Human temporal voice areas are sensitive to chimpanzee 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
Boulevard Pont-d'Arve 40
CH-1205 Geneva
Switzerland
e-mail: leonardo.ceravolo@unige.ch

Abstract

In recent years, research on voice processing, particularly the study of temporal voice areas (TVA), was dedicated almost exclusively to human voice. To characterize commonalities and differences regarding primate vocalization representations in the human brain, the inclusion of closely related primates, especially chimpanzees and bonobos, is needed. We hypothesized that commonalities would depend on both phylogenetic and acoustic proximity, with chimpanzees ranking the closest to *Homo*. Presenting human participants with four primate species vocalizations (rhesus macaques, chimpanzees, bonobos and humans) and taking into account acoustic distance or removing voxels explained solely by vocalization low-level acoustics, we observed within-TVA enhanced left and right anterior superior temporal gyrus activity for chimpanzee compared to all other species, and chimpanzee compared to human vocalizations. Our results provide evidence for a common neural basis in the TVA for the processing of phylogenetically and acoustically close vocalizations, namely those of humans and chimpanzees.

Introduction

The study of the cerebral mechanisms underlying speech and voice processing has gained steam since the early 2000s with the emergence of functional magnetic resonance imaging (fMRI)[1]. Voice-sensitive areas, generally referred to as ‘temporal voice areas’ (TVA), have been highlighted along the upper, superior part of the temporal cortex[2]. Since then, great effort has been put into better characterizing these TVA, with a specific focus on their spatial compartmentalization into functional subparts[3-5]. Repetitive transcranial magnetic stimulations over the right mid TVA lead to persistent voice detection impairment in a simple voice/non-voice discrimination task[6] and a rather large body of literature is aligned with the crucial role of the TVA in voice perception and processing[3,7-9]. Subparts of the TVA have also been directly linked to social perception[10], vocal emotion processing[11,12], voice identity[13,14] and gender[15] perception. The developmental axis of voice processing has also been studied in infants, revealing the existence of TVA as early as 7 but not 4 month-olds in the human brain[16] while *in utero* fetuses have been shown to be already able to recognize their parents’ voice[17]. With the constant development of brain imaging and analysis techniques[18], it is realistic to expect successful, though non-invasive, *in utero* ‘task-related’ voice perception fMRI results in the near future. Along the evolutionary axis, evidence for TVA or more generally voice-sensitive brain areas have emerged most notably for dogs[19] and monkeys[20,21] (*Macaca mulatta*), raising the questions of whether TVA are species-specific[22] and to which extent human and non-human primates share neural mechanisms enabling them to process conspecific vocalizations[23]. Less attention has however been devoted to paradigms presenting animal vocalizations to humans, and no study to date has ever reported human TVA activations for the processing of such auditory material, namely other animals’ vocalizations. Human processing of animal vocalizations has been studied using both monkey and cat material but no specific activations related to any of the species was observed[24]. Other studies have focused more specifically on phylogenetic distance, including as stimuli human great ape (chimpanzee, *Pan troglodytes*) and old-world-monkey (rhesus macaque, *Macaca mulatta*) vocalizations. Such studies could not identify species-specific brain activations in spite of the correct discrimination of chimpanzee affective vocalizations[25], and observed below[25] vs. above[26] chance discrimination of affective macaque vocalizations by human participants.

This scarce literature motivated the present study that aims at a reliable investigation of species-specific TVA activations in humans asked to categorize phylogenetically close and distant species' vocalizations while undergoing fast fMRI scanning. The importance of between-species acoustic differences and distance, especially fundamental frequency was also of major interest[27,28]. We therefore included vocalizations of our closet sister taxon, *Pan* (chimpanzees; bonobos, *Pan paniscus*), whose estimated split with *Homo* is only 6-8 million years ago as well as phylogenetically more distant species (cercopithecidae: rhesus macaques, with an estimated split with *Homo* 25 million years ago). In fact, any claim of human uniqueness for recruiting the TVA remains on hold and should be tested in light of these closely related species. Bonobo vocalizations are of particular interest, as this species is thought to have experienced evolutionary changes in their communication in part due to a neoteny process involving acoustic modifications (i.e., fundamental frequency)[27] even though they are as phylogenetically close to humans as chimpanzees[29]. Whether such changes would affect the abilities of human participants to recognize their calls should therefore be investigated in comparison to chimpanzee and rhesus macaque vocalizations. We therefore predicted: i) acoustic proximity for human and chimpanzee vocalizations, while more distance would separate those of bonobo and macaque vocalizations; ii) an overlap between brain networks of *Homo* and the *Pan* branch (chimpanzee, bonobo) but not the cercopithecidae (rhesus macaque) vocalizations; iii) shared and localized brain activations for the categorization of human and chimpanzee vocalizations extending to the TVA, depending on both phylogenetic proximity and acoustic distance. These hypotheses involve: a) a control of low-level acoustic differences, namely vocalization mean fundamental frequency and energy—included as trial-level covariates of 'no-interest' in the first neuroimaging statistical model; b) the inclusion of a measure of acoustic distance—included as trial-level covariate of 'interest' in a second neuroimaging statistical model.

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, leaving us with twenty-three participants (10 female, 13 male, mean age 24.65 years, SD 3.66). 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**). Therefore, eighteen human voices were selected and they were expressed by two male and two female actors, obtained from a nonverbal validated stimuli set of Belin and collaborators[30]. The eighteen selected chimpanzee, bonobo and rhesus macaque vocalizations contained single calls or call sequences produced by 6 to 8 different individuals in their natural environment. 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 naturality of the sounds[31] and to allow for low-level acoustic parameters of interest to be used in data modelling.

Experimental procedure and paradigm

Laying comfortably in a 3T scanner, participants listened to a total of seventy-two stimuli randomized and played binaurally using MRI compatible earphones at 70 dB SPL. 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 randomly assigned across participants. 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.

Temporal voice areas localizer task

Participants

One-hundred and fifteen right-handed, healthy, either native or highly proficient French-speaking participants (62 female, 54 male, mean age 25.34 years, SD 5.50) were included in this functional magnetic resonance task. Among these participants, twenty-two out of the twenty-three who performed the species categorization task were included (the temporal voice areas localizer task was not acquired for one of them due to technical issues). 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 Ethical Committee of the University of Geneva and was conducted according to the Declaration of Helsinki.

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 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 interblock interval was 8 s long.

Behavioral data analysis

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[32] Inc., Boston, MA, url: http://www.rstudio.com/). These data are reported in the supplementary materials (Fig.S1) since they 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.

Acoustic Mahalanobis distances

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)[33]. This 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. 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). Excluding the acoustical variables with the highest correlations ($r>.90$) to avoid redundancy of acoustic parameters, we retained 16 acoustic parameters.

We subsequently computed Mahalanobis distances to classify the 96 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**bc**).

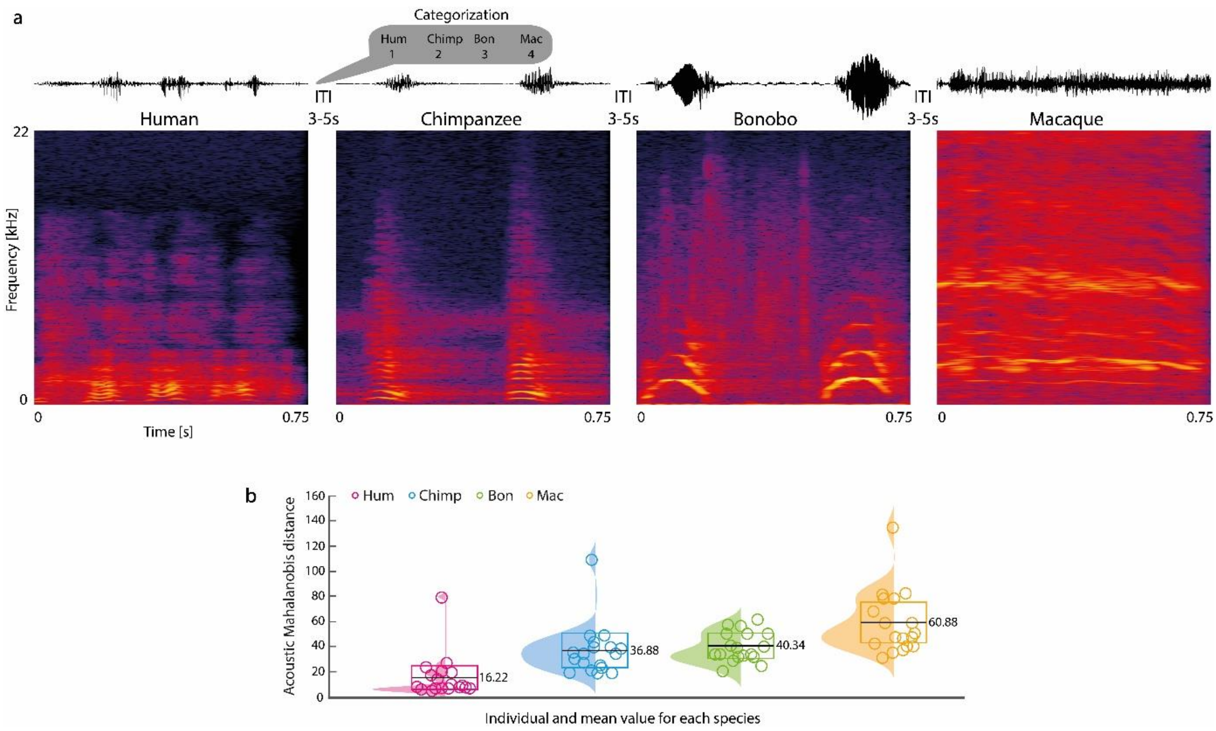

Fig.1: Timecourse of the species categorization task with stimuli example and acoustic distance data. a, Detail of the timecourse of four trials of the species categorization task in non-representative order, including waveform and spectrogram graphs for one example stimulus of each species. **b**, Scatter plot of the acoustic Mahalanobis distance data of each stimulus for each species including mean (numbers represent exact mean value) and box 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.

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 whole sample of one hundred and fifteen participants, 26450 volumes were acquired for a grand total of 952’200 slices.

Wholebrain data analysis

Species categorization task region-of-interest 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[33] (MNI) space using the DARTEL toolbox[34] 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 128 s. Two 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 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 including acoustic distance (the covariate was set to one in order to model it as 'of interest' regressor). For each model, each of their respective four simple contrasts were then taken to two flexible factorial second-level analyses. For both of these second-level analyses there were two factors: the 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=18). Brain region labelling was defined using xjView toolbox (http://www.alivelearn.net/xjview). 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.

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[33] (MNI) space using the DARTEL toolbox[34] 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 128 s. 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). All neuroimaging activations were thresholded in SPM12 by using a voxelwise family-wise error (FWE) correction at $p<.05$. Activation outline for vocal > nonvocal was precisely delineated and overlaid on brain displays of the species categorization task.

Results

Neuroimaging data within the temporal voice areas

We adopted a region-of-interest approach to uncover functional changes relative to species categorization and processing within the temporal voice areas, as delineated in our hypotheses. Low-level acoustics were used in two distinct models, namely vocalization mean energy and mean fundamental frequency (covariates of no-interest at the trial level, model 1) and a measure of acoustic distance (covariate of interest, model 2). We were particularly interested in brain activity while processing vocalizations of our closest relative (both acoustically and phylogenetically), the chimpanzee. The present study did not aim at uncovering wholebrain results underlying the processing of each species' vocalizations (see Fig.S2 and Fig.S3), although statistics presented in this section were computed with a voxelwise approach on the wholebrain for higher data reproducibility and generalizability.

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

In this model, we wanted to remove from species' processing brain activations the part of variance correlated with low-level acoustics of no-interest, namely mean voice energy and fundamental frequency. Brain activations common to human and chimpanzee vocalizations using the [human, chimpanzee > bonobo, macaque] contrast led to enhanced signal in the bilateral posterior, mid and anterior superior temporal cortex (Fig.2**abcd**, Table 1). Brain activity specific to chimpanzee vocalizations ([chimpanzee > human, bonobo, macaque]) led to enhanced activity in a cluster of the left anterior STG located within the temporal voice areas (Fig.2**c**). A similar result was observed when directly contrasting chimpanzee to human vocalizations ([chimpanzee > human]) in a slightly more medial area of the anterior STG, also located again within the voice-sensitive areas (Fig.2**g**, 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 either at the wholebrain level or within the TVA for both the [bonobo > human, chimpanzee, macaque] and the [macaque > human, chimpanzee, bonobo] contrasts.

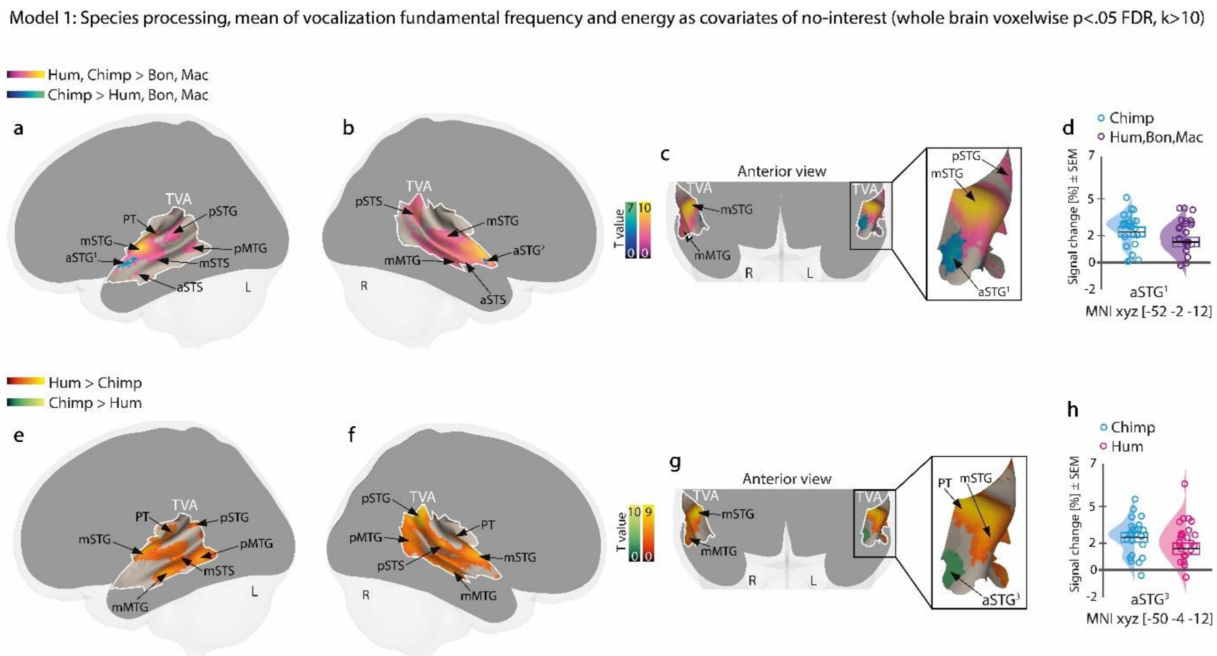

Fig.2: Wholebrain results when selectively contrasting processing of chimpanzee to other species' vocalizations with mean fundamental frequency and energy as trial-level covariates of no-interest. abc, Enhanced brain activity for human and chimpanzee compared to bonobo and macaque vocalizations (purple to yellow) on a sagittal view, overlaid with activity specific to chimpanzee

vocalizations (dark blue to green). **d**, Percentage of signal change for each individual and species 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 a more medial part of the anterior superior temporal gyrus (aSTG[3]) when contrasting chimpanzee to human vocalizations for each individual and species 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 (fundamental frequency 'F0' and mean energy of vocalizations). Data corrected for multiple comparison 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. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: temporal voice areas. 'a' prefix: anterior; 'm' prefix: mid; 'p' prefix: posterior; MTG: middle temporal gyrus; STG: superior temporal gyrus; STG: superior temporal gyrus; STS: superior temporal sulcus; PT: planum temporale; L: left; R: right.

Model 2: Effects of species processing with vocalization acoustic distance from human voice, per species, as covariate of interest at the trial level

In this second model, we wanted to add to species' processing brain activations the part of variance correlated with acoustic distance between each species and the human voice. Brain activations specific to human and chimpanzee vocalizations using the [human, chimpanzee > bonobo, macaque] contrast led to enhanced signal in the bilateral posterior, mid and anterior superior temporal cortex (Fig.3**abcd**, Table 2). Brain activity specific to chimpanzee vocalizations ([chimpanzee > human, bonobo, macaque]) led to enhanced activity in a cluster of the left anterior STG located within the temporal voice areas (Fig.3**c**). A similar result was observed when directly contrasting chimpanzee to human vocalizations ([chimpanzee > human]) in a slightly more medial area of the anterior STG, also located again within the voice-sensitive areas (Fig.3**g**, Table 2). 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.3**efg**, Table2). Again using this model, no voxels reached significance either at the wholebrain level or within the TVA for both the [bonobo > human, chimpanzee, macaque] and the [macaque > human, chimpanzee, bonobo] contrasts.

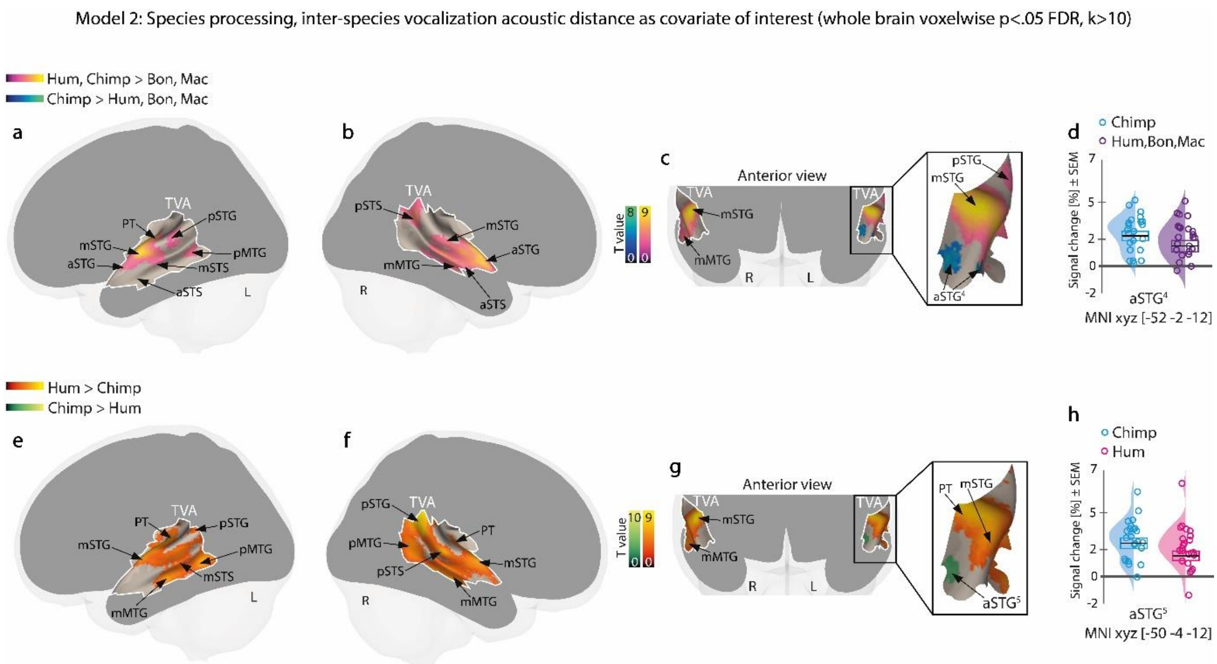

Fig.3: Wholebrain results when selectively contrasting processing of chimpanzee to other species' vocalizations with acoustic distance as trial-level covariate of interest. abc, Enhanced brain activity for human and chimpanzee compared to bonobo and macaque vocalizations (purple to yellow) on a sagittal view, overlaid with activity specific to chimpanzee vocalizations (dark blue to green). **d**, Percentage of signal change for each individual and species in the left anterior superior temporal gyrus (aSTG[4]). 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 a more medial part of the anterior superior temporal gyrus (aSTG[5]) when contrasting chimpanzee to human vocalizations for each individual and species with box plots representing mean value (black line) and the standard error of the mean with distribution fit. Brain activations are dependent of acoustic Mahalanobis distance between each species, see Methods for details. Data corrected for multiple comparison 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. Hum: human; Chimp: chimpanzee; Bon: bonobo; Mac: macaque. TVA: temporal voice areas. 'a' prefix: anterior; 'm' prefix: mid; 'p' prefix: posterior; MTG: middle temporal gyrus; STG: superior temporal gyrus; STG: superior temporal gyrus; STS: superior temporal sulcus; PT: planum temporale; L: left; R: right.

Discussion

The present study provides evidence of the sensitivity of the TVA to chimpanzee vocalizations, materialized by chimpanzee-specific enhanced activity in the left and right anterior STG. Second, our results highlight shared brain networks for the processing of both human voices and chimpanzee calls involving posterior, mid and anterior parts of bilateral superior temporal gyrus. Therefore, our results suggest that vocalizations expressed by another great ape species can also recruit subparts of the human temporal cortex normally dedicated to the processing of human voices, namely the anterior TVA.

Because we controlled our analyses for low-level acoustics and acoustic distance, we importantly demonstrate that similar TVA activity for the processing of human voices and chimpanzee vocalizations directly relate to both phylogenetic and acoustic proximity.

Often linked to the processing of conspecific vocalizations only (e.g., in humans[2]; macaques [20,21]; and dogs [19]), the present study questions the current view of TVA 'selectivity' showing that both human voices and chimpanzee calls enhance activity in the anterior TVA. Indeed, our neuroimaging analyses revealed the specific involvement of the left anterior STG when processing chimpanzee vocalizations. No specific results were observed for bonobo and macaque vocalizations, respectively. Furthermore, such anterior STG activity was also observed when a direct comparison between chimpanzee calls and human voice was made. This result adds to the specificity of subparts of the anterior TVA for the processing of species with human-like phylogeny and acoustics. Differences at the level of processing complexity between the two vocalizations could explain such observations. In fact, previous studies have shown the role of the left anterior STG and anterior STS in the conceptual representation of social context through the human voice[35-37]. Hence, our data could suggest that the anterior part of the left superior temporal cortex is recruited to process the social context of vocal stimuli expressed by human and chimpanzee species. Yet, this processing would be more automated for the perception of human voice due to our high exposition and expertise as humans, as opposed to chimpanzee calls that we do not encounter on a daily basis. For this reason, processing chimpanzee vocalizations and their context could trigger enhanced activity in the anterior superior temporal cortex, especially when compared to human voice[37].

Importantly, our data stress the importance of acoustic proximity between human and chimpanzee vocalizations: activity in the anterior STG and more generally in the anterior TVA would in fact depend on phylogenetic *and* acoustic proximity. If phylogenetic proximity was the only actor at play, bonobo calls should also trigger activity in the TVA, since they are similarly close to humans as chimpanzees as far as phylogeny is concerned. Concerning macaque calls, since they are both phylogenetically and acoustically distant from humans, the absence of TVA activity specific to this species was expected. This interpretation is strongly supported by the inclusion of acoustic Mahalanobis distance for each species compared to humans as covariate of interest. Additionally, previous research showed a higher pitch in young bonobo screams in comparison to chimpanzee and human baby cries[38], giving steam to the crucial role of acoustic Mahalanobis distance in our results. Therefore, it seems reasonable to hypothesize that TVA activity is not human-specific[2,6] *per se* but that it would instead be sensitive to the vocalizations of other primate species, provided that such vocalizations share sufficient acoustic (and phylogenetic) proximity with the human vocal signal.

We already mentioned that the interaction between phylogeny and acoustic distance or proximity would be at the origin of TVA enhancement for the processing of chimpanzee but not bonobo vocalizations. However, the absence of similar results for bonobo calls also support the evolutionary divergence of this peculiar species. In fact, according to the self-domestication hypothesis, bonobos would have evolved differently compared to chimpanzees due to selection against aggression[39]. Interestingly, differentiation in the evolutionary pathway of bonobos has affected both their behavior[29] and morphology. For instance, research has shown a shorter larynx in bonobos in comparison to chimpanzees resulting in a higher fundamental frequency in their calls[27], contributing to their greater acoustic distance from human or chimpanzee vocalizations. Putting into perspective the self-domestication hypothesis and our neuroimaging data, we can suppose that the calls of our common ancestor together with the other great apes 8 million year ago[40] would be close to the ones currently expressed by chimpanzees.

Taken together, our data allow us to draw the conclusion that both phylogenetic and acoustic proximity of primate vocalizations seem necessary to trigger activity in the human temporal voice areas. For this reason, anterior TVA activity was observed solely for the processing of chimpanzee but not bonobo or macaque vocalizations. Contrary to what was reported in recent years, we claim that the human TVA are also involved in the processing of heterospecific vocalizations, provided they share sufficient phylogenetic and acoustic proximity. Finally, our findings support a critical evolutionary continuity between the structure of human and chimpanzee vocalizations.

Acknowledgements

We thank the Swiss National Science foundation (SNSF) for supporting this interdisciplinary project (grants CR13I1_162720 / 1 to DG-TG), the National Centre of Competence in Research (NCCR) (51NF40-104897 - DG) hosted by the Swiss Center for Affective Sciences, as well as the Fondation Ernst et Lucie Schmidheiny supporting CD.

References

1 Ogawa, S., Lee, T.-M., Kay, A. R. & Tank, D. W. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences* **87**, 9868-9872 (1990).

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

3 Kriegstein, K. V. & Giraud, A.-L. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* **22**, 948-955 (2004).

4 Pernet, C. R. *et al.* The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* **119**, 164-174 (2015).

5 Aglieri, V., Chaminade, T., Takerkart, S. & Belin, P. Functional connectivity within the voice perception network and its behavioural relevance. *NeuroImage* **183**, 356-365 (2018).

6 Bestelmeyer, P. E., Belin, P. & Grosbras, M.-H. Right temporal TMS impairs voice detection. *Current Biology* **21**, R838-R839 (2011).

7 Frühholz, S., Trost, W., Grandjean, D. & Belin, P. Neural oscillations in human auditory cortex revealed by fast fMRI during auditory perception. *NeuroImage* **207**, 116401 (2020).

8 Latinus, M. & Belin, P. Human voice perception. *Current Biology* **21**, R143-R145 (2011).

9 Zäske, R., Hasan, B. A. S. & Belin, P. It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *cortex* **94**, 100-112 (2017).

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* **6**, 233 (2012).

11 Ethofer, T. *et al.* Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cerebral cortex* **22**, 191-200 (2012).

12 Witteman, J., Van Heuven, V. J. & Schiller, N. O. Hearing feelings: a quantitative meta-analysis on the neuroimaging literature of emotional prosody perception. *Neuropsychologia* **50**, 2752-2763 (2012).

13 Latinus, M., Crabbe, F. & Belin, P. Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex* **21**, 2820-2828 (2011).

14 Latinus, M., McAleer, P., Bestelmeyer, P. E. & Belin, P. Norm-based coding of voice identity in human auditory cortex. *Current Biology* **23**, 1075-1080 (2013).

15 Charest, I., Pernet, C., Latinus, M., Crabbe, F. & Belin, P. Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cerebral Cortex* **23**, 958-966 (2013).

16 Grossmann, T., Oberecker, R., Koch, S. P. & Friederici, A. D. The developmental origins of voice processing in the human brain. *Neuron* **65**, 852-858 (2010).

17 Kisilevsky, B. S. *et al.* Effects of experience on fetal voice recognition. *Psychological science* **14**, 220-224 (2003).

18 Hüppi, P. S. Cortical development in the fetus and the newborn: advanced MR techniques. *Topics in Magnetic Resonance Imaging* **22**, 33-38 (2011).

19 Andics, A., Gácsi, M., Faragó, T., Kis, A. & Miklósi, Á. Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology* **24**, 574-578 (2014).

20 Perrodin, C., Kayser, C., Logothetis, N. K. & Petkov, C. I. Voice cells in the primate temporal lobe. *Current Biology* **21**, 1408-1415 (2011).

21 Petkov, C. I. *et al.* A voice region in the monkey brain. *Nature neuroscience* **11**, 367-374 (2008).

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

23 Belin, P. Voice processing in human and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 2091-2107 (2006).

24 Belin, P. *et al.* Human cerebral response to animal affective vocalizations. *Proceedings of the Royal Society B: Biological Sciences* **275**, 473-481 (2008).

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* **111**, 145-150 (2018).

26 Linnankoski, I., Laakso, M., Aulanko, R. & Leinonen, L. Recognition of emotions in macaque vocalizations by children and adults. *Language & Communication* **14**, 183-192 (1994).

27 Grawunder, S. *et al.* Higher fundamental frequency in bonobos is explained by larynx morphology. *Current Biology* **28**, R1188-R1189 (2018).

28 Slocombe, K. E. & Zuberbühler, K. Chimpanzees modify recruitment screams as a function of audience composition. *Proceedings of the National Academy of Sciences* **104**, 17228-17233 (2007).

29 Gruber, T. & Clay, Z. A comparison between bonobos and chimpanzees: A review and update. *Evolutionary Anthropology: Issues, News, and Reviews* **25**, 239-252 (2016).

30 Belin, P., Fillion-Bilodeau, S. & Gosselin, F. The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods* **40**, 531-539 (2008).

31 Ferdenzi, C. *et al*. Voice attractiveness: Influence of stimulus duration and type. *Behavior research methods* **45**, 405-413 (2013).

32 Team, R. RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL http://www. rstudio. com* **42**, 14 (2015).

33 Collins, D. L., Neelin, P., Peters, T. M. & Evans, A. C. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography* **18**, 192-205 (1994).

34 Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95-113 (2007).

35 Mellem, M. S., Jasmin, K. M., Peng, C. & Martin, A. Sentence processing in anterior superior temporal cortex shows a social-emotional bias. *Neuropsychologia* **89**, 217-224 (2016).

36 Simmons, W. K., Reddish, M., Bellgowan, P. S. & Martin, A. The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex* **20**, 813-825 (2010).

37 Zahn, R. *et al.* Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences* **104**, 6430-6435 (2007).

38 Kelly, T. *et al.* Adult human perception of distress in the cries of bonobo, chimpanzee, and human infants. *Biological Journal of the Linnean Society* **120**, 919-930 (2017).

39 Hare, B., Wobber, V. & Wrangham, R. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour* **83**, 573-585 (2012).

40 Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet* **7**, e1001342 (2011).

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 temporal voice areas, wholebrain voxelwise *p*<.05 FDR corrected, k>10.

		MNI coordinates				
Region label	Hemisphere	X	Y	Z	T value	Cluster size (voxels)
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*	
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
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*	
Chimpanzee > Human						
Superior temporal gyrus ant[3]	L	-50	-4	-12	3.36	74

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

[1]Figure 2 cluster label: aSTG[1]

[2]Figure 2 cluster label: aSTG[2]

[3]Figure 2 cluster label: aSTG[3]

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 interest) in the temporal voice areas, wholebrain voxelwise *p*<.05 FDR corrected, k>10.

		MNI coordinates				
Region label	Hemisphere	X	Y	Z	T value	Cluster size (voxels)
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
Chimpanzee > Human, Bonobo, Macaque						
Superior temporal gyrus ant[4]	L	-52	-2	-12	4.74	71
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*	
Chimpanzee > Human						
Inferior temporal gyrus ant	L	-36	-2	-36	5.54	1310
Hippocampus	*L*	*-40*	*-24*	*-18*	*4.50*	
Superior temporal gyrus ant[5]	*L*	*-48*	*-6*	*-14*	*3.29*	

ant: anterior; mid: central part; post: posterior.
[4]Figure 3 cluster label: aSTG[4]
[5]Figure 3 cluster label: aSTG[5]