A.M., Kwong, K.K., Orban, G.A., Rosen, B.R., Tootell, R.B.H., Mandeville, J.B., 2002. Repeated fMRI using iron oxide contrast agent in awake, behaving macaques at 3 Tesla. Neuroimage 16, 283–294.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

Mazoyer, B., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., Mehler, J., 1993. The cortical representation of speech. J. Cogn. Neurosci. 5, 467–479.

Mazzoni, P., Bracewell, R.M., Barash, S., Andersen, R.A., 1996. Spatially tuned auditory responses in area lip of macaques performing delayed memory saccades to acoustic targets. J. Neurophysiol. 75, 1233–1241.

McLaren, D.G., Kosmatka, K.J., Oakes, T.R., Kroenke, C.D., Kohama, S.G., Matochik, J.A., Ingram, D.K., Johnson, S.C., 2009. A population-average MRI-based atlas collection of the rhesus macaque. Neuroimage 45, 52–59.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Curr. Biol. 17, 1692–1696.

Meyer, M., Zysset, S., von Cramon, D.Y., Alter, K., 2005. Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. Brain Res. Cogn. Brain Res. 24, 291–306.

Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb. Cortex 13, 1362–1368.

Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.A., 2005. Observing others: multiple action representation in the frontal lobe. Science 310, 332–336.

Otazu, G.H., Tai, L.H., Yang, Y., Zador, A.M., 2009. Engaging in an auditory task suppresses responses in auditory cortex. Nat. Neurosci. 12, 646–654.

Pallier, C., Dehaene, S., Poline, J.B., LeBihan, D., Argenti, A.M., Dupoux, E., Mehler, J., 2003. Brain imaging of language plasticity in adopted adults: can a second language replace the first? Cereb. Cortex 13, 155–161.

Papathanassiou, D., Etard, O., Mellet, E., Zago, L., Mazoyer, B., Tzourio-Mazoyer, N., 2000. A common language network for comprehension and production: a contribution to the definition of language epicenters with PET. Neuroimage 11, 347–357.

Patterson, R.D., Allerhand, M.H., Giguere, C., 1995. Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. J. Acoust. Soc. Am. 98, 1890–1894.

Perrodin, C., Kayser, C., Logothetis, N.K., Petkov, C.I., 2011. Voice cells in the primate temporal lobe. Curr. Biol. 21, 1408–1415.

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

Poremba, A., Saunders, R.C., Crane, A.M., Cook, M., Sokoloff, L., Mishkin, M., 2003. Functional mapping of the primate auditory system. Science 299 (5606), 568–572.

Poremba, A., Malloy, M., Saunders, R.C., Carson, R.E., Herscovitch, P., Mishkin, M., 2004. Species-specific calls evoke asymmetric activity in the monkey's temporal poles. Nature 29 (427(6973)), 448–451.

Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268, 111–114.

Remedios, R., Logothetis, N.K., Kayser, C., 2009. An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. J. Neurosci. 29, 1034–1045.

Romanski, L.M., 2004. Domain specificity in the primate prefrontal cortex. Cogn. Affect. Behav. Neurosci. 4, 421–429.

Romanski, L.M., Averbeck, B.B., 2009. The primate cortical auditory system and neural representation of conspecific vocalizations. Annu. Rev. Neurosci. 32, 315–346.

Romanski, L.M., Goldman-Rakic, P.S., 2002. An auditory domain in primate prefrontal cortex. Nat. Neurosci. 5, 15–16.

Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., Rauschecker, J.P., 1999. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat. Neurosci. 2, 1131–1136.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J. Neurophysiol. 93 (2), 734–747.

Ruggero, M.A., Temchin, A.N., 2005. Unexceptional sharpness of frequency tuning in the human cochlea. Proc. Natl. Acad. Sci. U.S.A. 102, 18614–18619.

Ryan, A.F., Miller, J.M., Pfingst, B.E., Martin, G.K., 1984. Effects of reaction time performance on single-unit activity in the central auditory pathway of the rhesus macaque. J. Neurosci. 4, 298–308.

Saleem, K., Logothetis, N., 2006. A Combined MRI and Histology Atlas of the Rhesus Monkey Brain. Academic Press.

Scott, B.H., Malone, B.J., Semple, M.N., 2007. Effect of behavioral context on representation of a spatial cue in core auditory cortex of awake macaques. J. Neurosci. 27, 6489–6499.

Serafin, J.V., Moody, D.B., Stebbins, W.C., 1982. Frequency selectivity of the monkey's auditory system: psychophysical tuning curves. J. Acoust. Soc. Am. 71, 1513–1518.

Seyfarth, R.M., Cheney, D.L., Marler, P., 1980. Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. Science 210, 801–803.

Sugihara, T., Diltz, M.D., Averbeck, B.B., Romanski, L.M., 2006. Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J. Neurosci. 26 (43), 11138–11147.

Sweet, R.A., Dorph-Petersen, K.A., Lewis, D.A., 2005. Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. J. Comp. Neurol. 491, 270–289.

Tsao, D.Y., Schweers, N., Moeller, S., Freiwald, W.A., 2008a. Patches of face-selective cortex in the macaque frontal lobe. Nat. Neurosci. 11 (8), 877–879.

Tsao, D.Y., Moeller, S., Freiwald, W.A., 2008b. Comparing face patch systems in macaques and humans. Proc. Natl. Acad. Sci. U.S.A. 105, 19514–19519.
Van Essen, D.C., 2005. A population-average, landmark- and surface-based (pals) atlas of human cerebral cortex. Neuroimage 28, 635–662.
Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. J. Am. Med. Inform. Assoc. 8, 443–459.
Vanduffel, W., Fize, D., Mandeville, J.B., Nelissen, K., Hecke, P.V., Rosen, B.R., Tootell, R.B., Orban, G.A., 2001. Visual motion processing investigated using contrast agent-enhanced fMRI in awake behaving monkeys. Neuron 32, 565–577.
Wallis, J.D., 2007. Orbitofrontal cortex and its contribution to decision-making. Annu. Rev. Neurosci. 30, 31–56.
Willems, R.M., Özürek, A., Hagoort, P., 2007. When language meets action: the neural integration of gesture and speech. Cereb. Cortex 17, 2322–2333.