Sensitivity of the human temporal voice areas to nonhuman primate vocalizations

Leonardo Ceravolo[1*], Coralie Debracque[1*], Thibaud Gruber[1‡] & Didier Grandjean[1‡]

[*]shared first authors

[‡]shared last authors

[1]University of Geneva, Geneva, Switzerland

Corresponding author:

Leonardo Ceravolo, PhD

University of Geneva, Unimail

Boulevard Pont-d'Arve 40

CH-1205 Geneva, Switzerland

e-mail: leonardo.ceravolo@unige.ch

orcid: https://orcid.org/0000-0003-0638-3981

Abstract

In recent years, research on voice processing in the human brain—particularly the study of temporal voice areas (TVA)—was dedicated almost exclusively to conspecific vocalizations. To characterize commonalities and differences regarding primate vocalization representations in the human brain, the inclusion of closely related nonhuman primates—namely chimpanzees and bonobos—is needed. We hypothesized that neural commonalities would depend on both phylogenetic and acoustic proximities, with chimpanzees ranking closest to *Homo*. Presenting human participants (N=23) with the vocalizations of four primate species (rhesus macaques, chimpanzees, bonobos and humans) and regressing-out relevant acoustic parameters using three distinct analyses, we observed within-TVA, sample-specific, bilateral anterior superior temporal gyrus activity for chimpanzee vocalizations compared to: all other species; nonhuman primates; human vocalizations. Within-TVA activity was also observed for macaque vocalizations. Our results provide evidence for subregions of the TVA that respond principally—but not exclusively—to phylogenetically and acoustically close nonhuman primate vocalizations, namely those of chimpanzees.

Introduction

The study of cerebral mechanisms underlying speech and voice processing has gained importance since the early 2000s with the advent of functional magnetic resonance imaging (fMRI) [1]. Voice-sensitive areas, commonly referred to as “temporal voice areas” (TVA) or simply “voice areas”, have been highlighted along the upper, superior portion of the temporal cortex [2]. Since then, great efforts have been made to better characterize these TVA, with particular attention to their spatial division into functional subregions [3-5]. A fairly large body of literature points to the critical role of the TVA in voice perception and processing in healthy participants [4, 6-8] as well as in lesioned patients [9]. Subregions of the TVA have also been directly linked to social perception [10], vocal emotion processing [11, 12], voice identity [13, 14], and gender perception [15]. The developmental axis of voice processing has also been studied in infants, demonstrating the existence of the TVA in the human brain as early as 7—but not 4—months of age [16], while the ability to respond specifically to the voice of their parents has been observed in fetuses in utero [17]. With the ongoing development of brain imaging and analysis techniques [18], it is realistic to expect successful, albeit noninvasive, fMRI results on task-related voice perception in utero in the near future. Along the evolutionary axis, evidence for TVA or, more generally, conspecific vocalization-sensitive brain areas has emerged primarily in dogs [19] and monkeys [20, 21] (*Macaca mulatta*), raising the question of whether such specialized brain areas are species-specific [22] and to what extent human and nonhuman primates share neural mechanisms that enable them to preferentially process conspecific vocalizations [23]. However, less attention has been paid to paradigms in which animal vocalizations are presented to humans, and to the best of our knowledge no study to date has reported selective human TVA activations for processing such auditory material, namely the vocalizations of other animals. Human processing of animal vocalizations has been studied with both monkey and cat material, but no *specific* cross-species activations have been observed within the TVA with respect to either species [24]. Other studies have focused more specifically on phylogenetic distance and have included nonhuman ape (chimpanzee, *Pan*

troglodytes) and 'Old World' monkey (rhesus macaque, *Macaca mulatta*) vocalizations as stimuli. Such studies failed to identify species-specific brain activations—despite correctly discriminating chimpanzee affective vocalizations [25]—and observed ambivalent results for below [25] vs. above [26] chance discrimination of macaque affective vocalizations by human participants. A recent exception is a study in which functionally homologous anterior TVA activity was observed in both humans and macaques: this region was indeed specific to macaque calls in the macaque's anterior TVA, and specific to human voices in the anterior TVA of humans, but no macaque-specific activity was observed in the human TVA [27]. This sparse literature motivated the present study, which aims to investigate cross-species TVA activations in humans when asked to categorize vocalizations from phylogenetically—and acoustically-close and -distant—species while undergoing fMRI scanning. The importance of acoustic differences between species and more specifically acoustic distance, particularly through fundamental frequency variations [28, 29] was indeed of great interest. Acoustic distance—calculated using Mahalanobis distance with 16 acoustic parameters extracted from the stimuli—was in fact a determining parameter in assessing affective cues recognition in nonhuman primate calls by human participants [30]. In this study, affiliative chimpanzee—but not bonobo—calls were acoustically the closest to positive human voice stimuli, suggesting a distinct evolution of bonobo calls [30]. Bonobo vocalizations are of particular interest because this species is thought to have undergone evolutionary changes in their communication, in part due to a neoteny process involving acoustic modifications, and although they are as phylogenetically close to humans as chimpanzees—with an estimated separation with the *Homo* lineage only 6-8 million years ago [31]. Previous research has shown that bonobos have a shorter larynx—a valid predictor of a species' mean fundamental frequency [32]—compared to chimpanzees, resulting in a higher fundamental frequency in their calls [28]. Such a difference has been demonstrated in juvenile bonobo calls compared to chimpanzee and human baby calls [33], arguing for a greater acoustic distance between bonobo calls and human or chimpanzee vocalizations. For these reasons, we included vocalizations from both Pan species (chimpanzees, *Pan troglodytes*; bonobos, *Pan paniscus*), as well as a

phylogenetically more distant species (*Cercopithecidae*: rhesus monkeys), with an estimated separation with the *Homo* lineage dating back to 25 million years ago. Indeed, any claim of human 'uniqueness' for TVA recruitment remains on hold and should be tested in light of these closely related species. Using the same stimuli, we previously investigated the specific frontal mechanisms involved in the categorization of nonhuman primate vocalizations independently of a selection of low-level acoustic parameters [34], but the possibility that acoustic differences would affect, at the auditory level, the ability of human participants to recognize nonhuman primate calls should be thoroughly examined, as we did in the present study. As suggested by research mentioned above, monkey vocalizations are overall less likely to be identified compared to ape vocalizations due to both phylogenetic and acoustic differences. Therefore, our mechanistic hypothesis of the difficulty for humans to recognize bonobo calls is that frequencies of the human tonotopic map in the auditory cortex—adapted and adjusted to the frequencies of the human voice during evolution—would not be tailored to process the frequencies generated by bonobo calls. It would also be the case for macaque calls, while frequencies of chimpanzee calls—being closer to the range of human voice fundamental frequency [28, 30]—would be better represented in the human auditory cortex and therefore more easily processed and better identified by humans.

According to the literature mentioned so far and to the mechanistic hypothesis underlying the processing of chimpanzee as opposed to bonobo and/or macaque calls by human participants, we therefore predicted: (i) more acoustic proximity between human and chimpanzee vocalizations, whereas more distance would separate those of bonobos and macaques from the human voice; (ii) a recruitment of temporal brain areas—within the TVA—for the processing of vocalizations from the Pan taxon (chimpanzee, bonobo) but not *Cercopithecidae* (rhesus monkey) vocalizations, taking into account acoustic features of interest through a discriminant analysis of the parameters that best underlie our stimuli.

Results

Our hypotheses involve a systematic and thorough control of phylogeny through the inclusion of specific primate species as well as the selection of specific acoustic features. We programmed a task in which the vocalizations of each species were presented randomly and for which the participants (N=23) had to specify to which species each stimulus corresponded. We therefore included equal numbers of trials (N=72) with human, chimpanzee, bonobo and macaque vocalizations—N=18 each—as well as trial-level acoustic features of the vocalizations, using three distinct statistical models with specific covariates. These models are sorted from the least to the most sophisticated modeling to uncover the role(s) of acoustic features on TVA activity potentially specific to each/some species (see the Methods; Model 1: mean of vocalization fundamental frequency and energy; Model 2: multi-dimensional Mahalanobis acoustic distance between the human voice and the calls of each nonhuman primate species [35]; Model 3: between-species most discriminant acoustic features of our stimuli, extracted using a general discriminant analysis [30]). Acoustical analyses involved in Model 2 allowed us to validate our first hypothesis according to which chimpanzees are acoustically the closest to humans, followed by the calls of bonobos and macaques (Fig.1**B**)—the main effect of Species on the acoustic distance was significant, $F(3,88)=15.84$, $p<.001$, as well as all comparisons (see Fig.1**B** and Table S2). In this study, we did not intend to focus on behavioral data since these have already been published with these stimuli in dedicated studies [30, 34]. Instead, we were interested in the neural processing associated with the exposition of human participants to primate vocalizations (Fig.1**A**).

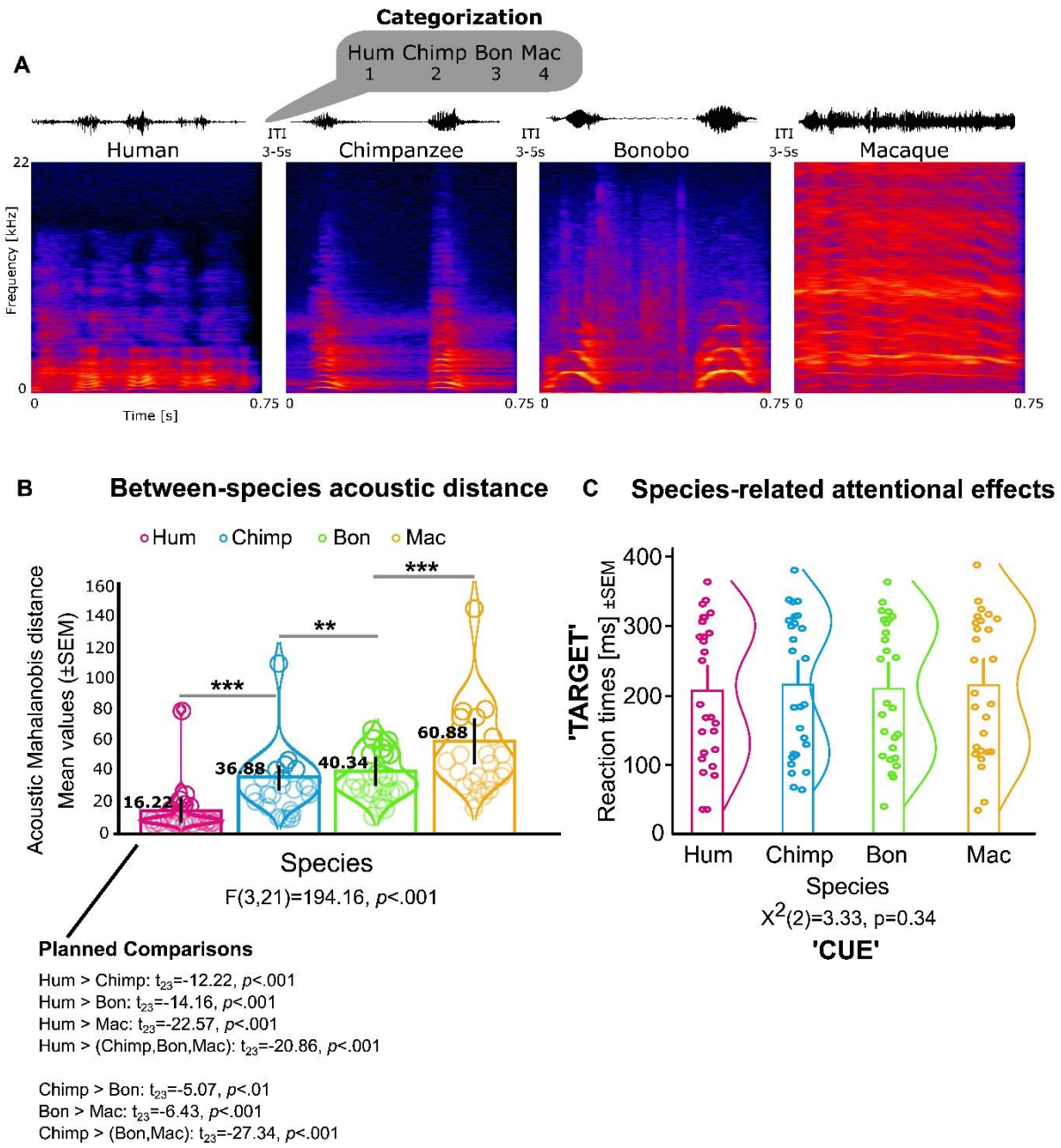

Fig.1: Timecourse of the species categorization task with stimuli example and acoustic distance data. (**A**) Detail of the timecourse of four trials of the species categorization task in non-representative order, including waveform and spectrogram graphs for one example stimulus of each species. (**B**) Scatter plot and histogram of the acoustic Mahalanobis distance data of each stimulus for each species including mean (numbers represent exact mean value) and violin plots of the standard error of the mean in addition to distribution fit. ITI: inter trial interval; Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque.

Neuroimaging data within the sample-specific temporal voice areas

We aimed at uncovering functional changes relative to species categorization and processing within sample-specific (N=23) TVA, as delineated in our hypotheses. As described above, we used three distinct statistical models including trial-level parametric modulators (Model 1-3). We were particularly interested in human brain activity while processing vocalizations of our closest relatives—both acoustically and phylogenetically—namely the chimpanzee but also the bonobo. The present study did not aim at uncovering wholebrain results underlying the processing of each species' vocalizations but rather focused on human voice-sensitive areas, namely the TVA, although corrected statistics (voxelwise $p<.05$ False Discovery Rate) presented in this section were computed with a wholebrain voxelwise approach for higher data reproducibility and generalizability, and not using region-of-interest (ROI) analyses. ROI analyses would most probably have artificially amplified the number of voxels in the TVA in this study. Clusters outside the bounds of the sample-specific TVA are therefore visible but in a desaturated hue to better highlight TVA activations. These clusters are even more visible in supplementary figures with the same contrasts as those presented in this section—with the addition of the [human,chimpanzee > bonobo,macaque] contrast—but with an outline of the TVA from an independent, larger sample of participants excluding the 23 participants of this study (N=98; Fig.S2-4). No difference in terms of a potential attentional bias towards any species of the stimuli used in this study was found (Independent sample of N=28, see Methods and Fig.S1 for detailed information on this aspect).

Model 1: Effects of species processing with vocalization mean fundamental frequency and mean energy as covariates of no-interest at the trial level

In this first 'simple' model, we first wanted to remove from brain activations the part of variance correlating with basic low-level acoustics—as reported in the literature [28, 30, 31, 33], namely mean voice fundamental frequency and energy. A total of four contrasts were overlaid in the figure to test our second hypothesis, according to which phylogenetic—and especially acoustic, see Model 2—proximity would trigger enhanced activity in the TVA, just as the human

voice does. Brain activity specific to chimpanzee vocalizations ([chimpanzee > human, bonobo, macaque]) led to enhanced activity in a cluster of the left anterior STG (aSTG[1], k=91 voxels, Fig.2**AD**) located within the TVA (Fig.2**AC**). A homologous cluster of the right anterior STG was found as well in this contrast (aSTG[2], Fig.2**B**). A similar result was observed when directly contrasting chimpanzee to human vocalizations ([chimpanzee > human]; Fig.2**EFH**) as well as chimpanzee to nonhuman primate calls ([chimpanzee > bonobo, macaque]) in three other clusters of the aSTG, located again within the TVA (Fig.2**ABC**, Table 1). Enhanced activity for human relative to chimpanzee vocalizations ([human > chimpanzee]) was observed in large parts of the anterior, mid and posterior superior and middle temporal cortex (Fig.2**EFG**, Table 1). No voxels reached significance at the wholebrain level for the [bonobo > human, chimpanzee, macaque], [bonobo > chimpanzee, macaque], [bonobo > human], [bonobo > chimpanzee], [bonobo > macaque], [macaque > human, chimpanzee, bonobo], [macaque > chimpanzee, bonobo], [macaque > human], [macaque > chimpanzee], [macaque > bonobo] contrasts. This analysis therefore revealed that the human anterior TVA are sensitive to cross-species primate vocalizations—specifically to chimpanzee but not bonobo or macaque calls using this regression model.

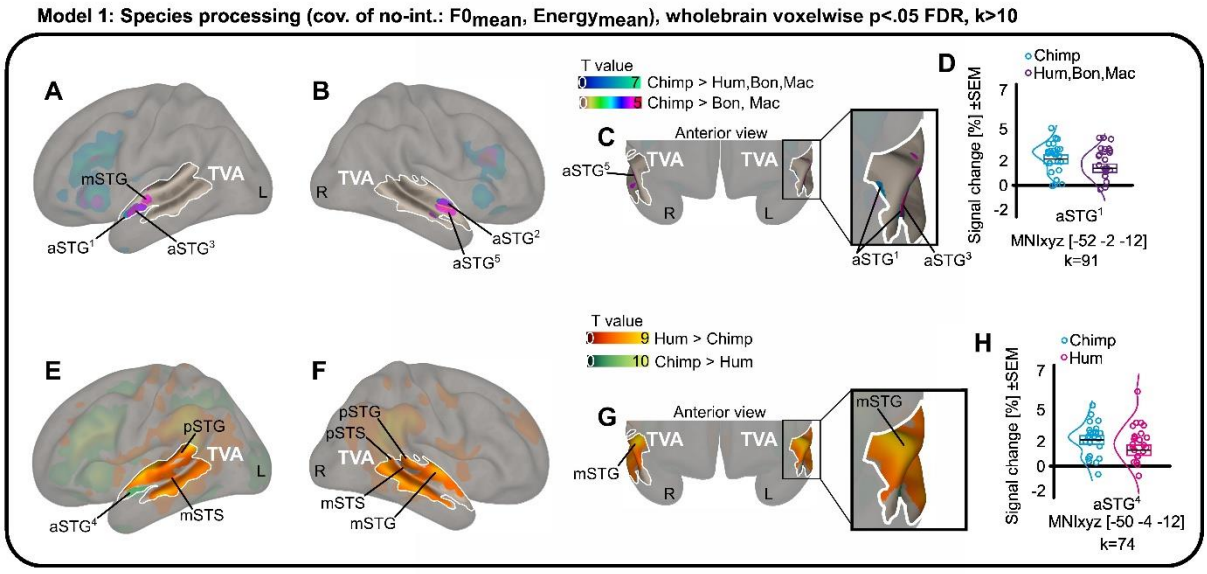

Fig.2: Wholebrain results when contrasting the processing of chimpanzee to other species' vocalizations with mean fundamental frequency and energy as trial-level covariates of no-interest (model 1). (**ABC**) Enhanced brain activity on a sagittal view with activity specific to chimpanzee vocalizations (dark blue to green) as well as between

chimpanzee calls vs bonobo and macaque calls (chimpanzee > bonobo and macaque: brown to red with light yellow outline). (**D**) Percentage of signal change for each individual and relevant species according to the contrast in the left anterior superior temporal gyrus (aSTG[1]). Box plots represent mean value (black line) and the standard error of the mean with distribution fit. (**EFG**) Direct comparison between human and chimpanzee vocalizations (human > chimpanzee: dark red to yellow; chimpanzee > human: dark green to yellow) on a sagittal render. (**H**) Percentage of signal change in the anterior superior temporal gyrus (aSTG[2]) when contrasting chimpanzee to human vocalizations for each individual and relevant species according to the contrast with box plots representing mean value (black line) and the standard error of the mean with distribution fit. Brain activations are independent of low-level acoustic parameters for all species (mean fundamental frequency 'F0' and mean energy of vocalizations). Data corrected for multiple comparisons using wholebrain voxelwise false discovery rate (FDR) at a threshold of $p<.05$. Percentage of signal change extracted at cluster peak including 9 surrounding voxels, selecting among these the ones explaining at least 85% of the variance using singular value decomposition. Circles represent individual values, boxplot represents the mean and its standard error, and half-violin plots show data distribution. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: sample-specific (N=23) temporal voice areas. 'a' prefix: anterior; 'm' prefix: mid; 'p' prefix: posterior; STG: superior temporal gyrus; STS: superior temporal sulcus; L: left hemisphere; R: right hemisphere.

Model 2: Effects of species processing with vocalization acoustic distance from human voice, per species, as covariate of no-interest at the trial level

In this second model, we wanted to remove from brain activations the part of variance correlating with the acoustic distance between each species and the human voice (see Methods for the detailed index of acoustic distance calculated with the human voice as reference). TVA brain activity specific to primate calls was triggered again for chimpanzee vocalizations ([chimpanzee > human, bonobo, macaque]) in a cluster of the left anterior STG within the TVA (Fig.3**ACD**). A similar result was observed when directly contrasting chimpanzee to human vocalizations ([chimpanzee > human]; Fig.3**EH**, Table 2). Enhanced activity for human relative to chimpanzee vocalizations ([human > chimpanzee]) was again observed in large parts of the anterior, mid and posterior superior and middle temporal cortex (Fig.3**EFG**, Table 2). Chimpanzee compared to other nonhuman primate calls ([chimpanzee > bonobo, macaque]) led to enhanced activity in the bilateral aSTG (aSTG[7] and aSTG[9], Fig.3**ABC**). Using this second modelling of the MRI data, no voxels reached significance at the wholebrain level for the [bonobo > human, chimpanzee, macaque], [bonobo > chimpanzee, macaque], [bonobo > chimpanzee], [bonobo > macaque], [macaque > human, chimpanzee,

bonobo], [macaque > chimpanzee, bonobo], [macaque > bonobo] contrasts. Contrasting [bonobo > human] and [macaque > human] yielded to enhanced activity outside the TVA while the [macaque > chimpanzee] comparison activated a very small cluster of within-TVA left planum temporale (see Fig.S5 for these contrasts). In this second model, again, only the calls of chimpanzees triggered specific activity in the anterior TVA.

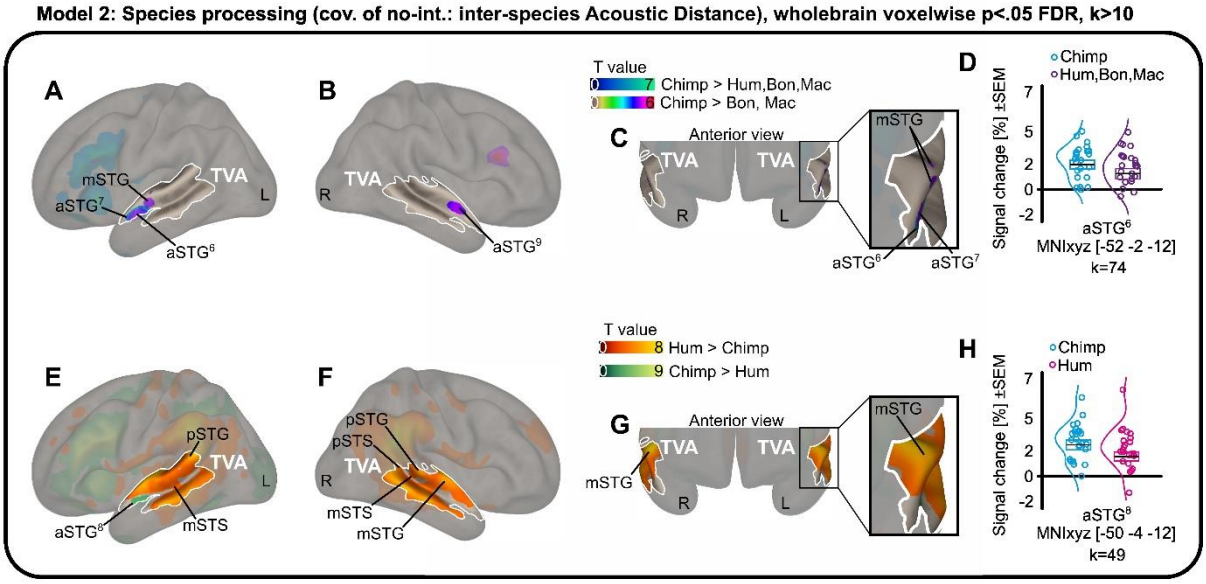

Fig.3: Wholebrain results when contrasting the processing of chimpanzee to other species' vocalizations with Mahalanobis acoustic distance as trial-level covariate of no-interest (model 2). (**ABC**) Enhanced brain activity on a sagittal view with activity specific to chimpanzee vocalizations (chimp > hum,bon,mac; dark blue to green) as well as between chimpanzee calls vs bonobo and macaque calls (chimpanzee > bonobo and macaque: brown to red with light yellow outline). (**D**) Percentage of signal change for each individual and relevant species according to the contrast in the left anterior superior temporal gyrus (aSTG^6). Box plots represent mean value (black line) and the standard error of the mean with distribution fit. (**EFG**) Direct comparison between human and chimpanzee vocalizations (human > chimpanzee: dark red to yellow; chimpanzee > human: dark green to yellow) on a sagittal render. (**H**) Percentage of signal change in the anterior superior temporal gyrus (aSTG^8) when contrasting chimpanzee to human vocalizations for each individual and relevant species according to the contrast with box plots representing mean value (black line) and the standard error of the mean with distribution fit. Brain activations are independent from the acoustic distance of each stimulus for all species. Data corrected for multiple comparisons using wholebrain voxelwise false discovery rate (FDR) at a threshold of *p*<.05. Percentage of signal change extracted at cluster peak including 9 surrounding voxels, selecting among these the ones explaining at least 85% of the variance using singular value decomposition. Circles represent individual values, boxplot represents the mean and its standard error, and half-violin plots show data distribution. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: sample-specific (N=23) temporal voice areas. 'a' prefix: anterior; 'm' prefix: mid; 'p' prefix: posterior; STG: superior temporal gyrus; STS: superior temporal sulcus; L: left hemisphere; R: right hemisphere.

Model 3: Effects of species processing with vocalization most discriminant acoustic parameters (N=6) as covariates of no-interest at the trial level

In this last model, we wanted to elaborate more on the discriminant factors that characterize the low-level acoustic parameters of our set of stimuli. This approach is complementary to the inclusion of acoustic distance in model 2 and extends and refines these results. To do so, we used as trial-level covariates of no-interest the acoustic parameters explaining the most variance ([r > 0.7] and [r < -0.7]) in factors 1-3 of a discriminant analysis of these stimuli [30]—see the methods section for details on this analysis. These parameters therefore include, in this specific order: vocalization loudness, intensity, change in spectrum, bandwidth contour of the second formant (F2), power of the fundamental frequency (F0) and finally the difference in intensity contour. Having these acoustic features as covariates, we ran the same contrasts as in models 1 & 2. As in previous modeling of the imaging data, TVA activity was triggered by chimpanzee vocalizations ([chimpanzee > human, bonobo, macaque]) in yet other, larger bilateral clusters of the aSTG within the TVA (aSTG[10] and aSTG[11], Fig.4**ABCD**), closely resembling activations of model 2. A left-lateralized similar cluster was observed when directly contrasting chimpanzee to human vocalizations ([chimpanzee > human]; aSTG[12], Fig.4**EH**, Table 3). Enhanced activity for human relative to chimpanzee vocalizations ([human > chimpanzee]) was similarly represented as in models 1 & 2 (anterior, mid and posterior superior and middle temporal cortex; Fig.4**EFG**, Table 3). Chimpanzee compared to other nonhuman primate calls in this model ([chimpanzee > bonobo, macaque]) led to the largest clusters observed in the aSTG—all models considered, still within the sample-specific TVA. Indeed, we observed a large left-lateralized cluster of the aSTG extending to the mid STG (aSTG[13], Fig.4**AC**) as well as a right-lateralized cluster (aSTG[14], Fig.4**BC**).

In this last modelling of the fMRI data, no voxels reached significance either at the wholebrain level or within the TVA for the [bonobo > human, chimpanzee, macaque] or [bonobo > chimpanzee, macaque], [bonobo > chimpanzee], [bonobo > macaque] contrasts. We however found activity specific to the processing of macaque calls only in the left TVA, more specifically

in a small cluster of the left mid STS ([macaque > human, chimpanzee, bonobo]) and in a small portion of the planum temporale adjacent to the primary auditory cortex for the [macaque > chimpanzee, bonobo] contrast (see Fig.S6). We also observed significant activations but outside the TVA for the [bonobo > human] and the [macaque > chimpanzee] contrast. Within-TVA activations were observed when contrasting macaque to human vocalizations in the left mid STS, as well as in the in the planum temporale, left mid STS and right mid STG when contrasting macaque to bonobo calls. See Fig.S7 for these results. Using this third model, we again observed chimpanzee-specific activity in the anterior TVA as well as mid STS activity specific to macaque calls, within the TVA.

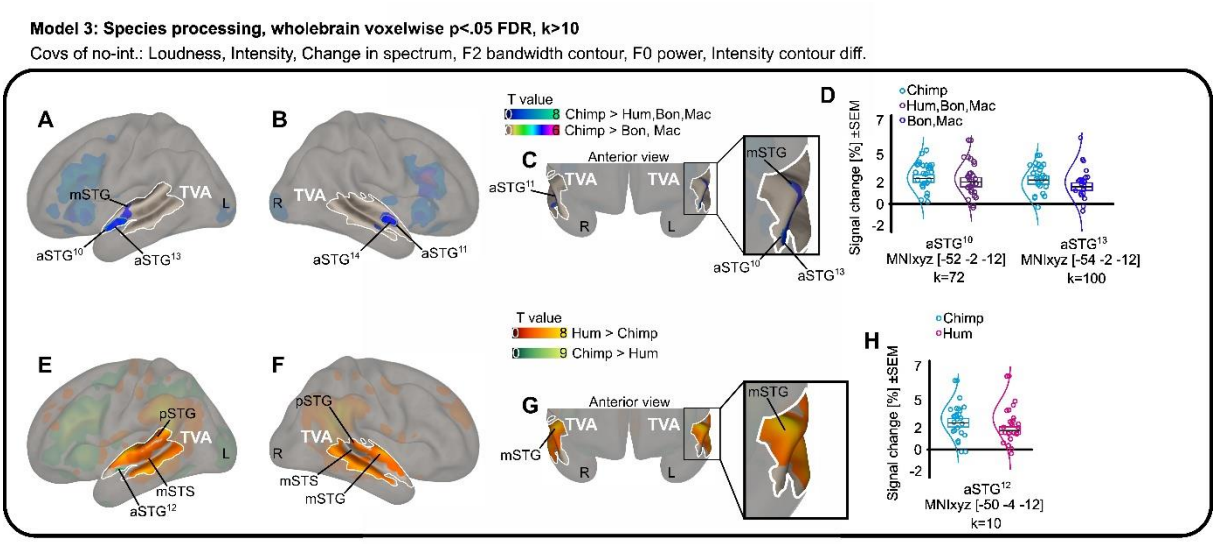

Fig.4: Wholebrain results when contrasting the processing of chimpanzee to other species' vocalizations with vocalization loudness, intensity, change in spectrum, F2 bandwidth contour, F0 power and intensity contour difference as trial-level covariates of no-interest (model 3). (**ABC**) Enhanced brain activity on a sagittal view with activity specific to chimpanzee vocalizations (dark blue to green) as well as between chimpanzee calls vs bonobo and macaque calls (chimpanzee > bonobo and macaque: brown to red with light yellow outline). (**D**) Percentage of signal change for each individual and relevant species according to the contrast in the left anterior superior temporal gyrus (aSTG^{10}). Box plots represent mean value (black line) and the standard error of the mean with distribution fit. (**EFG**) Direct comparison between human and chimpanzee vocalizations (human > chimpanzee: dark red to yellow; chimpanzee > human: dark green to yellow) on a sagittal render. (**H**) Percentage of signal change in the anterior superior temporal gyrus (aSTG^{12}) when contrasting chimpanzee to human vocalizations and when contrasting chimpanzee to bonobo and macaque calls (aSTG^{13}) for each individual and relevant species according to the contrast with box plots representing mean value (black line) and the standard error of the mean with distribution fit. Brain activations are independent of the most discriminant low-level acoustic parameters of the stimuli set [30]. Data corrected for multiple comparisons using wholebrain voxelwise false discovery rate (FDR) at a threshold of p<.05. Percentage of signal change extracted at cluster peak including 9 surrounding voxels, selecting among these the ones explaining at least 85%

of the variance using singular value decomposition. Circles represent individual values, boxplot represents the mean and its standard error, and half-violin plots show data distribution. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: sample-specific (N=23) temporal voice areas. ‘a’ prefix: anterior; ‘m’ prefix: mid; ‘p’ prefix: posterior; STG: superior temporal gyrus; STS: superior temporal sulcus; L: left hemisphere; R: right hemisphere.

A synthesis of the sensitivity of the human TVA to nonhuman primate calls

In the previous sections, we described three different models used to analyze our fMRI data. These models—from the simplest to the more sophisticated one—highlighted enhanced activity within sample-specific bilateral anterior TVA of our participants specifically when processing chimpanzee vocalizations—but also when processing macaque calls in model 3, in the bilateral mid STG, STS and planum temporale. When processing chimpanzee calls, TVA activity was especially enhanced in the aSTG but also in the anterior STS. We therefore regrouped these fourteen chimpanzee-specific aSTG clusters in Fig.5—most of them overlap greatly but we still named them individually according to each contrast and analysis for exhaustivity—overlaid with sample-specific TVA (Fig.5**CD**) and with the more general TVA from an independent sample of ninety-eight participants (Fig.5**AB**). Zooming closely, the area of maximal overlap between these regions (the orange surface) is located within the more general as well as within the sample-specific TVA. Interestingly, left-lateralized more medial clusters of aSTG were outside the outline of the sample-specific but not of the general TVA (Fig.5**AC**), while this was not the case for right-lateralized aSTG activations. Comparing the areas recruited when processing chimpanzee to bonobo and macaques calls, this contrast—especially in model 3, yielded to distinct clusters of aSTG. This result is visible when looking at the three ‘rich blue’ outlines in every panel of Fig.5. The results synthesized here highlight the important role of acoustic parameters and emphasize the role of the most discriminant acoustic features on TVA activity relating to nonhuman primate vocalizations, especially those of chimpanzees and macaques.

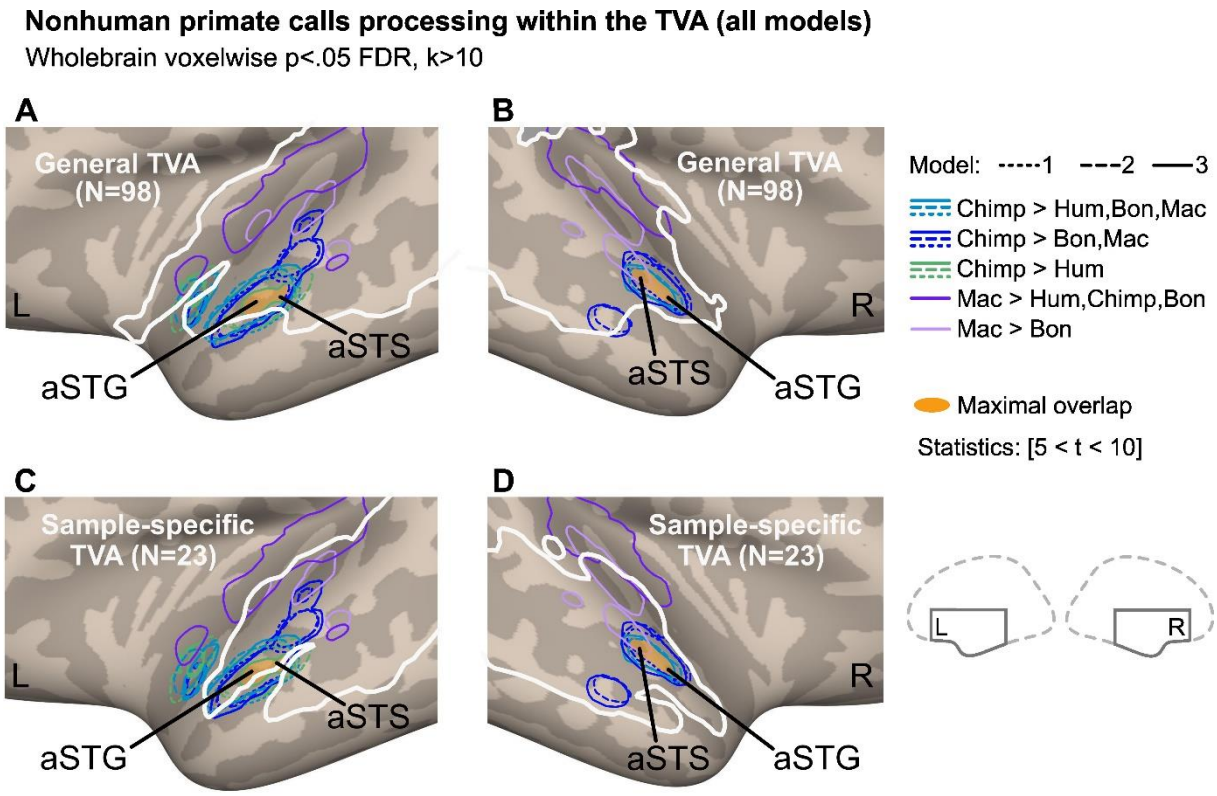

Fig.5: Synthesis of mid and anterior TVA clusters of activity recruited specifically by the processing of chimpanzee and macaque vocalizations (Models 1,2,3). aSTG and aSTS clusters recruited for the processing of chimpanzee calls as opposed to: human voices (green); bonobo, macaque calls (blue) and human voice; bonobo and macaque calls (turquoise) in the general TVA (**AB**, N=98) as well as in the sample-specific TVA (**CD**, N=23). Macaque results are only significant for Model 3 (purple: Macaque vs all other species; lilac: Macaque vs other nonhuman primates). Clusters are represented across all statistical models (Model 1: dotted line; Model 2: dashed line; Model 3: solid line). Model 1: mean of fundamental frequency and energy (covariates of no-interest, N=2); Model 2: acoustic distance (covariate of no-interest, N=1); Model 3: acoustic parameters that characterize low-level acoustics of our stimuli following a discriminant analysis (covariates of no-interest, N=6). Data are all corrected for multiple comparison using wholebrain voxelwise false discovery rate (FDR) at a threshold of *p*<.05. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: temporal voice areas. 'a' prefix: anterior; STG: superior temporal

These results are given even more weight by more fined-tuned comparisons of voice versus non-voice material in the voice-localizer task, namely by splitting the non-vocal blocks as a function of the auditory sounds they contain. In this more specific outline of TVA subregions, we observed that most chimpanzee- and macaque-specific STG and STS regions were still within the bounds of TVA, especially in the most relevant case when the outline represented a comparison of human voice signals to animal or nature sounds (Fig.6**A**-**D**), while the outlines

for human voice versus music or noise excluded most parts of the clusters of activation of these nonhuman primate species' calls in the TVA (Fig.6**E**-**H**).

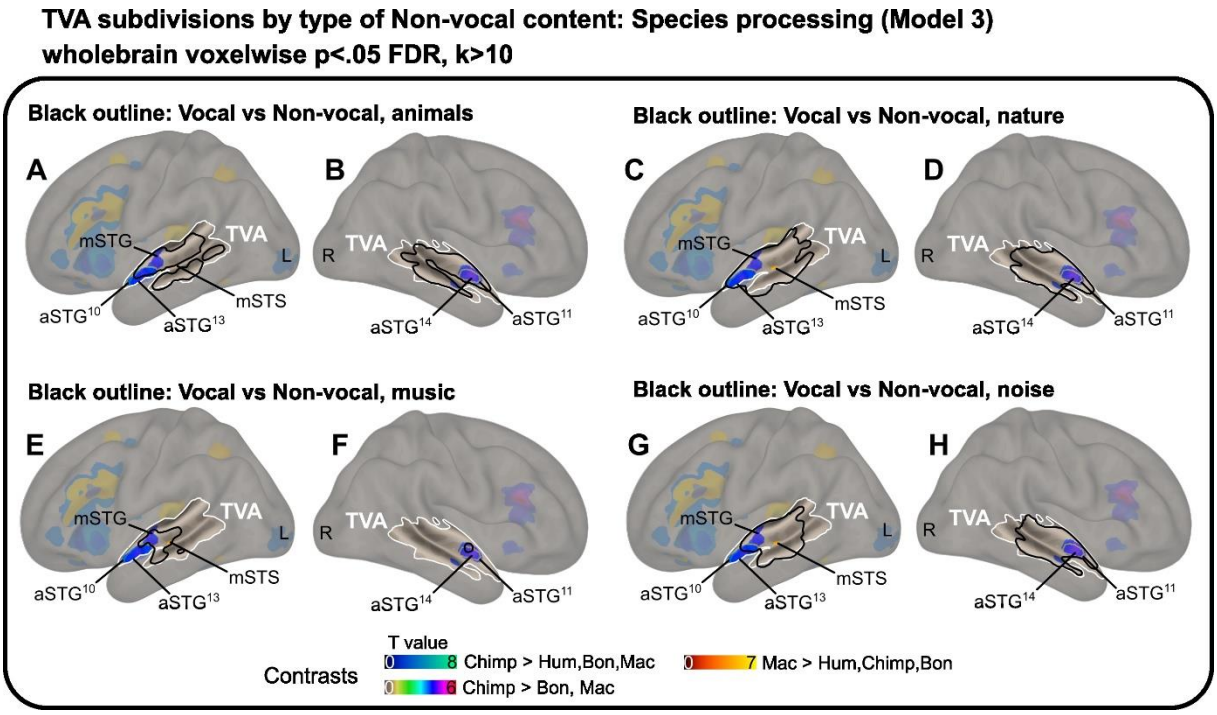

Fig.6: Clusters recruited specifically by the processing of chimpanzee and macaque vocalizations (Model 3) in subregions of the TVA, as a function of non-vocal material type. Enhanced brain activity on a sagittal views with activity specific to macaque vocalizations (red to yellow), specific to chimpanzee vocalizations (dark blue to green) as well as between chimpanzee calls vs bonobo and macaque calls (chimpanzee > bonobo and macaque: brown to red with light yellow outline). Brain activations are independent of the most discriminant low-level acoustic parameters of the stimuli set [30]. Data corrected for multiple comparisons using wholebrain voxelwise false discovery rate (FDR) at a threshold of *p*<.05. Black outline represents: voice compared to non-vocal stimuli of animal sounds (**A,B**), nature sounds (**C,D**), music (**E,F**), artificial noise (**G,H**). Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: sample-specific (N=23; white outline) temporal voice areas. STG: superior temporal gyrus; STS: superior temporal sulcus; 'a' prefix: anterior; 'm' prefix: mid; L: left hemisphere; R: right hemisphere.

Discussion

The present study provides evidence of the sensitivity of the human TVA to cross-species vocalizations, especially to chimpanzee calls but also to macaque vocalizations, as illustrated by specific enhanced activity in the bilateral mid and anterior STG and STS—within sample-specific TVA. These results were obtained through statistical modeling of the MRI data that

included either simple acoustics or the use of Mahalanobis acoustic distance between species and the most discriminant acoustic features specific to our stimuli as covariates. These two latter analyses converged and yielded to greatly overlapping results, especially in the anterior TVA. Therefore, our results suggest that vocalizations from another ape species recruits subregions of human temporal cortex that process species-specific voices in humans—namely the bilateral, sample-specific TVA. This evidence speaks in favor of cross-species primate vocalization processing in the anterior and mid TVA of humans—for chimpanzee and macaque calls, respectively. While our acoustic data confirmed the hypothesized hierarchy of acoustic distance as a function of phylogenetic distance between our species, we still observed mid STG and STS activity for macaque versus bonobo calls and a small cluster in the left mid STS specific to macaque calls in model 3—an unexpected result since macaques are the most distant species from humans both phylogenetically and acoustically in our study. Therefore, while we initially hypothesized that primate calls would exclusively recruit human TVA as a function of a combination of phylogenetic and acoustic proximity, our data also point toward the greater importance of the most discriminant acoustic features rather than acoustic distance alone. We discuss these aspects below in more detail and interpret their general meaning and subsequent scientific implications, in addition to highlighting the limitations of our study.

Often specifically associated with the processing of conspecific vocalizations (e.g., in humans [2, 22, 27], macaques [21, 27, 36], and dogs [19]), the present study challenges the common view of the TVA as ‘species-specific’ and illustrates that human voices, chimpanzee and macaque calls can enhance activity in the TVA. We think that the distinct locations of these TVA subregions recruited for processing the vocalizations of these primate species matters. In fact, there might be a possible association between anterior TVA—specific to processing chimpanzee calls—and the higher recognition performance of chimpanzee calls compared to those of bonobos or macaques in human participants [34]. Anterior TVA activity specific to the processing of chimpanzee calls occurred when these were compared to both human and nonhuman primate species, solely to other nonhuman primate vocalizations, or directly to the

human voice. However, homologous specific results were not observed for bonobo and they were more scarce—especially between models—for macaque vocalizations: we found macaque-specific activity in a small area of the planum temporale and in a small cluster of the left mid STS, congruent for instance with locations observed in the general processing of animal sounds, especially in the planum temporale [37, 38]. On the other hand, within-TVA anterior STG activity was also observed when chimpanzee vocalizations were directly compared to the human voice. We think this result highlights the cross-species specificity of this anterior subregion of the TVA for processing species phylogenetically close to humans and especially with human-like acoustics, namely the calls of chimpanzees in the case of our study. Because of their vocal proximity, the perception of human voices and chimpanzee calls in socio-affective contexts could involve a common 'social' core of the brain, which increases activity in brain regions such as the anterior TVA, as reported previously in studies pertaining to social contextual information processing in the anterior STG [39, 40]. Differences at the level of processing complexity between the two types of vocalizations could also explain this observation, while we demonstrated that saliency or attention-related effects do not exist between our species stimuli. Indeed, previous studies have shown the role of the anterior STG and the anterior STS in the conceptual representation of social context by the human voice [39-42]. Therefore, our data might suggest that the anterior part of the superior temporal cortex could be recruited to process the social context of human and chimpanzee vocal stimuli. However, this processing would be more automated for the perception of the human voice than for chimpanzee calls because of our high exposure and expertise as humans to these vocal signals, but these hypotheses should be addressed scientifically in studies dedicated to this topic. Our results are also complementary to and coherent with a 'voice patch' system in the brain of primates, as put forward by Belin and colleagues [43], and according to which distinct 'patches' or subregions of the temporal lobe—especially its anterior portion, would be interconnected and would allow for the processing of voice information. Such system would be present in many primate species such as humans, macaques and marmosets, with most recent evidence suggesting a population of neurons in the anterior STG of the macaque brain

selective to human voice [44], as also anticipated in another study on macaques and also in the anterior STG [45]. These fascinating and converging results mirror our present data—with 'chimpanzee-selective' responses in the anterior STG/TVA of our human participants—and strongly emphasize the need for pursuing a comparative approach in order to clarify the cross-species neural bases underlying the processing of human and nonhuman primate vocal signals. As we mentioned previously, these interpretations are, considering our results, free of any potential attentional bias towards one species over the others, since no effect was observed on that matter in a control, behavioral study involving an independent sample of twenty-eight participants in a species-specific exogenous cueing attentional paradigm—Methods and Fig.S1.

Importantly, our data also emphasize the influence of acoustic features and especially acoustic proximity between human and chimpanzee vocalizations: we show that activity in the anterior STG and more generally in the anterior TVA partly depends on phylogenetic and more importantly on key acoustic features and acoustic proximity. Consistent with previous studies [24, 25], we did not expect TVA activity for macaque calls processing because they are both phylogenetically and acoustically more distant from humans than the other species in this study—although, as above mentioned, we found a very small cluster in the left primary auditory cortex and mid STS for model 3. It is interesting to note that in model 2, with acoustic distance as trial-level covariate, we observed TVA activity only for chimpanzee but not macaque calls, giving further weight to the importance of acoustic distance in this context. Also, if only phylogenetic proximity mattered, bonobo calls should also elicit activity in the TVA because they are as phylogenetically close to humans as chimpanzees. But this viewpoint is rather reductive and our results show that this is not entirely correct, and that activity in the TVA crucially depends on the acoustic properties of the perceived vocalizations since we cannot infer phylogeny from vocalizations. This interpretation is strongly supported by the inclusion of acoustic Mahalanobis distance for each species compared to the human voice as a trial-level covariate of no-interest. Using such modelling, differential neuroimaging results between

chimpanzee and bonobo vocalizations were explained by both acoustic and phylogenetic proximity in the TVA. These results are consistent with the recent proposal—and recent findings [30]—that there are substantial differences between chimpanzee and bonobo vocalizations. These encompass fundamental frequency range and mean due to larynx length [28, 32]—despite the evolutionary relatedness to chimpanzees [28]. Therefore, the interaction between phylogeny and acoustic distance or proximity would explain the anterior TVA expansion for processing specifically chimpanzee but not bonobo vocalizations. This argument however falls short to explain the recruitment of the TVA by macaque calls in model 3.

Overall, it seems reasonable to hypothesize that TVA activity is not *per se* human-specific [2, 41] but that TVA are instead sensitive to vocalizations from other primate species, provided that these vocalizations have sufficient acoustic proximity to human vocal signals—which would in itself be related to anatomical and/or behavioral changes throughout phylogenetic evolution. This integrative view is again consistent with the concept of a ‘voice patch’ system in the primate brain [43]. We therefore propose that the mid and anterior TVA, unlike the rest of the TVA, would be heterospecific—sensitive to vocalization acoustics triggered by evolution. This proposition also implies a validation of our mechanistic hypothesis according to which the mean fundamental frequency of chimpanzee but not bonobo calls—the former being much closer to the mean fundamental frequency of the human voice [28], would allow for a better identification and recognition of chimpanzee calls by humans. This advantage would rely on neurons of the human auditory cortex—both the primary and more secondary regions—being specialized in the processing of low to mid fundamental frequencies such as those of the human voice and chimpanzee calls. In our third analysis model, we looked further into this aspect and included several acoustic properties of our stimuli as a function of the four species in our stimuli. A discriminant analysis [30] allowed us to select specific acoustic features that best discriminate between our species stimuli. Namely, we took the six parameters explaining the most the differences between our stimuli, including vocalization loudness, intensity—similar to our ‘energy’ covariate of model 1, in addition to change in spectrum, F2 bandwidth

contour, F0 power and intensity contour difference. Using these more sophisticated acoustic features as covariates of no-interest, we still obtained brain imaging results very similar to those of model 1 and even closer to model 2—with acoustic distance as covariate, yet with some subtle differences in anterior STG cluster size and location. The peaks were indeed located more ventral and were larger—as compared to results of model 1 & 2, especially for the processing of chimpanzee-specific and macaque-specific vocalizations compared to all primate species and to nonhuman primates alone. These results suggest that the inclusion of spectrum change to intensity- and frequency-related acoustical parameters of the vocal signals slightly shifted and enlarged activation locations in the anterior STG. This result is again congruent with the proposed existence of 'voice patches' in the temporal lobe of primate species [43], with the interconnectivity of these patches highly depending on very fine-grained acoustic aspects of primate vocal signals. This step motivated the inclusion of these parameters as covariates of no-interest in neuroimaging model 3, to retain brain activations marginally independent of such acoustics. The congruence between these data should be explored in more detail in the future by the combination of computational bioacoustics and functional neuroimaging, due to the high relevance and sensitivity of combining these techniques to investigate primate social communication [46].

A final but maybe more secondary interpretation arising from our results regarding bonobo calls also supports the evolutionary divergence of this peculiar species. According to the self-domestication hypothesis, bonobos would have evolved differently than chimpanzees due to selection against aggression [47]. Interestingly, differentiation in the evolutionary path of bonobos has influenced both their behavior [31] and morphology, leading to differences at the level of call production [28, 33]. Considering these documented acoustic differences and putting them in perspective with our neuroimaging data, the calls of our last common ancestor with the other Pan species 8 million years ago [48], may have been closer to those uttered by modern chimpanzees than to those of bonobos. Our data indeed show that modern human brains remain more sensitive to the acoustic characteristics of the calls of the former compared

to the latter, arguing for more conserved calls between modern chimpanzees and humans. This aspect is also in line with significant differences between, for instance, the fundamental frequency of human baby cries or babbling (~250-600Hz) compared to that of bonobos (~1000-3500Hz) [49, 50], while they correspond more closely to the fundamental frequency of chimpanzee calls (~500-1000Hz) [28]. In our study, bonobo calls definitely are so much different than those uttered by the species of our other stimuli, that they presumably fall outside of the phylogeny and acoustic proximity factors that we outlined so far. This would also put into perspective the recruitment of mid TVAs for macaque calls.

In a sense, we therefore validate our first hypothesis regarding the existence of acoustic distance between each primate species used in our study. We also partially validate our second hypothesis, albeit not completely. In fact, macaque compared to bonobo or other primate calls in model 3 revealed mid TVA activations, and we think that these activations may depend specifically on the importance of the most discriminant acoustic features. Several TVA subregions or 'patches' underlying cross-species primate vocalization processing might therefore exist, and our data highlight at least one of them in the mid and anterior portion of the TVA. We will now discuss in further detail task-related limitations that might account for the partial divergence between our results and hypotheses.

Even though we tried to control for critical acoustic features, species categories and their related evolutionary distance, several limitations should be mentioned. These limitations are both theoretical and methodological. First, we cannot rule out the fact that including more primate species in our set of stimuli would not have influenced the results. In fact, even though our species categories were specifically chosen for this task, the inclusion of vocalizations from other great apes—such as gorillas or orangutans—would have broadened the scope of our results. Related to this aspect, we can also mention that tackling primate phylogeny, which spans over millions of years, with only four species restricts the possible inference based on our results. Second, we observed improved sensitivity of our data by the use of more sophisticated acoustic modeling, namely the inclusion of both between-species acoustic

distance and of the most discriminant acoustic features in the functional imaging data. However, we did not include as stimuli—or in a control task—the synthesized acoustic parameters of interest, for instance by using species-specific F0 contour or its spectral content in other neutral, comparable auditory stimuli. We cannot therefore *completely* rule out that such task would not trigger brain activations that overlap with our results—although such data would not be mutually exclusive with our data and interpretation. Future work should therefore address with the greatest level of detail the specific question of acoustics in primate vocalization processing, in addition to adding more—as well as synthesized—stimuli from other great ape species. The origin of these acoustic differences should also be investigated, since we can assume that these differences originated at least partially from evolutionary processes as well as survival and adaptation mechanisms. Finally, individual differences in the processing and preference of one species over another or over all the others cannot be ruled out, even though we provide evidence that attentional effects toward the vocalizations of a specific species did likely not exist in our data. Therefore, individual differences should be assessed in more detail in the future, with the inclusion of participant-level covariates such as questionnaire scores assessing the familiarity with primate vocalizations or the hedonic value of these vocalizations for each individual. Among the more general limitations of nonhuman primate neuroscience lies the fact that more inclusive and large-scale collaborations would be needed. Such collaborations and framework would lead to a better study and understanding of primate neuroscience, and previous initiatives have recently been put forward in this direction [51, 52].

Taken together, our data suggest that phylogeny-driven specific acoustic features appear to be necessary to trigger cross-species activity in the human temporal voice areas—especially in subregions of the TVA in which increased activity underlies voice signals compared to animal and nature sounds. We provide evidence for specific anterior and mid TVA subregions that underlie the processing of the calls of one of our closest relatives, namely chimpanzees but also of macaques, respectively. In line with recently reported literature, we contend that the human TVA are also involved in the processing of heterospecific primate vocalizations,

provided they exhibit sufficient phylogenetic and especially spectro-temporal acoustic proximity to the human voice; as such, we predict that other similarities will be uncovered in the processing of human and nonhuman primate communicative signals. Finally, our results support a critical evolutionary continuity between the structure of human and chimpanzee vocalizations, possibly reflecting one of their common ancestors, as opposed to bonobo vocalizations that underwent more recent and critical changes within the last 1-2 million years. In contrast, the chimpanzee vocal system may be closer to the one of the common ancestors of humans and chimpanzees, as shown by the conserved activation in the human modern brain.

Material and Methods

Species categorization task

Participants

Twenty-five right-handed, healthy, either native or highly proficient French-speaking participants took part in the study. One participant was excluded because he had no correct response at all and may have fallen asleep, while another participant was excluded due to incomplete scanning and technical issues at the MRI scanner, leaving us with twenty-three participants (10 female, 13 male, mean age 24.65 years, SD 3.66). With this sample size and our study design, we achieved a power of 75.12% for a between-means comparison with Effect size dz=0.5 and alpha=0.05 as calculated in G*Power version 3.1.9.7 [53]. All participants were naive to the experimental design and study, had normal or corrected-to-normal vision, normal hearing and no history of psychiatric or neurologic incidents. Participants gave written informed consent for their participation in accordance with ethical and data security guidelines of the University of Geneva. The study was approved by the Ethics Cantonal Commission for Research of the Canton of Geneva, Switzerland (CCER) and was conducted according to the Declaration of Helsinki.

Stimuli

Seventy-two vocalizations of four primate species (human, chimpanzee, bonobo, and rhesus macaque) were used in this study (see Fig.1**A**). The eighteen selected chimpanzee, bonobo and rhesus macaque vocalizations contained single calls or call sequences produced by 6 to 8 different individuals in agonistic (threat, distress) or affiliative ('positive') social contexts. These were randomly selected—in a between-participants fashion—among our full database of primate stimuli, containing specifically: 15 chimpanzee individuals (recorded in the wild in the Budongo forest, Uganda), 10 bonobo individuals (recorded in the wild in the Salonga national park, Democratic Republic of Congo, DRC) and 16 macaque individuals (recorded in the wild from semi-free monkeys on Cayo Santiago, Puerto Rico). We then selected eighteen human voices obtained from a nonverbal validated stimuli set of Belin and collaborators [54], which were expressed by two male and two female adults expressing positive or negative social interactions. All vocal stimuli were standardized to 750 milliseconds using PRAAT (www.praat.org) but were not normalized in any way in order to preserve the naturalness of the sounds [55] and to allow for low-level acoustic parameters to be used in neuroimaging data modelling.

Experimental procedure and paradigm

Laying comfortably in a 3T scanner, participants listened to a total of seventy-two stimuli pseudo-randomized and played binaurally using MRI compatible earphones at 70 dB SPL (Model 'S14', Sensimetrics Corporation, Gloucester, MA, USA). At the beginning of the experiment, participants were instructed to identify the species that expressed the vocalizations using a keyboard. For instance, the instructions could be "Human – press 1, Chimpanzee – press 2, Bonobo – press 3 or Macaque – press 4". The pressed keys were pseudo-randomly assigned across participants (response box: fORP, Cortech Solutions, Inc., Wilmington, NC, USA). In a 3-5 second interval (jittering of 400 ms) after each stimulus, participants were asked to categorize the species. If the participant did not respond during this

interval, the next stimulus followed automatically. See Fig.1**A** for a detailed illustration of the paradigm.

Temporal voice areas localizer task

Participants

Two independent samples of participants performed the task while undergoing fMRI scanning: the sample of this study (10 female, 13 male, mean age 24.65 years, SD 3.66), leading to the delineation of sample-specific TVA; an independent sample of ninety-eight right-handed, healthy, either native or highly proficient French-speaking participants (52 female, 46 male, mean age 24.66 years, SD 4.97) leading to the delineation of more general and representative TVA. All participants were naive to the experimental design and study, had normal or corrected-to-normal vision, normal hearing and no history of psychiatric or neurologic incidents. Participants gave written informed consent for their participation in accordance with ethical and data security guidelines of the University of Geneva. The study was approved by the Ethics Cantonal Commission for Research of the Canton of Geneva, Switzerland (CCER) and was conducted according to the current regulations in Switzerland.

Stimuli and paradigm

Auditory stimuli consisted of sounds from a variety of sources [2]. Vocal stimuli were obtained from 47 speakers: 7 babies, 12 adults, 23 children and 5 older adults. Stimuli included 20 blocks of vocal sounds and 20 blocks of non-vocal sounds. Vocal stimuli within a block could be either speech 33%: words, non-words, foreign language or non-speech 67%: laughs, sighs, various onomatopoeia. Non-vocal stimuli consisted of natural sounds 14%: wind, streams, animals, 29%: cries, gallops, the human environment, 37%: cars, telephones, airplanes or musical instruments, 20%: bells, harp, instrumental orchestra. The paradigm, design and stimuli were obtained through the Voice Neurocognition Laboratory website (http://vnl.psy.gla.ac.uk/resources.php). Stimuli were presented through earphones (Model

‘S14’, Sensimetrics Corporation, Gloucester, MA, USA) at an intensity that was kept constant throughout the experiment 70 dB sound-pressure level. Participants were instructed to actively listen to the sounds. The silent inter-block interval was 8 s long.

Exogenous attention task using species as cues

This task was performed to control for potential attentional biases or specific salience of one species compared to the others. This task was therefore a control task and the results are reported in Fig.S1. The task was designed according to work on attention orienting following vocal material presentation [56]. More specifically, a cue is first presented and is quickly followed by the presentation of a neutral target to detect. The detection of this target can reliably be delayed or accelerated depending on the nature of the cue. In this control study, the cue corresponded to the stimuli of the main species categorization task followed by a sine wave tone (a ‘bip’) that had to be detected as fast as possible (the target). Any specific attentional bias of a species would therefore trigger reaction times differences for target detection, while no differences would invalidate any attentional effect or increased salience linked to a specific species. We did not observe any difference between cue species in this task ($\chi^2(2)=3.33$, $p=0.34$, Fig.S1).

Participants

Twenty-eight participants (independent of the samples presented so far) took part in this behavioral study (15 female, 13 male, mean age 22.63 years, SD 5.00). With this sample size and our study design, we achieved a power of 82.48% for a between-means comparison with Effect size dz=0.5 and alpha=0.05 as calculated in G*Power version 3.1.9.7 [53]. All participants were naive to the experimental design and study, had normal or corrected-to-normal vision, normal hearing and no history of psychiatric or neurologic incidents. Participants gave written informed consent for their participation in accordance with ethical and data security guidelines of the University of Geneva. The study was approved by the Ethics

committee of the University of Geneva (Switzerland), Department of Psychology and Educational Sciences, and was conducted according to the Declaration of Helsinki.

Stimuli and paradigm

The cues of the task were the species stimuli (human voice, chimpanzee, bonobo and macaque calls) used in the species categorization task above. The target 'bip' was a sine wave tone created using Matlab (Matlab 2020a, The Mathworks, Inc., Natick, MA, USA) with a wave frequency of 600Hz, a fade-in and fade-out of 10 ms each and a total duration of 100 ms. The auditory material was presented through headphones (Sennheiser HD-25 II, Sennheiser electronic SE & Co. KG, Germany) at a constant sound-pressure level of 70dB. The procedure unfolded as follows, on a computer with a light grey screen background: for each trial (N=172), a cue (human voice, chimpanzee, bonobo or macaque call) was presented for 750 ms while a black fixation cross was presented at the center of the screen. Following a jittered blank screen of duration 100 to 250 ms (in steps of 50 ms), the target bip was presented for 100 ms. Right at the end of the presented of the bip, the fixation cross turned white, indicating that the response screen had started and that a response was expected as fast and accurately as possible as instructed, by using the 'space' key of the keyboard. A varying inter-trial interval of 1 s to 2.5 s was used (in steps of 500 ms). Among the total of 172 trials, there were 24 trials per species (12 in agonistic and 12 in affiliative social contexts) for a total of 96 stimuli (4 species * 24 trials=96) that were each presented twice (N=96*2=172).

Behavioral data analysis

Species categorization task

Accuracy

Behavioral data were exclusively used to exclude participants who had below chance level categorization of human voices. Therefore, data from twenty-three participants mentioned in the *Species Categorization Task - Participants* section above were analyzed using R studio

software (R Studio team [57] Inc., Boston, MA, url: http://www.rstudio.com/). These data can be found in a published article focused on decisional aspects in the frontal cortex—region-of-interest analysis and computational modelling of the probability of correct species categorization—using the same species stimuli as in this study [34]. Since behavioral data are not part of the questions of interest of this paper addressing neural correlates of the species-specific processing of vocalizations within the temporal voice areas in human participants, and since they are published elsewhere [34], these are not presented.

Exogenous attention task

Reaction times

This study did not contain any good or bad response, and the dependent variable of interest was the reactions times to detect the target 'bip'. In order to remove extreme values, we discarded for each participant the values below the 5th percentile and above the 95th percentile. In average, the number of trials per participants therefore went down from 172 to 165 (~4.1% of trials removed). Data were then analyzed using R studio software (R Studio team [57] Inc., Boston, MA, url: http://www.rstudio.com/) using linear mixed effects modeling of the lme4 package [58]. The formula was the following:

$$model < -lmer(RT \sim Species * Context + (1 + StimID|ParticipantID) + (1|Gender) + (1|Age), data = mydata)$$

in which *RT* is the reaction times (dependent variable), *Species* is the four species of the stimuli (human, chimpanzee, bonobo, macaque) interacting with the *Context* of production (affiliative, agonistic) (fixed effects). The random effects were: the random slope of the identity of each stimulus (*StimID*) as a function of participants (*ParticipantID*), participant gender (*Gender*) and age (*Age*). This model explained 57.54% of the variance in the data (R2c).

There was no significant effect of Species ($\chi^2(3)=3.33$, $p=.34$), a significant effect of Context ($\chi^2(2)=11.39$, $p<.01$) and no interaction between Species and Context ($\chi^2(6)=3.06$, $p=.80$). The effect of Context was explained by slower reaction times for affiliative than agonistic

vocalizations, independent of Species ($\chi^2(1)=6.083.33$, $p<.05$). Descriptive statistics per Species for the reaction times were the following: Human, mean=207.28, SD=118.43; Chimpanzee, mean=211.31, SD=127.07; Bonobo, mean=209.45, SD=151.85; Macaque, mean=214.05, SD=117.84. Descriptive statistics per Context for the reaction times were: Affiliative, mean=215.03, SD=115.41; Agonistic, mean=208.16, SD=135.39. See illustration in Fig.S1.

Acoustic analysis of the vocalizations

Mahalanobis acoustic distance (Neuroimaging model 2)

To quantify the impact of acoustic similarities in human recognition of affective vocalizations of other primates, we extracted 88 acoustic parameters from all vocalizations using the extended Geneva Acoustic parameters set defined as the optimal acoustic indicators related to voice analysis (GeMAPS [59]). This open-source set of acoustical parameters was selected based on: i) their potential to index affective physiological changes in voice production, ii) their proven value in former studies as well as their automatic extractability, and iii) their theoretical significance. GeMAPS relies on an automatic extraction system, which therefore automatically extracts an acoustic parameter set from an audio file—in an unsupervised, minimalistic manner. Then, to assess the acoustic distance between vocalizations of all species, we ran a General Discriminant Analysis model (GDA). More precisely, we used the 88 acoustical parameters in a GDA in order to discriminate our stimuli based on the different species (human, chimpanzee, bonobo, and rhesus macaque). Among these 88 acoustical parameters, we excluded those that were strongly correlating—i.e., with correlation scores r>.90—to avoid redundancy and minimize multicollinearity. Following this selection process and the GDA, we eventually retained 16 acoustic parameters (Table S1).

We subsequently computed multidimensional Mahalanobis distances to classify the 72 stimuli on these selected acoustical features. A Mahalanobis distance is a generalized pattern

analysis comparing the distance of each vocalization from the centroids of the different species vocalizations. This analysis allowed us to obtain an acoustical distance matrix used to test how the acoustical distances were differentially related to the different species (see Fig.1**B**) and we used it as a covariate of no-interest in neuroimaging model 2. Using a one-way ANOVA with the Distance as the dependent variables and the Species as independent variable, the main effect of Species was significant ($F(3,88)=15.84$, $p<.001$). All between-species differences were significant ($.01<p<.001$; see Fig.1 and Table S2).

These data are the topic of a publication about the impact of acoustic parameters on the recognition of the affective cues of primate vocalizations by human participants [30]. All details are described in this article for the present stimuli.

Most discriminant low-level acoustic parameters (Neuroimaging model 3)

Following the GDA on the 88 acoustic parameters of the species stimuli presented above, we decided to use as covariates of no-interest the most discriminant low-level acoustic features of our stimuli to maximize brain activations that are independent of these features. We therefore included the most significant acoustic features ([r>.70] and [r<-.70]) of the first three factors of the GDA, that explained 27.14%, 21.63% and 18.99% of the variance, respectively [30]. Such selection left us with the following acoustic features: Factor 1, (1) vocalization loudness (r=0.92), (2) intensity (r=0.87), (3) change in spectrum (r=0.72); Factor 2, (4) bandwidth contour of the second formant (F2; r=0.79); Factor 3, (5) power of the fundamental frequency (F0; r=0.80) and finally (6) the difference in intensity contour (r=-0.71). The acoustic parameters were used as covariates of no-interest (N=6) in that specific order—namely, from the highest to lowest factor saturation—in neuroimaging model 3.

Again, all details of this analysis are described in detail in a dedicated article for the present stimuli [30] and the values are reported in Table S1.

Imaging data acquisition

Species categorization task

Structural and functional brain imaging data were acquired by using a 3T scanner Siemens Trio, Erlangen, Germany with a 32-channel coil. A 3D GR\IR magnetization-prepared rapid acquisition gradient echo sequence was used to acquire high-resolution (0.35 x 0.35 x 0.7 mm^3) T1-weighted structural images (TR = 2400 ms, TE = 2.29 ms). Functional images were acquired by using fast fMRI, with a multislice echo planar imaging sequence with 79 transversal slices in descending order, slice thickness 3 mm, TR = 650 ms, TE = 30 ms, field of view = 205 x 205 mm2, 64 x 64 matrix, flip angle = 50 degrees, bandwidth 1562 Hz/Px. In total for this task, 636 functional volumes of 79 slices were acquired for each participant for a total of 50244 slices per participant. For our whole sample of twenty-three participants, 14628 volumes were acquired for a grand total of 1’155’612 slices.

Temporal voice areas localizer task

Structural and functional brain imaging data were acquired by using a 3T scanner Siemens Trio, Erlangen, Germany with a 32-channel coil. A magnetization-prepared rapid acquisition gradient echo sequence was used to acquire high-resolution (1 x 1 x 1 mm^3) T1-weighted structural images TR = 1,900 ms, TE = 2.27 ms, TI = 900 ms. Functional images were acquired by using a multislice echo planar imaging sequence with 36 transversal slices in descending order, slice thickness 3.2 mm, TR = 2,100 ms, TE = 30 ms, field of view = 205 x 205 mm2, 64 x 64 matrix, flip angle = 90°, bandwidth 1562 Hz/Px. In total for this task, 230 functional volumes of 36 slices were acquired for each participant for a total of 8280 slices per participant. For our sample of ninety-eight participants, 22’540 volumes were acquired for a grand total of 811’440 slices. For the sample-specific data (N=23), 5290 volumes were acquired for a grand total of 190’440 slices.

Wholebrain data analysis

Species categorization task analysis within the temporal voice areas

Functional images were analyzed with Statistical Parametric Mapping software (SPM12, Wellcome Trust Centre for Neuroimaging, London, UK). Preprocessing steps included realignment to the first volume of the time series, slice timing, normalization into the Montreal Neurological Institute [60] (MNI) space using the DARTEL toolbox [61] and spatial smoothing with an isotropic Gaussian filter of 8 mm full width at half maximum. To remove low-frequency components, we used a high-pass filter with a cutoff frequency of 1/128Hz. Three general linear models were used to compute first-level statistics, in which each event was modeled by using a boxcar function and was convolved with the hemodynamic response function, time-locked to the onset of each stimulus. In model 1, separate regressors were created for all trials of each species (Species factor: human, chimpanzee, bonobo, macaque vocalizations) and two covariates of no-interest each (mean fundamental frequency and mean energy of each species) for a total of 12 regressors. Finally, six motion parameters were included as regressors of no interest to account for movement in the data and our design matrix therefore included a total of 18 columns plus the constant term. The species regressors were used to compute simple contrasts for each participant, leading to separate main effects of human, chimpanzee, bonobo, and macaque vocalizations. Covariates were set to zero in order to model them as no-interest regressors. In model 2, separate regressors were created for all trials of each species (Species factor: human, chimpanzee, bonobo, macaque vocalizations) and one covariate of no-interest for each species (acoustic distance for each species relative to human voice stimuli) for a total of 8 regressors. Finally, six motion parameters were included as regressors of no interest to account for movement in the data and our design matrix therefore included a total of 14 columns plus the constant term. The species regressors were used to compute simple contrasts for each participant, leading to separate main effects of human, chimpanzee, bonobo, and macaque vocalizations excluding acoustic distance (the covariate was set to zero in order to model it as ‘of no-interest’). In model 3, separate

regressors were created for all trials of each species (Species factor: human, chimpanzee, bonobo, macaque vocalizations) and six covariates of no-interest each (vocalization loudness, intensity, change in spectrum, bandwidth contour of the second formant (F2), power of the fundamental frequency (F0) and finally the difference in intensity contour) for a total of 28 regressors. Finally, six motion parameters were included as regressors of no-interest to account for movement in the data and our design matrix therefore included a total of 34 columns plus the constant term. The species regressors were used to compute simple contrasts for each participant, leading to separate main effects of human, chimpanzee, bonobo, and macaque vocalizations. Covariates were set to zero in order to model them as no-interest regressors.

For each first-level model, each of their respective four simple contrasts were then taken to two flexible factorial second-level analyses. For all of these second-level analyses there were two factors: Participants factor (independence set to yes, variance set to unequal) and the Species factor (independence set to no, variance set to unequal). For these analyses and to be consistent, we only included participants who were above chance level (25%) in the species categorization task (N=23). Brain region labelling was defined using xjView toolbox (http://www.alivelearn.net/xjview) implementing the Automated anatomical labelling ('aal3') atlas [62]. All neuroimaging activations were thresholded in SPM12 by using a voxelwise false discovery rate (FDR) correction at $p<.05$ and an arbitrary cluster extent of k>10 voxels to remove very small clusters of activity. The TVA highlighted in the figures are therefore only visual outlines of these regions, but no region-of-interest (ROI) analysis was performed here in order to maximize data representativeness—which is impacted negatively by ROI analysis.

Temporal voice areas localizer task

Functional images were analyzed with Statistical Parametric Mapping software (SPM12, Wellcome Trust Centre for Neuroimaging, London, UK). Preprocessing steps included realignment to the first volume of the time series, slice timing, normalization into the Montreal Neurological Institute [60] (MNI) space using the DARTEL toolbox [61] and spatial smoothing

with an isotropic Gaussian filter of 8 mm full width at half maximum. To remove low-frequency components, we used a high-pass filter with a cutoff frequency of 1/128Hz. A general linear model was used to compute first-level statistics, in which each block was modeled by using a block function and was convolved with the hemodynamic response function, time-locked to the onset of each block. Separate regressors were created for each condition (vocal and non-vocal; condition factor). Finally, six motion parameters were included as regressors of no interest to account for movement in the data. The condition regressors were used to compute simple contrasts for each participant, leading to a main effect of vocal and non-vocal at the first-level of analysis: [1 0] for vocal, [0 1] for non-vocal. These simple contrasts were then taken to a flexible factorial second-level analysis in which there were two factors: Participants factor (independence set to yes, variance set to unequal) and the Condition factor (independence set to no, variance set to unequal). An identical analysis architecture was used to delineate more specific subregions of the TVA according to the type of non-vocal material, with categories: animal sounds, music, nature sounds, artificial noise sounds. Timing onsets of these newly created “events”—including their duration—within each non-vocal block of the task were determined and each main effect contrast computed at the first-level was then taken to a second-level flexible factorial analysis with settings identical to the above. All neuroimaging activations were thresholded in SPM12 by using a voxelwise false discovery rate (FDR) correction at $p<.05$ and an arbitrary cluster extent of $k>10$ voxels to remove very small clusters of activity. Activation outline for vocal > non-vocal—the contrast revealing the TVA—was precisely delineated for the N=98 and the N=23 samples and overlaid on brain displays of the species categorization task (see Fig.S8). TVA subregions according to the auditory material (specific non-vocal condition blocks material) are reported in Fig.S9.

Acknowledgements

We thank the Swiss National Science foundation (SNSF) for supporting this interdisciplinary project (grant CR13I1_162720 / 1 to DG-TG), the National Centre of Competence in Research

(NCCR) (51NF40-104897 to DG) hosted by the Swiss Center for Affective Sciences, as well as the Fondation Ernst et Lucie Schmidheiny supporting CD. TG was additionally supported by a grant of the SNSF during the final editing of this article (grant PCEFP1_186832). We also thank Katie Slocombe and Zanna Clay for providing the nonhuman primates auditory stimuli and Daphne Bavelier for her advice on the design of the species categorization task. We would like also to acknowledge the staff of the Brain and Behavior Laboratory at the University of Geneva where all data were acquired.

Author contributions

LC designed the task, acquired part of the data, analyzed the data, designed the figures, wrote and edited the manuscript. CD designed the task, acquired the data, analyzed the data, wrote and edited parts of the manuscript. TG provided theoretical background and edited the manuscript. DG helped design the task, the analyses and edited the manuscript.

Declaration of interests

The authors declare no competing interests whatsoever.

Data availability statement

All data, stimuli and codes used in this article are available in the FAIR-compliant open repository YARETA (URL: https://doi.org/10/g8wkrb).

References

1. Ogawa, S., et al., *Brain magnetic resonance imaging with contrast dependent on blood oxygenation.* proceedings of the National Academy of Sciences, 1990. **87**(24): p. 9868-9872.
2. Belin, P., et al., *Voice-selective areas in human auditory cortex.* Nature, 2000. **403**(6767): p. 309-312.

3. Aglieri, V., et al., *Functional connectivity within the voice perception network and its behavioural relevance.* NeuroImage, 2018. **183**: p. 356-365.
4. Kriegstein, K.V. and A.-L. Giraud, *Distinct functional substrates along the right superior temporal sulcus for the processing of voices.* Neuroimage, 2004. **22**(2): p. 948-955.
5. Pernet, C.R., et al., *The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices.* Neuroimage, 2015. **119**: p. 164-174.
6. Frühholz, S., et al., *Neural oscillations in human auditory cortex revealed by fast fMRI during auditory perception.* NeuroImage, 2020. **207**: p. 116401.
7. Latinus, M. and P. Belin, *Human voice perception.* Current Biology, 2011. **21**(4): p. R143-R145.
8. Zäske, R., B.A.S. Hasan, and P. Belin, *It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content.* cortex, 2017. **94**: p. 100-112.
9. Roswandowitz, C., et al., *Obligatory and facultative brain regions for voice-identity recognition.* Brain, 2017. **141**(1): p. 234-247.
10. Lahnakoski, J.M., et al., *Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception.* Frontiers in human neuroscience, 2012. **6**: p. 233.
11. Ethofer, T., et al., *Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI.* Cerebral cortex, 2012. **22**(1): p. 191-200.
12. Witteman, J., V.J. Van Heuven, and N.O. Schiller, *Hearing feelings: a quantitative meta-analysis on the neuroimaging literature of emotional prosody perception.* Neuropsychologia, 2012. **50**(12): p. 2752-2763.
13. Latinus, M., F. Crabbe, and P. Belin, *Learning-induced changes in the cerebral processing of voice identity.* Cerebral Cortex, 2011. **21**(12): p. 2820-2828.
14. Latinus, M., et al., *Norm-based coding of voice identity in human auditory cortex.* Current Biology, 2013. **23**(12): p. 1075-1080.
15. Charest, I., et al., *Cerebral processing of voice gender studied using a continuous carryover fMRI design.* Cerebral Cortex, 2013. **23**(4): p. 958-966.
16. Grossmann, T., et al., *The developmental origins of voice processing in the human brain.* Neuron, 2010. **65**(6): p. 852-858.
17. Kisilevsky, B.S., et al., *Effects of experience on fetal voice recognition.* Psychological science, 2003. **14**(3): p. 220-224.
18. Hüppi, P.S., *Cortical development in the fetus and the newborn: advanced MR techniques.* Topics in Magnetic Resonance Imaging, 2011. **22**(1): p. 33-38.
19. Andics, A., et al., *Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI.* Current Biology, 2014. **24**(5): p. 574-578.
20. Perrodin, C., et al., *Voice cells in the primate temporal lobe.* Current Biology, 2011. **21**(16): p. 1408-1415.
21. Petkov, C.I., et al., *A voice region in the monkey brain.* Nature neuroscience, 2008. **11**(3): p. 367-374.
22. Fecteau, S., et al., *Is voice processing species-specific in human auditory cortex? An fMRI study.* Neuroimage, 2004. **23**(3): p. 840-848.
23. Belin, P., *Voice processing in human and non-human primates.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2006. **361**(1476): p. 2091-2107.
24. Belin, P., et al., *Human cerebral response to animal affective vocalizations.* Proceedings of the Royal Society B: Biological Sciences, 2008. **275**(1634): p. 473-481.
25. Fritz, T., et al., *Human behavioural discrimination of human, chimpanzee and macaque affective vocalisations is reflected by the neural response in the superior temporal sulcus.* Neuropsychologia, 2018. **111**: p. 145-150.
26. Linnankoski, I., et al., *Recognition of emotions in macaque vocalizations by children and adults.* Language & Communication, 1994. **14**(2): p. 183-192.
27. Bodin, C., et al., *Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques.* Current Biology, 2021. **31**(21): p. 4839-4844. e4.

28. Grawunder, S., et al., *Higher fundamental frequency in bonobos is explained by larynx morphology.* Current Biology, 2018. **28**(20): p. R1188-R1189.
29. Slocombe, K.E. and K. Zuberbühler, *Chimpanzees modify recruitment screams as a function of audience composition.* Proceedings of the National Academy of Sciences, 2007. **104**(43): p. 17228-17233.
30. Debracque, C., et al., *Humans recognize affective cues in primate vocalizations: acoustic and phylogenetic perspectives.* Scientific reports, 2023. **13**(1): p. 10900.
31. Gruber, T. and Z. Clay, *A comparison between bonobos and chimpanzees: A review and update*. Evolutionary Anthropology: Issues, News, and Reviews, 2016. **25**(5): p. 239-252.
32. Titze, I., T. Riede, and T. Mau, *Predicting achievable fundamental frequency ranges in vocalization across species.* PLoS computational biology, 2016. **12**(6): p. e1004907.
33. Kelly, T., et al., *Adult human perception of distress in the cries of bonobo, chimpanzee, and human infants.* Biological Journal of the Linnean Society, 2017. **120**(4): p. 919-930.
34. Ceravolo, L., et al., *Frontal mechanisms underlying primate calls recognition by humans.* Cerebral Cortex Communications, 2023. **4**(4): p. tgad019.
35. De Maesschalck, R., D. Jouan-Rimbaud, and D.L. Massart, *The mahalanobis distance*. Chemometrics and intelligent laboratory systems, 2000. **50**(1): p. 1-18.
36. Perrodin, C., et al., *Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex.* Proceedings of the National Academy of Sciences, 2015. **112**(1): p. 273-278.
37. Engel, L.R., et al., *Different categories of living and non-living sound-sources activate distinct cortical networks.* Neuroimage, 2009. **47**(4): p. 1778-1791.
38. Lewis, J.W., et al., *Distinct cortical pathways for processing tool versus animal sounds.* Journal of neuroscience, 2005. **25**(21): p. 5148-5158.
39. Simmons, W.K., et al., *The selectivity and functional connectivity of the anterior temporal lobes.* Cerebral Cortex, 2010. **20**(4): p. 813-825.
40. Zahn, R., et al., *Social concepts are represented in the superior anterior temporal cortex.* Proceedings of the National Academy of Sciences, 2007. **104**(15): p. 6430-6435.
41. Bestelmeyer, P.E., P. Belin, and M.-H. Grosbras, *Right temporal TMS impairs voice detection.* Current Biology, 2011. **21**(20): p. R838-R839.
42. Mellem, M.S., et al., *Sentence processing in anterior superior temporal cortex shows a social-emotional bias.* Neuropsychologia, 2016. **89**: p. 217-224.
43. Belin, P., C. Bodin, and V. Aglieri, *A "voice patch" system in the primate brain for processing vocal information?* Hearing research, 2018. **366**: p. 65-74.
44. Giamundo, M., et al., *A population of neurons selective for human voice in the monkey brain.* Proceedings of the National Academy of Sciences, 2024. **121**(25): p. e2405588121.
45. Perrodin, C., et al., *Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices.* Journal of Neuroscience, 2014. **34**(7): p. 2524-2537.
46. Cauzinille, J., et al., *Applying machine learning to primate bioacoustics: Review and perspectives.* American Journal of Primatology, 2024. **86**(10): p. e23666.
47. Hare, B., V. Wobber, and R. Wrangham, *The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression.* Animal Behaviour, 2012. **83**(3): p. 573-585.
48. Perelman, P., et al., *A molecular phylogeny of living primates.* PLoS Genet, 2011. **7**(3): p. e1001342.
49. Dewi, S.P., A.L. Prasasti, and B. Irawan. *The study of baby crying analysis using MFCC and LFCC in different classification methods*. in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*. 2019. IEEE.
50. Jagtap, S.S., P.K. Kadbe, and P.N. Arotale. *System propose for Be acquainted with newborn cry emotion using linear frequency cepstral coefficient*. in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. 2016. IEEE.
51. Hartig, R., et al., *A framework and resource for global collaboration in non-human primate neuroscience*. Current research in neurobiology, 2023. **4**: p. 100079.
52. Milham, M., et al., *Toward next-generation primate neuroscience: a collaboration-based strategic plan for integrative neuroimaging.* Neuron, 2022. **110**(1): p. 16-20.

53. Kang, H., *Sample size determination and power analysis using the G* Power software*. Journal of educational evaluation for health professions, 2021. **18**.
54. Belin, P., S. Fillion-Bilodeau, and F. Gosselin, *The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing.* Behavior research methods, 2008. **40**(2): p. 531-539.
55. Ferdenzi, C., et al., *Voice attractiveness: Influence of stimulus duration and type.* Behavior research methods, 2013. **45**(2): p. 405-413.
56. Ceravolo, L., S. Frühholz, and D. Grandjean, *Modulation of auditory spatial attention by angry prosody: an fMRI auditory dot-probe study.* Frontiers in neuroscience, 2016. **10**: p. 216.
57. Team, R., *RStudio: integrated development for R.* RStudio, Inc., Boston, MA URL http://www. rstudio. com, 2015. **42**: p. 14.
58. Bates, D., *Fitting linear mixed-effects models using lme4.* arXiv preprint arXiv:1406.5823, 2014.
59. Eyben, F., et al., *The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing.* IEEE transactions on affective computing, 2015. **7**(2): p. 190-202.
60. Collins, D.L., et al., *Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space.* Journal of computer assisted tomography, 1994. **18**(2): p. 192-205.
61. Ashburner, J., *A fast diffeomorphic image registration algorithm.* Neuroimage, 2007. **38**(1): p. 95-113.
62. Rolls, E.T., et al., *Automated anatomical labelling atlas 3.* Neuroimage, 2020. **206**: p. 116189.

Tables

Table 1: Activations, cluster size and coordinates for each contrast of interest of model 1 (mean of vocalization fundamental frequency and energy as trial-level covariates of no-interest) in the sample-specific temporal voice areas, wholebrain voxelwise *p*<.05 FDR corrected, k>10.

		MNI coordinates				
Region label	Hemisphere	X	Y	Z	T value	Cluster size (voxels)
Chimpanzee > Human, Bonobo, Macaque						
Superior temporal gyrus ant[1]	L	-52	-2	-12	4.84	91
Superior temporal gyrus mid	L	-50	-8	-12	4.12	
Superior temporal gyrus ant[2]	R	54	0	-12	3.63	18
Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus ant[3]	R	56	-2	-8	4.89	74
Superior temporal gyrus mid	L	-50	-8	-12	4.12	
Superior temporal gyrus ant[5]	L	-52	-2	-12	5.18	78
Chimpanzee > Human						
Superior temporal gyrus ant[4]	L	-50	-4	-12	3.36	74
Human, Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus mid	L	-58	-12	0	9.21	1664
Superior temporal sulcus post	*L*	*-58*	*-42*	*-2*	*4.30*	
Superior temporal sulcus mid	*L*	*-54*	*-24*	*-6*	*3.92*	
Superior temporal gyrus mid	R	56	-8	2	8.86	3051
Superior temporal gyrus ant	*R*	*60*	*-2*	*-8*	*7.22*	
Superior temporal sulcus mid	*R*	*56*	*-10*	*-14*	*6.37*	
Superior temporal sulcus post	*R*	*58*	*-40*	*-4*	*3.75*	
Human > Chimpanzee						
Supramarginal gyrus	R	56	-42	28	8.41	5941
Superior temporal gyrus mid	*R*	*56*	*-10*	*2*	*8.09*	
Superior temporal gyrus post	*R*	*58*	*-46*	*14*	*6.76*	
Superior temporal sulcus post	*R*	*66*	*-32*	*0*	*6.50*	
Middle temporal gyrus mid	*R*	*68*	*-20*	*-10*	*5.52*	
Superior temporal sulcus ant	*R*	*66*	*-14*	*-6*	*5.18*	
Middle temporal gyrus ant	*R*	*60*	*4*	*-20*	*4.45*	
Supramarginal gyrus	L	-56	-40	34	7.51	6109
Superior temporal gyrus mid	*L*	*-50*	*-18*	*4*	*7.37*	
Superior temporal gyrus post	*L*	*-56*	*-52*	*18*	*6.53*	
Middle temporal gyrus mid	*L*	*-64*	*-22*	*-10*	*6.10*	
Middle temporal gyrus post	*L*	*-62*	*-44*	*-6*	*4.87*	
Superior temporal sulcus mid	*L*	*-54*	*-32*	*-4*	*4.76*	

ant: anterior; mid: central part; post: posterior.
[1-5]Figure 2 cluster labels

Table 2: Activations, cluster size and coordinates for each contrast of interest of model 2 (inter-species vocalization acoustic distance as trial-level covariate of no-interest) in the sample-specific temporal voice areas, wholebrain voxelwise *p*<.05 FDR corrected, k>10.

Region label	Hemisphere	MNI coordinates X	Y	Z	T value	Cluster size (voxels)
Chimpanzee > Human, Bonobo, Macaque						
Superior temporal gyrus ant[6]	L	-52	-2	-12	4.75	74
Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus ant[7]	L	-54	-2	-12	5.15	31
Superior temporal gyrus ant[9]	R	56	-2	-8	4.68	30
Chimpanzee > Human						
Inferior temporal gyrus ant	L	-36	-2	-36	5.54	1310
Hippocampus	*L*	*-40*	*-24*	*-18*	*4.50*	
Superior temporal gyrus ant[8]	*L*	*-50*	*-4*	*-12*	*3.30*	*49*
Human, Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus mid	L	-58	-12	0	9.35	1112
Superior temporal gyrus mid	*L*	*-50*	*-18*	*4*	*7.27*	
Superior temporal sulcus ant	*L*	*-54*	*-2*	*-12*	*4.19*	
Superior temporal gyrus mid	R	56	-8	2	8.52	1619
Superior temporal gyrus ant	*R*	*58*	*-2*	*-10*	*6.66*	
Superior temporal sulcus mid	*R*	*56*	*-10*	*-14*	*6.09*	
Superior temporal sulcus post	*R*	*58*	*-40*	*-4*	*3.75*	
Superior temporal gyrus post	R	58	-46	14	5.32	641
Human > Chimpanzee						
Supramarginal gyrus	R	56	-42	28	8.91	6411
Superior temporal gyrus mid	*R*	*56*	*-8*	*2*	*8.62*	
Superior temporal gyrus post	*R*	*58*	*-44*	*20*	*7.80*	
Middle temporal gyrus mid	*R*	*68*	*-20*	*-10*	*5.76*	
Superior temporal sulcus ant	*R*	*58*	*4*	*-20*	*5.00*	
Supramarginal gyrus	L	-62	-48	24	7.66	6527
Superior temporal gyrus mid	*L*	*-50*	*-18*	*6*	*7.50*	
Middle temporal gyrus mid	*L*	*-58*	*-32*	*-6*	*5.24*	
Superior temporal gyrus ant	*L*	*-54*	*0*	*2*	*5.18*	

ant: anterior; mid: central part; post: posterior.

[6-9]Figure 3 cluster labels

Table 3: Activations, cluster size and coordinates for each contrast of interest of model 3 (vocalization loudness, intensity, change in spectrum, F2 bandwidth contour, F0 power and intensity contour difference as trial-level covariate of no-interest) in the sample-specific temporal voice areas, wholebrain voxelwise *p*<.05 FDR corrected, k>10.

		MNI coordinates				
Region label	Hemisphere	X	Y	Z	T value	Cluster size (voxels)
Chimpanzee > Human, Bonobo, Macaque						
Superior temporal gyrus ant[10]	L	-52	-2	-12	4.54	72
Superior temporal gyrus ant[11]	R	54	0	-12	3.19	19
Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus ant[13]	L	-54	-2	-12	5.03	100
Superior temporal gyrus ant	*L*	*-58*	*-12*	*-2*	*4.98*	
Superior temporal gyrus ant	*L*	*-52*	*2*	*-12*	*4.38*	
Superior temporal gyrus ant[14]	R	58	-2	-12	4.97	91
Superior temporal sulcus ant	*R*	*54*	*-8*	*-12*	*4.30*	
Chimpanzee > Human						
Superior temporal gyrus ant[12]	L	-50	-4	-12	2.70	10
Superior temporal gyrus ant	*L*	*-48*	*-10*	*-12*	*2.60*	
Human, Chimpanzee > Bonobo, Macaque						
Superior temporal gyrus mid	L	-56	-14	2	10.00	1780
Superior temporal sulcus post	*L*	*-60*	*-42*	*-2*	*4.73*	
Superior temporal sulcus post	*L*	*-62*	*-34*	*-4*	*4.31*	
Superior temporal sulcus mid	*L*	*-54*	*-24*	*-6*	*4.27*	
Middle temporal gyrus ant	*L*	*-58*	*4*	*-22*	*4.19*	
Superior temporal sulcus mid	*L*	*-64*	*-28*	*-4*	*4.14*	
Superior temporal sulcus ant	*L*	*-58*	*-4*	*-14*	*4.10*	
Middle temporal sulcus ant	*L*	*-60*	*-4*	*-18*	*4.02*	
Superior temporal sulcus mid	*L*	*-50*	*-28*	*-8*	*3.74*	
Superior temporal gyrus mid	R	58	-8	2	8.75	3659
Superior temporal sulcus mid	*R*	*58*	*-16*	*-10*	*6.85*	
Superior temporal gyrus ant	*R*	*60*	*-2*	*-6*	*6.51*	
Superior temporal sulcus ant	*R*	*58*	*-2*	*-12*	*6.44*	
Middle temporal gyrus mid	*R*	*56*	*-10*	*-14*	*6.13*	
Superior temporal sulcus mid	*R*	*50*	*-24*	*-8*	*6.02*	
Superior temporal gyrus post	*R*	*58*	*-46*	*14*	*5.89*	
Superior temporal sulcus post	*R*	*48*	*-30*	*-6*	*5.80*	
Superior temporal gyrus post	*R*	*52*	*-44*	*14*	*5.36*	
Superior temporal sulcus mid	*R*	*64*	*-22*	*-8*	*5.27*	
Middle temporal gyrus post	*R*	*66*	*-32*	*-2*	*4.76*	
Human > Chimpanzee						
Superior temporal gyrus mid	L	-54	-14	4	8.55	5392
Supramarginal gyrus	*L*	*-60*	*-48*	*24*	*7.54*	
Superior temporal gyrus post	*L*	*-54*	*-58*	*18*	*6.88*	
Middle temporal gyrus mid	*L*	*-66*	*-22*	*-8*	*6.39*	

Middle temporal gyrus post	*L*	*-66*	*-44*	*-2*	*5.49*	
Supramarginal gyrus	R	56	-42	28	8.22	5250
Supramarginal gyrus	*R*	*56*	*-42*	*36*	*7.52*	
Superior temporal gyrus mid	*R*	*56*	*-8*	*2*	*7.50*	
Superior temporal gyrus post	*R*	*50*	*-48*	*22*	*6.23*	
Superior temporal gyrus post	*R*	*58*	*-46*	*14*	*6.22*	
Superior temporal gyrus mid	*R*	*50*	*-18*	*6*	*6.05*	
Superior temporal gyrus post	*R*	*60*	*-46*	*18*	*5.99*	
Middle temporal gyrus post	*R*	*70*	*-32*	*0*	*5.84*	
Middle temporal gyrus mid	*R*	*64*	*-22*	*-12*	*5.75*	
Middle temporal gyrus mid	*R*	*66*	*-20*	*-18*	*5.42*	

ant: anterior; mid: central part; post: posterior.

[10-14]Figure 4 cluster labels