PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

**Cite this article:** Bodin C, Belin P. 2019 Exploring the cerebral substrate of voice perception in primate brains. *Phil. Trans. R. Soc. B* **375**: 20180386. http://dx.doi.org/10.1098/rstb.2018.0386

Accepted: 26 August 2019

One contribution of 15 to a theme issue 'What can animal communication teach us about human language?'

**Subject Areas:**
evolution, neuroscience

**Keywords:**
language evolution, primates, voice patch system, voice perception, comparative neuroimaging

**Author for correspondence:**
Clémentine Bodin
e-mail: clementine.bodin@univ-amu.fr

THE ROYAL SOCIETY
PUBLISHING

# Exploring the cerebral substrate of voice perception in primate brains

Clémentine Bodin[1] and Pascal Belin[1,2]

[1]Institut de Neurosciences de la Timone, UMR 7289 Centre National de la Recherche Scientifique and Aix-Marseille Université, Marseille, France
[2]Département de Psychologie, Université de Montréal, Montréal, Canada

CB, 0000-0003-3394-0632

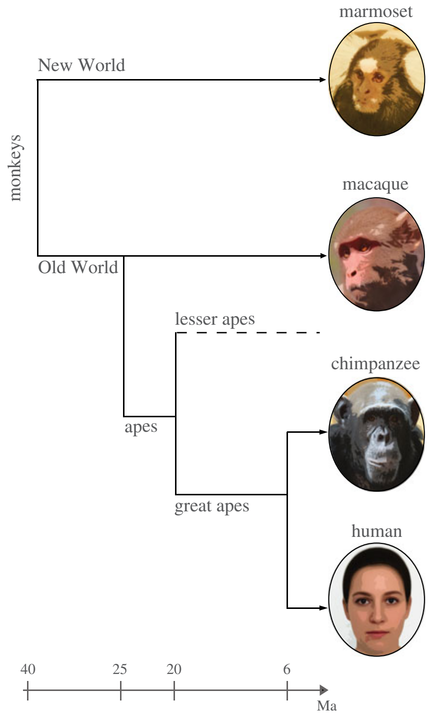
One can consider human language to be the Swiss army knife of the vast domain of animal communication. There is now growing evidence suggesting that this technology may have emerged from already operational material instead of being a sudden innovation. Sharing ideas and thoughts with conspecifics via language constitutes an amazing ability, but what value would it hold if our conspecifics were not first detected and recognized? Conspecific voice (CV) perception is fundamental to communication and widely shared across the animal kingdom. Two questions that arise then are: is this apparently shared ability reflected in common cerebral substrate? And, how has this substrate evolved? The paper addresses these questions by examining studies on the cerebral basis of CV perception in humans' closest relatives, non-human primates. Neuroimaging studies, in particular, suggest the existence of a 'voice patch system', a network of interconnected cortical areas that can provide a common template for the cerebral processing of CV in primates.

This article is part of the theme issue 'What can animal communication teach us about human language?'

## 1. Introduction

The question of language evolution is certainly one of the most challenging questions of our times. Since Charles Darwin's pioneering ideas almost two centuries ago [1], extensive research now supports a scenario where language has been gradually shaped from animal precursors instead of a sudden and recent emergence in the human lineage. Rather than a single encapsulated entity, language is considered as a set of cognitive components that may or may not be present at varying degrees in other animal species [2]. By building cognitive phylogenies of these components, we can thus capture crucial information about the emergence of language and the factors that influenced it, which could have been neglected if considered as a whole [3]. Recent breakthroughs in the field of vocal communication highlighted several of these components in non-human primates (NHP), including vocal learning [4,5], a rudimentary form of grammar [6–10], together with sequence learning abilities [11–13] and also the presence of a semantic content [14–16] and intentionally in their vocalizations [17].

What emerges as a connecting thread across these different components of vocal communication is the perception of voice, i.e. the processing of information carried by the caller's voice. This fundamental ability is a key element of communication for a wide range of species and is therefore particularly appropriate to bridge the gap between animal communication and human language. In particular, an efficient processing of conspecific vocal signals is crucial in a number of situations such as competition for territory, parental care, reproduction or predator avoidance. From these essential behaviours arises the need to infer the vocalizer's size, age, sex, group membership,



| left hemi | right hemi | contrast | awake | *n* | ref. |
|---|---|---|---|---|---|
| aTP | aTP | CV > phase scrambled sounds | ✗ | 6 | 118 |
| aSTP | aSTP, mSTP, pSTG | CV > max (Env + HV + scrambled controls) | ✓ | 2 | 107 |
| aTP, aSTP | aLS, aSTP, mSTP | CV > Env | | 2 | 108 |
| aSTP | aTP, aSTP | CV > scrambled CV | | 2 | |
| mSTP | mSTP | CV > scrambled CV | | 3 | 109 |
| - | aSTP | adaptation to speaker | | 2 | 107 |
| aSTP, aSTS | aSTP, pSTS/G | CV > HV+ Env | | 3 | 82 |
| - | mSTG, pMTG, ITG | proximal + distal CV > time reversed CV | ✗ | 3 | 119 |
| aSTG, mSTG/S, pSTS | aSTG, mSTG/S, pSTS | CV > Env | ✓ | 218 | 89 |
| - | mSTG/S | CV > auditory chimeras | | 22 | 90 |
| - | aSTG | adaptation to speaker | | 14 | 123 |
| aTP, mSTS, pSTS, deep pSTS | aTP, mSTS, pSTS | distant > close morphing | | 24 | 122 |
| aSTG, mSTG/S, pSTS/G | aSTG, mSTG/S, pSTS/G | CV > HV+ Env | | 3 | 82 |

**Figure 1.** Neuroimaging evidence of the temporal lobe regions showing sensitivity to conspecific voice (CV) in primates. A simplified phylogenetic tree of the species of interest is represented on the left (common marmoset, rhesus macaque, chimpanzee and human). On the right side, the table summarizes the regions found in neuroimaging studies (references in the last column) for the left and right hemispheres separately. The other columns indicate the main differences in the experimental procedures used in these studies (contrast of interest, anaesthesia and number of individuals). Three main categories of contrasts emerge: CV > non-CV (red), CV > acoustic controls (green) and identity sensitivity (orange). In black, we added recent results that we obtained in a comparative study between human and macaques. a, m, p, anterior, middle, posterior; Env, environmental sounds; HV, heterospecific voice; ITG, inferior temporal gyrus; LS, lateral sulcus; MTG, middle temporal gyrus; Ma, million years ago; STP, superior temporal plane; STG/S superior temporal gyrus/sulcus; TP, temporal pole.

individual identity or inner state, to adjust behaviour accordingly. This non-verbal content of voice can be distinguished from speech in human language and is also present in other animal vocal signals. Although vocalizations exhibit a certain specificity due to species-specific ecological constraints, the perceptual mechanisms involved in the processing of non-verbal information are probably more conserved. Since neither behaviour nor its brain substrate can be directly investigated from fossils, comparing humans to the closest extant species, NHP, can be used to infer the recent evolution of voice perception before the emergence of language. Here, particular attention is given to what is shared across primates rather than what separates them.

In the present paper, voice perception will refer primarily to the processing of information in conspecific vocalizations (CV) despite evidence that primates are also able to extract information from heterospecific vocalizations [18–22]. CV perception is assumed to include several processing stages that are organized in a similar way to those employed to extract information from faces [23,24], from distinguishing CV among non-CV sounds (initial 'structural encoding' stage) to processing different types of information contained in CVs (e.g. species, identity, gender, emotional state, etc.) in interacting but segregated functional pathways. Here, a particular interest will be given to the speaker/caller identity recognition as involving high-level processing stages in both systems. Since it is only in humans that voice perception abilities also include speech perception, this particular type of CV information will not be discussed.

We start by summarizing behavioural evidence of voice perception in primate species ranging from New World monkeys to apes (marmoset, macaque, chimpanzee and human). These species were selected based on the available neuroimaging literature on the cerebral basis of voice perception, developed in the second part of the paper. From this evidence, we develop our hypothesis of a conserved 'voice patch' system in primates dedicated to process CV information. This network of voice-sensitive areas can be compared to the face-processing system of the visual cortex [25–28]. More generally, we assume that cross-species similarities constitute evidence for homologous mechanisms inherited from a common ancestor and a gradual evolution of voice perception [29,30]. In the final section, we suggest future directions for comparative research on voice perception.

## 2. Behavioural evidence of conspecific voice perception

The primate auditory channel, together with vision, evolved as the main communication mode relative to the olfactory and chemical channels predominant in other animal species. Marmosets, macaques and chimpanzees diverge from the human lineage about 40 and 25 and 6 Ma, respectively (figure 1, left). These species have complex, albeit fairly different, social behaviours that can be regulated using a specific set of conspecific vocalizations. Humans, in particular, have remarkable abilities to extract verbal and also non-verbal information from CV, such as identity [31–33], gender [34] or personality [35]. This process is already operational in early infancy to recognize parents' voice [36–38] and the emotional content of voice [39]. Extraction of caller information is known to reflect the source-filter theory of voice production [40] where acoustic cues derived both from the larynx (fundamental frequency, f0) and the upper vocal tract (mostly formants or vocal tract resonances) are involved [31,41,42]. Surprisingly, it is only quite recently that a behavioural advantage of voice detection has been experimentally demonstrated in humans: both categorization and detection have been seen to improve when human voices (CV) are the targets compared to non-vocal sounds [43–45].

In macaques, belonging to the Old World monkeys, vocalizations are mainly produced to regulate and coordinate group activities using a rich call repertoire divided into a dozen of classes according to the social context and the motivational state [46–50]. One early series of studies in Japanese macaques employing category identification tasks showed that they can discriminate different CV from the 'coo' class of their repertoire and in a more efficient way than the compared species [51–55]. Nevertheless, further research on monkeys not intensively trained for this discrimination suggests a gradual transition through the different CV within the 'coo' and 'screams' classes rather than discrete boundaries [49,55–57]. In addition, the intrinsic variability of each class is not equal [58] and, hence, potentially conveys distinct information. As in humans, there is clear behavioural evidence that macaques can use identity information from CV [47,59–61]. They seem to rely on formant frequency information, in view of their ability to perceive formant frequency changes in playback trials [62] and associate these changes to differences in perceived body size [63,64]. Although macaques show moderate sexual dimorphism in body size, it is not clear whether macaques can recognize gender from CV.

Recently, there has been a renewed interest in New World monkeys such as marmosets that form a group of social and territorial species living in the upper canopy of South American forests. As highly vocal animals, they are in almost constant vocal communication, even in captivity [65,66], and their repertoire is now well-characterized acoustically [67–69]. This includes multiple vocalization types, from both simple to compound calls composed of sequences of simple calls [67]. This sequential production is highly variable in the temporal domain and can be modulated by the emotional state [70]. However, the processing of complex sequences remains limited in these species [71]. One long-distance contact call, the 'phee call', is produced by visually separated congeners in a reciprocal exchange known as 'antiphonal calling' [72]. This call contains potentially significant information on the caller's identity [68] or gender [73]. Recent evidence demonstrated that marmosets can process identity from these calls. The 'Virtual Monkey' approach, an automated playback technique, exploited the antiphonal calling behaviour: changes in the identity of synthetic phees were followed by changes in the frequency and latency of antiphonal calling by the subject, demonstrating identity discrimination [74]. Others have shown that changes in caller identity during playback can induce exploratory behaviour in marmosets [75].

Behavioural evidence of CV perception in great apes is much less documented, especially because they are more rarely studied in the laboratory. Chimpanzee vocalizations have been described in association with their facial displays as graded among acoustically defined call categories from simple, well defined, categories to more blended ones according to the internal motivations [76]. Considered separately, a large proportion of compound calls have been recorded, some of them exhibiting a different function than their simple counterpart [77]. A limited number of studies suggest that identity can be processed from chimpanzee CV. One of them found markers of individuality in the acoustics of one long-range call [78].

Two others employed vocal-to-facial matching tasks and reported correct identification of the caller from both long-range and short-range calls [79,80]. In contrast with this sparse literature on CV perception, higher-level properties of communication like intentionality and meaning have been described for chimpanzee vocalizations [15,17].

Thus, the evidence presented in this section indicates that CV of other NHP convey relevant information to their social interactions, as does non-verbal information in the human voice. The perceptual mechanisms involved in CV processing could then be conserved to some extent due to selective constraints inherent to this social life. Thanks to the recent advances in non-invasive neuroimaging techniques, the cerebral mechanisms of CV processing in primates can now be more precisely examined and compared with those of humans.

## 3. Cerebral evidence of conspecific voice processing in the temporal lobe

### (a) Clarifying where and what we are looking at

In primates, the anatomical organization of the auditory cortex reflects a functional hierarchy where information flows through primary regions (the 'core'), secondary regions ('belt' to 'parabelt') and auditory related fields, extending principally from the lateral sulcus (LS) to the superior temporal sulcus and to the extra-temporal regions [81]. The primary auditory cortex is located in Heschl's gyrus in humans but is deeply hidden in the LS, behind the parietal opercula, in monkeys. Different cytoarchitectonic parcellations of the auditory cortex have given rise to various nomenclatures. This potential source of confusion makes the interspecies comparison difficult to assess. Here, we chose to report the data from the literature (figures 1 and 2) based on simple morphological references. The medial-lateral axis is represented in order by the LS, the superior temporal plane (STP), the superior temporal gyrus (STG) and the STS. The positions on the rostral to caudal axis are noted as anterior (a), middle (m) and posterior (p). Another potential source of confusion in the cross-species comparison concerns the nature of the contrast used to highlight a CV sensitivity. Two principal methods exist: the first (CV versus non-CV) compares CV with categories of complex sounds such as heterospecific (HV) or environmental (Env) sounds; while the second (CV versus control sounds) compares CV with acoustically matched control sounds for which a specific set of acoustic parameters are kept unchanged (e.g. temporal envelop in phase-scrambled sounds). A third type of contrast specifically examines the sensitivity to identity information. Figure 1 summarizes the approximate anatomical location of contrasts reported in key selected studies for each species, classified into three main categories: CV > non-CV (red), CV > acoustic controls (green) and sensitivity to speaker/caller identity (orange). Each selected study is further developed in the following section.

### (b) Neuroimaging evidence of conspecific voice perception

The human cerebral substrate for voice perception is centred on secondary auditory cortical regions located bilaterally

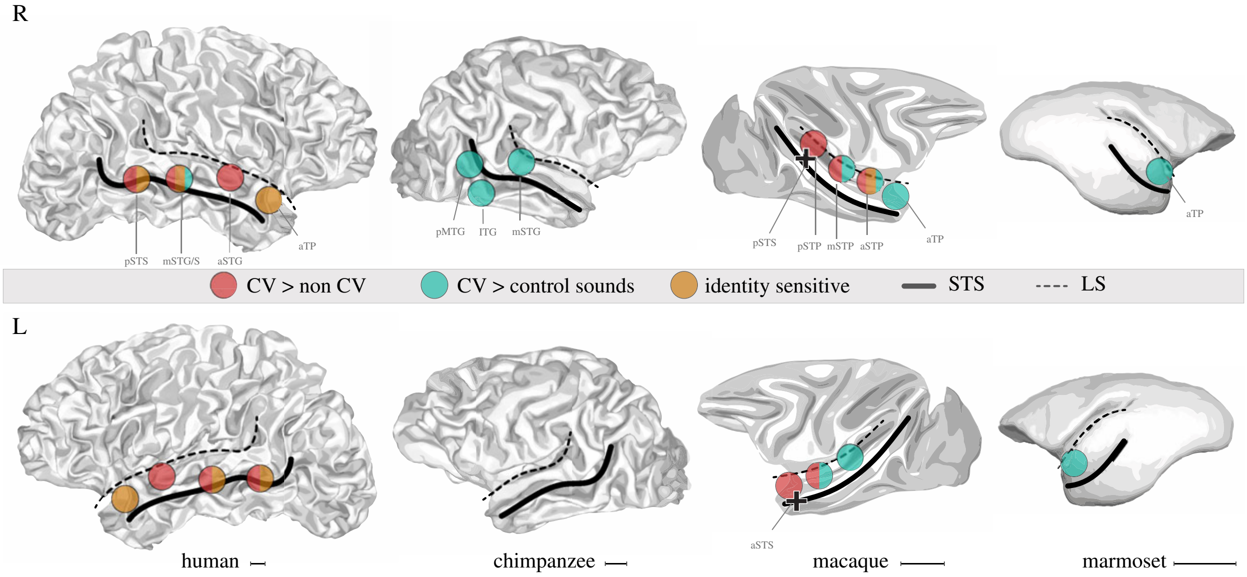


**Figure 2.** Organization of the conspecific voice (CV) patches along the temporal lobe in primates. The location of the regions is based on the selected neuroimaging studies in figure 1. Coloured spheres represent the location of a region more sensitive to CV than: non-CV (red), acoustic controls (green), a particular sensitivity to identity (orange) or to several of these contrasts (mixed colour). Black crosses illustrate the position of the newly identified CV-sensitive STS sites [82]. The regions are represented on white matter surfaces to reveal the inner part of the folds. The brain surfaces were modified from personal data for human, one individual image of the National Chimpanzee Brain Resource (NS092988) for chimpanzees, the NMT atlas white matter surface for macaques [83] and the segmented Brain/MINDS atlas for marmosets [84]. The black bar indicates 1 cm scale. L, left hemisphere; R, right hemisphere; a, m, p, anterior, middle, posterior; CV, conspecific voice; ITG, inferior temporal gyrus; LS, lateral sulcus; MTG, middle temporal gyrus; STP, superior temporal plane; STG/S, superior temporal gyrus/sulcus; TP, temporal pole.

along the superior temporal gyrus (STG) and sulcus (STS). These 'temporal voice areas' (TVAs) show greater fMRI signal in response to vocal sounds, whether they contain speech or not, compared to other categories of non-vocal sounds such as environmental sounds, amplitude-modulated noise [85–87], or to hetero-specific vocalizations (HV) [88]. Although their exact anatomical location in the temporal lobe varies considerably across individuals, a cluster analysis of voice-sensitivity peaks in several hundred subjects highlighted an organization in three 'voice patches' along STG/STS bilaterally (TVAa, TVAm and TVAp) [89] (figure 1). To test their selectivity, Agus *et al.* [90] used 'auditory chimeras' that matched vocal sounds for a large subsets of acoustical features and showed that the right TVAs (middle STG/STS) were preferentially activated only for the natural human voice. Very little is known about the cortical anatomy underlying the TVAs, on average centred bilaterally around the deep portion of the STS but exhibiting an important variability at the individual level [91]. Functional connectivity investigations have highlighted both intra- and inter-hemispheric connections between the different TVAs during passive voice listening [92] or a voice recognition task [87]. The three patches of the right STS were also shown to be structurally connected to each other [93]. Thus, while it seems clear that vocal information flows through the voice areas, little is known about their individual roles to date. Nevertheless, a growing body of literature indicates that they are recruited differently to process identity (reviewed in [27]).

As for behavioural evidence, our knowledge of the cerebral substrate for voice perception in NHPs mainly comes from the studies on rhesus macaques and marmosets, which are standard animal models in neuroscience. Nevertheless, pioneering studies were initially conducted in the auditory cortex of the squirrel monkey revealing CV sensitivity at the neuronal scale [94,95]. Electrophysiological recordings in awake macaques evidenced neurons in belt and parabelt areas that show a strong sensitivity to CVs [96–99] with latencies and selectivity increasing in the caudo-rostral direction towards the temporal pole [100,101]. Conversely, a strong sensitivity to CVs was found in the core areas of freely moving marmosets using electrophysiology [102–104]. Despite the crucial information provided by electrophysiology, direct comparison between species remains difficult because electrophysiological recordings in humans are mostly extracranial and, therefore, at low-spatial resolution. Recent advances in functional imaging hold much promise as they allow scanning of awake animals using protocols comparable to those of humans [105,106].

Petkov *et al.* [107] was the first fMRI study revealing macaque voice patches with responses analogous to the human TVAs, i.e. areas with significantly stronger response to macaque CV than other categories of environmental (CV versus non-CV) or control sounds (CV versus control sounds). Using an auditory cortex parcellation, a bilateral but mostly rightward activity of the anterior STP was found together with two clusters in the right middle STP (close to primary auditory cortex) and posterior STG. A right anterior patch was observed in the same location in two individuals and was still observed in anaesthetized monkeys, removing the possible effect of attention to sounds. It was then targeted in a subsequent single-cell recording electrophysiology study that revealed 'voice cells' in that anterior voice patch [97]. Surprisingly, no further consideration was given to the other voice patches.

Another study by Ortiz Rios *et al.* [108] also reported CV selectivity: compared to environmental sounds (CV versus non-CV) CV activated more bilateral anterior STP and right middle STP in addition to more anterior regions of the temporal pole. However, they recruited only the anterior patches when compared with their scrambled version (CV versus control sounds), altered spectrally and temporally.

Joly *et al*. [109] performed a pioneering comparative study in which human ($n = 20$) and macaque ($n = 3$) subjects were scanned in the same scanner while exposed to the same stimuli, including vocalizations of both species. Areas along the middle STP, close to A1, showed greater response to CV compared to their scrambled controls, but not compared to human vocalizations.

It has been suggested that the position of the anterior voice area described in Petkov *et al*. [107] is different from what was expected from human data [30,110]. Yet this comparison only focused on the most anterior patch observed in two macaques, leaving the possibility that the other, not yet examined, voice patches could be more similar to those of humans. In fact, taking into account all relevant studies and combining the different contrasts and nomenclatures used, several voice patches clearly emerge along the STG, in both hemispheres of the macaque brain. In a recent fMRI study, we scanned three awake macaques and three humans using a similar paradigm of passive auditory listening [82]. We found that CV elicited more activity in the bilateral anterior STP and the right posterior STG than environmental and heterospecific sounds, replicating earlier findings. For the first time, however, we found that the CV versus non-CV contrast also recruited the STS in its left anterior and right posterior portions (black crosses in figure 2), as in humans.

These novel results support the existence of several voice patches in monkeys; they also point towards a recruitment of STS domains. Indeed, this region has often been neglected as a classified multisensory region only. However, neuronal recordings already reported unimodal auditory responses in macaque STS [111–113]. Using face and voice stimuli, the anterior STS was shown to contain a more balanced proportion of auditory and visual unimodal neurons than the anterior STP; however, multisensory interaction was found to be equally prominent in both regions [113].

The past 5 years have seen increasingly rapid advances in the field of marmoset brain imaging [114–117], particularly as its small size is compatible with highfield (7T) rodent MRI allowing for higher signal and spatial resolution. Remarkably, a recent fMRI study in anaesthetized marmosets ($n = 6$) [118] revealed a gradient of sensitivity to CV along a caudal-ventral axis [118], with areas of high selectivity (here compared to scrambled controls) in the most anterior parts of the temporal lobe bilaterally. Such CV sensitivity accords with the anterior patches previously described in other species and with the neuronal gradient of selectivity of the macaque STP [101]. However, this is inconsistent with the other patches and previous reports of CV sensitivity in the marmoset primary cortex [102–104], located more dorsally. Although speculative, it is still possible that anaesthesia may lower the signal strength and allow only the regions that are least sensitive to the level of vigilance to be seen. One PET (positron emission tomography) study started to fill the gap between humans and monkeys by studying CV sensitivity in anaesthetized chimpanzees [119]. By grouping both proximal (short-range) and distal (long-range) categories of calls in contrast to their time-reversed controls (CV versus controls), the activity was lateralized on the right posterior temporal lobe, with peaks extending from the superior to the inferior temporal gyrus. However, this posterior activity was mainly driven by proximal calls and an important variability seemed to exist across conditions and individuals. Three methodological factors can have induced such variability and could explain the divergence with other species. First, PET acquisition was necessarily delayed after the auditory listening task during which the radioactive tracer is injected and this delay may have varied between individuals. Second, the auditory listening task was performed on freely moving animals without control of interaural differences. Finally, time-reversed controls may have involved different processes than other matched controls described in figure 1. For instance, in macaques, temporal inversion was shown to induce distinct behavioural responses depending on the acoustical symmetry of the call [120,121]. Hence, although a promising investigation, there is abundant room for further progress in determining the localization of voice areas in chimpanzees and it would be premature to interpret the observed lateralized activity.

Neuroimaging studies in the different primate species mentioned above are summarized in figure 2, which illustrates our current knowledge of the putative location of voice patches in the temporal lobe of humans, chimpanzees, macaques and marmosets.

### (c) Identity processing

Speaker identity processing in humans involves both temporal and prefrontal regions with strong right-hemispheric lateralization [122,123]. The most anterior voice-sensitive region of the right temporal lobe (right TVAa) in particular shows adaptation to speaker identity, i.e. smaller response to syllables spoken by a single speaker than to syllables spoken by multiple speakers [123]. A similar adaptation procedure evidenced a sensitivity to gender in this region [124]. Subsequently, Andics *et al*. [122] showed that bilateral TVAs are recruited by contrasting close versus distant identities morphed along an acoustic continuum. Nevertheless, all clusters in the right hemisphere, but only the deep left STS, were positively correlated with recognition performance. This is in line with the idea that unfamiliar voices are coded in the TVAs in a multidimensional acoustical 'voice-space' [31,125]. In particular, voices acoustically close to their (own-gender) average prototype elicit smaller TVA activity than more distinctive, acoustically dissimilar voices as a 'norm-based' coding [125]. It is worth noting that familiar and unfamiliar voice may be processed through dissociate pathways and thus make the prototype model more complex than expected (as reviewed in [33]). Andics *et al*. [122] also described interesting adaptation effects (response reduction to stimuli perceived as similar) along the STS axis: a short-term acoustic adaptation in the bilateral middle/posterior STS but a longer-term identity effect in the anterior temporal poles and the deep posterior STS. This may suggest that CV is primarily processed acoustically in middle and posterior TVAs then addressed to the anterior patches to extract identity-relevant information. A preponderant involvement of the right anterior region in that processing is suggested by adaptation mechanisms [123] and information-decoding procedures [122,123]. Contrastingly, the role of the right posterior region is nuanced by two contradictory lesion studies: from a cohort of patients, one classified it as an obligatory structure for voice-identity recognition [126], whereas another case study reported no effect on voice perception or identity recognition after a complete right pSTS resection [127]. As TVAs are functionally connected to each other but also to frontal regions during voice perception [92], it would be interesting to explore if compensatory mechanisms exist after such a lesion.

In macaques, the right anterior voice area described earlier exhibits the same speaker adaptation response to that observed in the human right anterior temporal lobe: greater response to CV from different individuals than CV from a single individual [107]. Some of the voice cells in that region also show some degree of caller selectivity, differentiating between individuals more than call type [113]. Advances in the processing of identity in these monkeys have been pushed forward by the discovery of multisensory regions integrating vocal and facial information and converging toward the temporal pole (reviewed in [128]; see section *Parallel with the face-processing system*).

Despite behavioural evidence that identity is relevant to marmosets, cerebral evidence of such processing is not obvious. In a recent PET study [75], extra-temporal regions were found to be associated with the perception of phee calls from a single subject compared to multiple subjects' stimuli. However, an adaptation effect could have been expected in their CV-sensitive anterior temporal poles [118] by contrasting these two conditions. To our knowledge, there is no experimental evidence to the neural coding of caller identity in chimpanzees.

### (d) Conspecific voice perception in extra-temporal regions

Neuroimaging studies also revealed extra-temporal regions, sensitive—although less consistently—to CV. Three bilateral patches were identified in the human frontal cortex as the 'frontal voice areas', more sensitive to voice than non-CV stimuli [92]. Especially, voice recognition performance was related to the functional connectivity into this fronto-temporal network in the right hemisphere [92]. The previously cited literature in macaques (figure 1) also reported CV sensitivity in parietal and ventro-lateral prefrontal cortex together with higher-level visual areas. In particular, electrophysiological recordings provide clear evidence that prefrontal cortex contains CV-sensitive neuronal populations; however, it is not yet clear whether they constitute higher-level areas in voice processing than those of the temporal lobe (reviewed in [98]). A study combining fMRI with neuronal microsimulation suggests that the level of processing of temporal areas cannot predict their effective connectivity with frontal areas [129]. Nevertheless, it is highly possible that anterior temporal and prefrontal cortices collaborate during CV processing as part of the same ventral pathway for complex sounds processing [130,131].

## 4. A voice patch system hypothesis

The literature overviewed here support the notion of a conserved system for the perception of CV in primates. Several discrete areas along the primate STG exhibit a specific sensitivity to voice compared to other categories of natural sounds or matched controls. These areas extend from belt to auditory related fields up to polar regions of the temporal lobe [81]. Future work is needed to provide a generic model linking the diverse results gathered so far in a coherent picture. First, it becomes essential to better determine the role of STS in the processing of CV in monkeys. Too long considered as only multimodal, this view is challenged by the presence of unimodal auditory neurons [111–113] and now by new data [82]. Second, additional neuroimaging investigations should be carried out in chimpanzees and marmoset to counterbalance the increasing amount of evidence in humans and macaques.

We further discuss three points in relation to the hypothesis of a potentially conserved voice patch system: (i) the voice patch system [132] is organized into a network of interconnected voice patches comparable to those of the face-processing system of visual cortex [25–28]; (ii) supporting an ecologically relevant ability in primates, it could be part of a broader social network in their brain; (iii) speech emerged from primitive roots including the voice patch system.

### (a) Analogies and interactions with the face-processing system

Full characterization of the voice patch system could allow further testing of the hypothesis of similar coding strategies for processing face and voice [24], converging across sensory modalities to extract, for example, identity information. Studies in humans, macaques and, more recently, marmosets together demonstrate the existence of a system of discrete, interconnected face-sensitive areas containing 'face cells' and supporting a series of increasingly abstract (identity-invariant) face representations [25–28,133]. The overall arrangement of face patches and their approximate distribution in the occipito-temporal cortex appears quite similar across species, although there is an overall shift of areas ventrally from the STS in humans compared to macaques [134]. This shift also seems to apply to a lesser extent to voice areas, with areas more deeply located in the STP in monkeys, but mostly around the STS in humans. Although we still do not know if there is a vocal equivalent of the view-point invariance gradient observed in the face network, adaptation and multivariate paradigms [107,123,135] suggest that invariant, word-independent, vocal identity information is mainly processed in anterior temporal regions in both humans and monkeys. Further work would determine whether NHP represent different callers in a measurable 'voice identity space' similar to humans and if they rely on norm-based coding strategies. From the face processing literature, behavioural evidence indicates that chimpanzees but not macaques rely on a norm-based coding of facial identity [136], whereas neuronal recordings suggest that this coding is also present in macaques [137]. As for the face patches, structural [93] and functional connections [92] were also found between the different voice patches, meaning that both systems could constitute an interconnected network.

Yet, in contrast to the notion of a conserved face-patch system across primate brains, behavioural studies report striking differences between great apes and monkeys in the way they encode faces [138]. An investigation of the cerebral basis of face processing in chimpanzees identified bilateral face-sensitive regions mostly localized around orbitofrontal and posterior STS areas [139], the latter being close to the voice-sensitive regions in this species (figure 2). Although new investigations are required to confirm this result, it indicates some inconsistencies across primates as observed for voice in our case (figure 2). Perhaps the key lies in the way voice and face processing systems interact with each other.

Several analogies between the two systems have been established from behaviour to cerebral bases between humans and monkeys [24,128]. The most direct evidence of their interaction is probably their connectivity. Blank *et al*. [93] reported direct structural connections between voice and ventral face areas in humans, and von Kriegstein *et al*. [140] a functional coupling of these regions during familiar speaker recognition. Both findings indicate a multimodal integration for identity recognition. But where could this integration take place? A model largely inspired by monkey data suggests a multimodal interaction that would converge towards the macaque temporal pole region to extract identity information [128], which is also supported by electrophysiological recordings and neuroimaging studies (figure 2). Extracting identity from voice in humans was previously shown to occur in the right anterior STS and over a large part of the TVAs by adaptation and morphing paradigms respectively [122,123] (figure 1). However, extracting identity from both face and voice could engage the right posterior STS in a modality general representation (e.g. [141–143]) by an audio-visual integration phenomenon (e.g. [141,144]), although others claim that this multimodal association is preferentially processed in inferior parietal areas [126]. Hence two identity processing streams emerge from the human literature. On the one hand, voice is processed along the STG/S antero-posterior axis using specific circuits depending on its familiarity (reviewed in [33]) to be fully recognized and stored [122] in the anterior temporal lobe. On the other hand, a multimodal pathway involving the posterior STS and inferior parietal areas integrates information carried by the face and voice. The latter lacked evidence in monkeys and could be inextricably linked to the social nature of the transmitted information such as emotions [144] and social stimuli perception [145,146] in humans.

Hence, although face and voice processing systems would be conserved in primates, their functional interactions may differ between species. Strikingly, the temporo-parietal junction (TPJ) has undergone an increasing cortical expansion from New World monkeys to great apes [147], along with a major restructuring of its anatomical and functional organization [148]. This expansion may have influenced a ventral shift of both the voice- and the face-patch systems from monkeys to humans and favoured the emergence of a new functional pathway combining multimodal social information in and beyond the pSTS in humans [148].

### (b) Integrated into a broader social network?

From a larger perspective, we assume that the vocal exchanges of information essential to the primates' social life may have shaped their voice processing system in a similar way. As an example, similar encoding strategies in their vocalizations [149] could allow a generalization across call categories but also across the vocal repertoires of other species [18–22]. To what extent has social life been able to determine these similarities in voice processing among primates? The social networks involved in the perception of interacting faces and in orofacial movement have been recently described in macaques, highlighting similarities with the speech-production system in humans [150]. At least two of these networks also interact with voice in humans and monkeys: the face-processing system (see previous section) and the lateral prefrontal cortex (e.g. [80,86,92,117]). In daily life, faces, voices and orofacial movements are in constant interaction during communication, which can suggest that the CV-processing system is also part of a broader social network in the primate brain. Importantly, whereas the cerebral substrates (see previous sections) and coding strategies [149] appear relatively conserved through evolution, voice and face processing systems are still permeable to the early social environment. For example, a study on face-deprived infant macaques [151] demonstrated that exposure to faces was necessary for the emergence of the face patches and a behavioural interest in those stimuli. In marmosets, parental feedback can affect vocal development and the acoustic structure of the infant's calls [152]. In humans, voice areas become functional between four and seven months of age [153], while a behavioural tuning to speech over heterospecific vocalizations seems to arise at three months of age [154]. The influence of the auditory environment on the development of CV-processing networks in primates remains to be investigated experimentally.

### (c) How did speech processing integrate into the voice patch system?

Evidencing similarities between human and NHP can help bridge the gap between animal communication and human language. However, this purpose is made challenging by the tight relationship between speech and voice in humans. The identity- and speech-processing pathways of voice perception constantly interact, in no small part because the same acoustical cues (formant frequencies) allow perceiving both what is being said [155] and who is speaking [156]. In the brain, speech stimuli elicit higher activity in the temporal voice areas than vocal sounds without speech [86]. Yet neuroimaging evidence suggests a general dissociation between speech-related processes (mostly left hemisphere) and speaker identification (mostly right hemisphere) [33,122,157–159]. Thus, the genius of human language could lie both in the interhemispheric dissociation of its components and in the ability to connect spatially segregated regions during communication. Inter-species comparison of the fibre bundles connecting temporal and frontal lobes, such as the arcuate fasciculus, provide useful information: from monkeys to humans, projections are increasingly widespread along the STG and in prefrontal areas [160]. This type of evidence supports the hypothesis that the left dorsal pathway, devoted to the processing of complex sounds in monkeys [131] and to the articulation and production of language in humans [159] has become more complex during the evolution of primates [13,131]. The temporo-parietal regions of the right hemisphere, however, may have evolved in a way that favours multimodal associations and the processing of high-level social information ([148]; see two previous sections). In this scenario, a conserved voice patch system could constitute one of the primitive foundations on which language would have emerged asymmetrically.

## 5. Future directions for comparative research in voice processing

Future work is needed to understand which features drive neuronal responses in the acoustically complex and variable CV and, thus, determine if NHP represent different callers in a measurable 'voice identity space' similar to that of humans. The 'Virtual Monkey' approach [74] exploiting the natural vocal behaviours of marmosets is a promising technique for this purpose and could be used on new species. A clearer appreciation of the different functional roles of each voice patch will provide crucial information on both an interaction with those of the face-processing system and with language-related areas. In the same line, neuroimaging studies using comparable paradigms across species should also be increasingly conducted to reliably estimate their similarities/differences. Finally, collaborative research in ethology and neuroscience will be essential in the future to improve our knowledge of the environmental and social factors that have influenced the emergence of language and of their respective contributions.

Data accessibility. This article has no additional data.

Authors' contributions. C.B. and P.B. wrote and revised the paper.

Competing interests. We declare we have no competing interests.

Funding. This work was funded by the French Fondation pour la Recherche Medicale (AJE201214 to P.B. and FDT201805005141 to C.B.), the Agence Nationale de la Recherche by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX), and the European Research grant (ERC COVOPRIM) to P.B.

## References

1. Darwin C. 1871 *The descent of man and selection in relation to sex*. London, UK: John Murray.
2. Fitch WT. 2017 Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.* **24**, 3–33. (doi:10.3758/s13423-017-1236-5)
3. Fitch WT, Huber L, Bugnyar T. 2010 Social cognition and the evolution of language: constructing cognitive phylogenies. *Neuron* **65**, 795–814. (doi:10.1016/j.neuron.2010.03.011)
4. Tyack PL. 2019 A taxonomy for vocal learning. *Phil. Trans. R. Soc. B* **375**, 20180406. (doi:10.1098/rstb.2018.0406)
5. Fischer J, Hammerschmidt K. 2019 Towards a new taxonomy of primate vocal production learning. *Phil. Trans. R. Soc. B* **375**, 20190045. (doi:10.1098/rstb.2019.0045)
6. Zuberbühler K. 2002 A syntactic rule in forest monkey communication. *Anim. Behav.* **63**, 293–299. (doi:10.1006/anbe.2001.1914)
7. Ouattara K, Lemasson A, Zuberbühler K. 2009 Campbell's monkeys concatenate vocalizations into context-specific call sequences. *Proc. Natl Acad. Sci. USA* **106**, 22 026–22 031. (doi:10.1073/pnas.0908118106)
8. Arnold K, Zuberbühler K. 2012 Call combinations in monkeys: compositional or idiomatic expressions? *Brain Lang.* **120**, 303–309. (doi:10.1016/j.bandl.2011.10.001)
9. Townsend SW, Engesser S, Stoll S, Zuberbühler K, Bickel B. 2018 Compositionality in animals and humans. *PLoS Biol.* **16**, e2006425. (doi:10.1371/journal.pbio.2006425)
10. Zuberbühler K. 2019 Syntax and compositionality in animal communication. *Phil. Trans. R. Soc. B* **375**, 20190062. (doi:10.1098/rstb.2019.0062)
11. Wilson B *et al.* 2015 Auditory sequence processing reveals evolutionarily conserved regions of frontal cortex in macaques and humans. *Nat. Commun.* **6**, 8901. (doi:10.1038/ncomms9901)
12. Kikuchi Y *et al.* 2017 Sequence learning modulates neural responses and oscillatory coupling in human and monkey auditory cortex. *PLoS Biol.* **15**, e2000219. (doi:10.1371/journal.pbio.2000219)
13. Fitch WT. 2018 What animals can teach us about human language: the phonological continuity hypothesis. *Curr. Opin. Behav. Sci.* **21**, 68–75. (doi:10.1016/j.cobeha.2018.01.014)
14. Seyfarth RM, Cheney DL, Marler P. 1980 Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Anim. Behav.* **28**, 1070–1094. (doi:10.1016/S0003-3472(80)80097-2)
15. Crockford C, Wittig RM, Mundry R, Zuberbühler K. 2012 Wild chimpanzees inform ignorant group members of danger. *Curr. Biol.* **22**, 142–146. (doi:10.1016/j.cub.2011.11.053)
16. Schlenker P, Chemla E, Zuberbühler K. 2016 What do monkey calls mean? *Trends Cogn. Sci.* **20**, 894–904. (doi:10.1016/j.tics.2016.10.004)
17. Graham KE, Wilke C, Lahiff NJ, Slocombe KE. 2019 Scratching beneath the surface: intentionality in great ape signal production. *Phil. Trans. R. Soc. B* **375**, 20180403. (doi:10.1098/rstb.2018.0403)
18. Hauser MD. 1988 How infant vervet monkeys learn to recognize starling alarm calls: the role of experience. *Behaviour* **105**, 187–201. (doi:10.1163/156853988X00016)
19. Zuberbühler K. 2000 Interspecies semantic communication in two forest primates. *Proc. R. Soc. Lond. B* **267**, 713–718. (doi:10.1098/rspb.2000.1061)
20. Belin P, Fecteau S, Charest I, Nicastro N, Hauser MD, Armony JL. 2008 Human cerebral response to animal affective vocalizations. *Proc. R. Soc. Lond. B* **275**, 473–481. (doi:10.1098/rspb.2007.1460)
21. Candiotti A, Zuberbühler K, Lemasson A. 2013 Voice discrimination in four primates. *Behav. Processes* **99**, 67–72. (doi:10.1016/j.beproc.2013.06.010)
22. Filippi P, Gogoleva SS, Volodina EV, Volodin IA, de Boer B. 2017 Humans identify negative (but not positive) arousal in silver fox vocalizations: implications for the adaptive value of interspecific eavesdropping. *Curr. Zool.* **63**, 445–456. (doi:10.1093/cz/zox035)
23. Belin P, Bestelmeyer PEG, Latinus M, Watson R. 2011 Understanding voice perception. *Br. J. Psychol.* **102**, 711–725. (doi:10.1111/j.2044-8295.2011.02041.x)
24. Yovel G, Belin P. 2013 A unified coding strategy for processing faces and voices. *Trends Cogn. Sci.* **17**, 263–271. (doi:10.1016/j.tics.2013.04.004)
25. Moeller S, Freiwald WA, Tsao DY. 2008 Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science* **320**, 1355–1359. (doi:10.1126/science.1157436)
26. Freiwald WA, Tsao DY, Livingstone MS. 2009 A face feature space in the macaque temporal lobe. *Nat. Neurosci.* **12**, 1187–1196. (doi:10.1038/nn.2363)
27. Freiwald WA, Tsao DY. 2010 Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851. (doi:10.1126/science.1194908)
28. Chang L, Tsao DY. 2017 The code for facial identity in the primate brain. *Cell* **169**, 1013–1028. (doi:10.1016/j.cell.2017.05.011)
29. Fitch WT. 2000 The evolution of speech: a comparative review. *Trends Cogn. Sci.* **4**, 258–267. (doi:10.1016/S1364-6613(00)01494-7)
30. Ghazanfar AA. 2008 Language evolution: neural differences that make a difference. *Nat. Neurosci.* **11**, 382–384. (doi:10.1038/nn0408-382)
31. Latinus M, Belin P. 2011 Human voice perception. *Curr. Biol.* **4**, 143–145. (doi:10.1016/j.cub.2010.12.033)
32. Mathias SR, von Kriegstein K. 2014 How do we recognise who is speaking? *Front. Biosci.* **6**, 92–109. (doi:10.2741/S417)
33. Maguinness C, Roswandowitz C, Von Kriegstein K. 2018 Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia* **116**, 179–193. (doi:10.1016/j.neuropsychologia.2018.03.039)
34. Mullennix JW, Johnson KA, Topcu-Durgun M, Farnsworth LM. 1995 The perceptual representation of voice gender. *J. Acoust. Soc. Am.* **98**, 3080–3095. (doi:10.1121/1.413832)
35. McAleer P, Todorov A, Belin P. 2014 How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE* **9**, e90779. (doi:10.1371/journal.pone.0090779)
36. DeCasper AJ, Fifer WP. 1980 Of human bonding: newborns prefer their mothers' voices. *Science* **208**, 1174–1176. (doi:10.1126/science.7375928)
37. Ockleford EM, Vince MA, Layton C, Reader MR. 1988 Responses of neonates to parents' and others' voices. *Early Hum. Dev.* **18**, 27–36. (doi:10.1016/0378-3782(88)90040-0)
38. Kisilevsky BS, Hains SMJ, Lee K, Xie X, Huang H, Ye HH, Zhang K, Wang Z. 2003 Effects of experience on

fetal voice recognition. *Psychol. Sci.* **14**, 220–224. (doi:10.1111/1467-9280.02435)

39. Flom R, Bahrick L. 2007 The development of infant discrimination of affect in multimodal and unimodal stimulation: the role of intersensory redundancy. *Dev. Psychol.* **43**, 238–252. (doi:10.1037/0012-1649.43.1.238)
40. Fant G. 1970 *Acoustic theory of speech production*. Berlin, Germany: Walter de Gruyter.
41. Baumann O, Belin P. 2010 Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol. Res.* **74**, 110–120. (doi:10.1007/s00426-008-0185-z)
42. Taylor AM, Reby D. 2010 The contribution of source–filter theory to mammal vocal communication research. *J. Zool.* **280**, 221–236. (doi:10.1111/j.1469-7998.2009.00661.x)
43. Agus TR, Suied C, Thorpe SJ, Pressnitzer D. 2012 Fast recognition of musical sounds based on timbre. *J. Acoust. Soc. Am.* **131**, 4124–4133. (doi:10.1121/1.3701865)
44. Isnard V. 2016 L'efficacité du système auditif humain pour la reconnaissance de sons naturels. Thesis, Paris [cité 10 nov 2018]. See http://www.theses.fr/2016PA066458
45. Suied C, Agus TR, Thorpe SJ, Mesgarani N, Pressnitzer D. 2014 Auditory gist: recognition of very short sounds from timbre cues. *J. Acoust. Soc. Am.* **135**, 1380–1391. (doi:10.1121/1.4863659)
46. Green S. 1975 Variation of vocal pattern with social situation in the Japanese monkey (*Macaca fuscata*): a field study. *Primate Behav.* **4**, 1–102.
47. Hauser MD. 1991 Sources of acoustic variation in rhesus macaque (*Macaca mulatta*) vocalizations. *Ethology* **89**, 29–46. (doi:10.1111/j.1439-0310.1991.tb00291.x)
48. Hauser MD, Marler P. 1993 Food-associated calls in rhesus macaques (*Macaca mulatta*): I. Socioecological factors. *Behav. Ecol.* **4**, 194–205. (doi:10.1093/beheco/4.3.194)
49. Rowell TE, Hinde RA. 1962 Vocal communication by the rhesus Mojsxey (*Macaca mulatta*). *Proc. Zool. Soc. Lond.* **138**, 279–294. (doi:10.1111/j.1469-7998.1962.tb05698.x)
50. Katsu N, Yamada K, Nakamichi M. 2016 Function of grunts, girneys and coo calls of Japanese macaques (*Macaca fuscata*) in relation to call usage, age and dominance relationships. *Behaviour* **153**, 125–142. (doi:10.1163/1568539X-00003330)
51. Beecher MD, Petersen MR, Zoloth SR, Moody DB, Stebbins WC. 1979 Perception of conspecific vocalizations by Japanese macaques. *Brain Behav. Evol.* **16**, 443–460. (doi:10.1159/000121881)
52. Zoloth SR, Petersen MR, Beecher MD, Green S, Marler P, Moody DB, Stebbins W. 1979 Species-specific perceptual processing of vocal sounds by monkeys. *Science* **204**, 870–873. (doi:10.1126/science.108805)
53. Petersen MR, Beecher MD, Zoloth SR, Green S, Marler PR, Moody DB, Stebbins WC. 1984 Neural lateralization of vocalizations by Japanese macaques: communicative significance is more important than acoustic structure. *Behav. Neurosci.* **98**, 779–790. (doi:10.1037/0735-7044.98.5.779)
54. May B, Moody DB, Stebbins WC. 1988 The significant features of Japanese macaque coo sounds: a psychophysical study. *Anim. Behav.* **36**, 1432–1444. (doi:10.1016/S0003-3472(88)80214-8)
55. May B, Moody DB, Stebbins WC. 1989 Categorical perception of conspecific communication sounds by Japanese macaques, *Macaca fuscata*. *J. Acoust. Soc. Am.* **85**, 837–847. (doi:10.1121/1.397555)
56. Le Prell CG, Moody DB. 1997 Perceptual salience of acoustic features of Japanese monkey coo calls. *J. Comp. Psychol.* **111**, 261–274. (doi:10.1037/0735-7036.111.3.261)
57. Le Prell CG, Hauser MD, Moody DB. 2002 Discrete or graded variation within rhesus monkey screams? Psychophysical experiments on classification. *Anim. Behav.* **63**, 47–62. (doi:10.1006/anbe.2001.1888)
58. Christison-Lagay KL, Bennur S, Blackwell J, Lee JH, Schroeder T, Cohen YE. 2014 Natural variability in species-specific vocalizations constrains behavior and neural activity. *Hear. Res.* **312**, 128–142. (doi:10.1016/j.heares.2014.03.007)
59. Gouzoules S, Gouzoules H, Marler P. 1984 Rhesus monkey (*Macaca mulatta*) screams: representational signalling in the recruitment of agonistic aid. *Anim. Behav.* **32**, 182–193. (doi:10.1016/S0003-3472(84)80336-X)
60. Hauser MD. 1996 *The evolution of communication*. Cambridge, MA: MIT Press.
61. Rendall D, Rodman PS, Emond RE. 1996 Vocal recognition of individuals and kin in free-ranging rhesus monkeys. *Anim. Behav.* **51**, 1007–1015. (doi:10.1006/anbe.1996.0103)
62. Fitch WT, Fritz JB. 2006 Rhesus macaques spontaneously perceive formants in conspecific vocalizations. *J. Acoust. Soc. Am.* **120**, 2132–2141. (doi:10.1121/1.2258499)
63. Fitch WT. 1997 Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J. Acoust. Soc. Am.* **102**, 1213–1222. (doi:10.1121/1.421048)
64. Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. 2007 Vocal-tract resonances as indexical cues in rhesus monkeys. *Curr. Biol.* **17**, 425–430. (doi:10.1016/j.cub.2007.01.029)
65. Eliades SJ, Miller CT. 2017 Marmoset vocal communication: behavior and neurobiology. *Dev. Neurobiol.* **77**, 286–299. (doi:10.1002/dneu.22464)
66. Miller CT, Freiwald WA, Leopold DA, Mitchell JF, Silva AC, Wang X. 2016 Marmosets: a neuroscientific model of human social behavior. *Neuron* **90**, 219–233. (doi:10.1016/j.neuron.2016.03.018)
67. Agamaite JA, Chang C-J, Osmanski MS, Wang X. 2015 A quantitative acoustic analysis of the vocal repertoire of the common marmoset (*Callithrix jacchus*). *J. Acoust. Soc. Am.* **138**, 2906–2928. (doi:10.1121/1.4934268)
68. Miller CT, Mandel K, Wang X. 2010 The communicative content of the common marmoset phee call during antiphonal calling. *Am. J. Primatol.* **72**, 974–980. (doi:10.1002/ajp.20854)
69. Pistorio AL, Vintch B, Wang X. 2006 Acoustic analysis of vocal development in a New World primate, the common marmoset (*Callithrix jacchus*). *J. Acoust. Soc. Am.* **120**, 1655–1670. (doi:10.1121/1.2225899)
70. Kato Y, Gokan H, Oh-Nishi A, Suhara T, Watanabe S, Minamimoto T. 2014 Vocalizations associated with anxiety and fear in the common marmoset (*Callithrix jacchus*). *Behav. Brain Res.* **275**, 43–52. (doi:10.1016/j.bbr.2014.08.047)
71. Wakita M. 2019 Auditory sequence perception in common marmosets (*Callithrix jacchus*). *Behav. Processes.* **162**, 55–63. (doi:10.1016/j.beproc.2019.01.014)
72. Miller CT, Wang X. 2006 Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. *J. Comp. Physiol. A* **192**, 27–38. (doi:10.1007/s00359-005-0043-z)
73. Norcross JL, Newman JD, Cofrancesco LM. 1999 Context and sex differences exist in the acoustic structure of phee calls by newly-paired common marmosets (*Callithrix jacchus*). *Am. J. Primatol.* **49**, 165–181. (doi:10.1002/(SICI)1098-2345(199910)49:2<165::AID-AJP7>3.0.CO;2-S)
74. Miller CT, Wren TA. 2012 Individual recognition during bouts of antiphonal calling in common marmosets. *J. Comp. Physiol. A* **198**, 337–346. (doi:10.1007/s00359-012-0712-7)
75. Kato M, Yokoyama C, Kawasaki A, Takeda C, Koike T, Onoe H, Iriki A. 2018 Individual identity and affective valence in marmoset calls: in vivo brain imaging with vocal sound playback. *Anim. Cogn.* **21**, 331–343. (doi:10.1007/s10071-018-1169-z)
76. Parr LA, Cohen M, de Waal F. 2005 Influence of social context on the use of blended and graded facial displays in chimpanzees. *Int. J. Primatol.* **26**, 73–103. (doi:10.1007/s10764-005-0724-z)
77. Crockford C, Boesch C. 2005 Call combinations in wild chimpanzees. *Behaviour* **142**, 397–421. (doi:10.1163/1568539054012047)
78. Marler P, Hobbett L. 1975 Individuality in a long-range vocalization of wild chimpanzees. *Z. Tierpsychol.* **38**, 97–109. (doi:10.1111/j.1439-0310.1975.tb01994.x)
79. Bauer HR, Philip MM. 1983 Facial and vocal individual recognition in the common chimpanzee. *Psychol. Rec.* **33**, 161–170. (doi:10.1007/BF03394834)
80. Kojima S, Izumi A, Ceugniet M. 2003 Identification of vocalizers by pant hoots, pant grunts and screams in a chimpanzee. *Primates* **44**, 225–230. (doi:10.1007/s10329-002-0014-8)
81. Hackett TA. 2015 Anatomic organization of the auditory cortex. In *Handbook of clinical neurology*, vol. 129 (eds MJ Aminoff, F Boller, DF Swaab), pp. 27–53. Amsterdam, The Netherlands: Elsevier.
82. Trapeau R, Bodin C, Belin P. 2019 A comparative fMRI study of voice-selective regions in primates. In *Organization for Human Brain Mapping Conf., 9–13 June, Rome*.
83. Seidlitz J, Sponheim C, Glen D, Ye FQ, Saleem KS, Leopold DA, Ungerleider L, Messinger A. 2018

A population MRI brain template and analysis tools for the macaque. *Neuroimage* **170**, 121–131. (doi:10.1016/j.neuroimage.2017.04.063)

84. Woodward A, Hashikawa T, Maeda M, Kaneko T, Hikishima K, Iriki A, Okano H, Yamaguchi Y. 2018 The Brain/MINDS 3D digital marmoset brain atlas. *Sci. Data* **5**, 180009. (doi:10.1038/sdata.2018.9)
85. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000 Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312. (doi:10.1038/35002078)
86. Belin P, Zatorre RJ, Ahad P. 2002 Human temporal-lobe response to vocal sounds. *Cogn. Brain Res.* **13**, 17–26. (doi:10.1016/S0926-6410(01)00084-2)
87. Kriegstein K, Giraud A. 2004 Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* **22**, 948–955. (doi:10.1016/j.neuroimage.2004.02.020)
88. Fecteau S, Armony JL, Joanette Y, Belin P. 2004 Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* **23**, 840–848. (doi:10.1016/j.neuroimage.2004.09.019)
89. Pernet CR *et al.* 2015 The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* **119**, 164–174. (doi:10.1016/j.neuroimage.2015.06.050)
90. Agus TR, Paquette S, Suied C, Pressnitzer D, Belin P. 2017 Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Sci. Rep.* **7**, 11526. (doi:10.1038/s41598-017-11684-1)
91. Bodin C, Takerkart S, Belin P, Coulon O. 2017 Anatomo-functional correspondence in the superior temporal sulcus. *Brain Struct. Funct.* **223**, 221–232. (doi:10.1007/s00429-017-1483-2)
92. Aglieri V, Chaminade T, Takerkart S, Belin P. 2018 Functional connectivity within the voice perception network and its behavioural relevance. *Neuroimage* **183**, 356–365. (doi:10.1016/j.neuroimage.2018.08.011)
93. Blank H, Anwander A, von Kriegstein K. 2011 Direct structural connections between voice- and face-recognition areas. *J. Neurosci.* **31**, 12 906–12 915. (doi:10.1523/JNEUROSCI.2091-11.2011)
94. Glass I, Wollberg Z. 1983 Responses of cells in the auditory cortex of awake squirrel monkeys to normal and reversed species-specific vocalizations. *Hear. Res.* **9**, 27–33. (doi:10.1016/0378-5955(83)90131-4)
95. Wollberg Z, Newman JD. 1972 Auditory cortex of squirrel monkey: response patterns of single cells to species-specific vocalizations. *Science* **175**, 212–214. (doi:10.1126/science.175.4018.212)
96. Ghazanfar AA, Chandrasekaran C, Logothetis NK. 2008 Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J. Neurosci.* **28**, 4457–4469. (doi:10.1523/JNEUROSCI.0541-08.2008)
97. Perrodin C, Kayser C, Logothetis NK, Petkov CI. 2011 Voice cells in the primate temporal lobe. *Curr. Biol.* **16**, 1408–1415. (doi:10.1016/j.cub.2011.07.028)
98. Romanski LM, Averbeck BB. 2009 The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu. Rev. Neurosci.* **32**, 315–346. (doi:10.1146/annurev.neuro.051508.135431)
99. Tian B, Reser D, Durham A, Kustov A, Rauschecker JP. 2001 Functional specialization in rhesus monkey auditory cortex. *Science* **292**, 290–293. (doi:10.1126/science.1058911)
100. Fukushima M, Saunders RC, Leopold DA, Mishkin M, Averbeck BB. 2014 Differential coding of conspecific vocalizations in the ventral auditory cortical stream. *J. Neurosci.* **34**, 4665–4676. (doi:10.1523/JNEUROSCI.3969-13.2014)
101. Kikuchi Y, Horwitz B, Mishkin M. 2010 Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J. Neurosci.* **30**, 13 021–13 030. (doi:10.1523/JNEUROSCI.2267-10.2010)
102. Nagarajan SS, Cheung SW, Bedenbaugh P, Beitel RE, Schreiner CE, Merzenich MM. 2002 Representation of spectral and temporal envelope of twitter vocalizations in common marmoset primary auditory cortex. *J. Neurophysiol.* **87**, 1723–1737. (doi:10.1152/jn.00632.2001)
103. Wang X, Merzenich MM, Beitel R, Schreiner CE. 1995 Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J. Neurophysiol.* **74**, 2685–2706. (doi:10.1152/jn.1995.74.6.2685)
104. Wang X, Kadia SC. 2001 Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J. Neurophysiol.* **86**, 2616–2620. (doi:10.1152/jn.2001.86.5.2616)
105. Chen G, Wang F, Dillenburger BC, Friedman RM, Chen LM, Gore JC, Avison MJ, Roe AW. 2012 Functional magnetic resonance imaging of awake monkeys: some approaches for improving imaging quality. *Magn. Reson. Imaging.* **30**, 36–47. (doi:10.1016/j.mri.2011.09.010)
106. Vanduffel W, Farivar R. 2014 Functional MRI of awake behaving macaques using standard equipment. In *Advanced brain neuroimaging topics in health and disease - methods and applications* (eds TD Papageorgiou, GI Christopoulos, SM Smirnakis). InTech. [cité 4 août 2016]. See http://www.intechopen.com/books/advanced-brain-neuroimaging-topics-in-health-and-disease-methods-and-applications/functional-mri-of-awake-behaving-macaques-using-standard-equipment.
107. Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. 2008 A voice region in the monkey brain. *Nat. Neurosci.* **11**, 367–374. (doi:10.1038/nn2043)
108. Ortiz-Rios M, Kuśmierek P, DeWitt I, Archakov D, Azevedo FAC, Sams M, Jääskeläinen IP, Keliris GA, Rauschecker JP. 2015 Functional MRI of the vocalization-processing network in the macaque brain. *Front. Neurosci.* **9**, 113. (doi:10.3389/fnins.2015.00113)
109. Joly O, Pallier C, Ramus F, Pressnitzer D, Vanduffel W, Orban GA. 2012 Processing of vocalizations in humans and monkeys: a comparative fMRI study. *Neuroimage* **62**, 1376–1389. (doi:10.1016/j.neuroimage.2012.05.070)
110. Ghazanfar AA, Eliades SJ. 2014 The neurobiology of primate vocal communication. *Curr. Opin. Neurobiol.* **28**, 128–134. (doi:10.1016/j.conb.2014.06.015)
111. Benevento LA, Fallon J, Davis BJ, Rezak M. 1977 Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Exp. Neurol.* **57**, 849–872. (doi:10.1016/0014-4886(77)90112-1)
112. Hikosaka K, Iwai E, Saito H, Tanaka K. 1988 Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *J. Neurophysiol.* **60**, 1615–1637. (doi:10.1152/jn.1988.60.5.1615)
113. Perrodin C, Kayser C, Logothetis NK, Petkov CI. 2014 Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *J. Neurosci.* **7**, 2524–2537. (doi:10.1523/JNEUROSCI.2805-13.2014)
114. Belcher AM, Yen CC, Stepp H, Gu H, Lu H, Yang Y, Silva AC, Stein EA. 2013 Large-scale brain networks in the awake, truly resting marmoset monkey. *J. Neurosci.* **33**, 16 796–16 804. (doi:10.1523/JNEUROSCI.3146-13.2013)
115. Hung C-C, Yen CC, Ciuchta JL, Papoti D, Bock NA, Leopold DA, Silva AC. 2015 Functional MRI of visual responses in the awake, behaving marmoset. *Neuroimage* **120**, 1–11. (doi:10.1016/j.neuroimage.2015.06.090)
116. Papoti D, Yen CC-C, Hung C-C, Ciuchta J, Leopold DA, Silva AC. 2017 Design and implementation of embedded 8-channel receive-only arrays for whole-brain MRI and fMRI of conscious awake marmosets. *Magn. Reson. Med.* **78**, 387–398. (doi:10.1002/mrm.26339)
117. Silva AC. 2017 Anatomical and functional neuroimaging in awake, behaving marmosets. *Dev. Neurobiol.* **77**, 373–389. (doi:10.1002/dneu.22456)
118. Sadagopan S, Temiz-Karayol NZ, Voss HU. 2015 High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Sci. Rep.* **5**, 10950. (doi:10.1038/srep10950)
119. Taglialatela JP, Russell JL, Schaeffer JA, Hopkins WD. 2009 Visualizing vocal perception in the chimpanzee brain. *Cereb. Cortex* **19**, 1151–1157. (doi:10.1093/cercor/bhn157)
120. Le Prell CG, Moody DB. 2000 Factors influencing the salience of temporal cues in the discrimination of synthetic Japanese monkey (*Macaca fuscata*) coo calls. *J. Exp. Psychol. Anim. Behav. Process.* **26**, 261–273. (doi:10.1037/0097-7403.26.3.261)
121. Ghazanfar AA, Smith-Rohrberg D, Hauser MD. 2001 The role of temporal cues in rhesus monkey vocal recognition: orienting asymmetries to reversed calls. *Brain Behav. Evol.* **58**, 163–172.
122. Andics A, McQueen J, Petersson K, Gal V, Rudas G, Vidnyanszky Z. 2010 Neural mechanisms for voice recognition. *Neuroimage* **52**, 1528–1540. (doi:10.1016/j.neuroimage.2010.05.048)

123. Belin P, Zatorre R. 2003 Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* **14**, 2105–2109. (doi:10.1097/00001756-200311140-00019)

124. Charest I, Pernet C, Latinus M, Crabbe F, Belin P. 2013 Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cereb. Cortex* **23**, 958–966. (doi:10.1093/cercor/bhs090)

125. Latinus M, Mc Aleer P, Bestelmeyer PEG, Belin P. 2013 Norm-based coding of voice identity in human auditory cortex. *Curr. Biol.* **23**, 1075–1080. (doi:10.1016/j.cub.2013.04.055)

126. Roswandowitz C, Kappes C, Obrig H, von Kriegstein K. 2018 Obligatory and facultative brain regions for voice-identity recognition. *Brain* **141**, 234–247. (doi:10.1093/brain/awx313)

127. Jiahui G, Garrido L, Liu RR, Susilo T, Barton JJS, Duchaine B. 2017 Normal voice processing after posterior superior temporal sulcus lesion. *Neuropsychologia* **105**, 215–222. (doi:10.1016/j.neuropsychologia.2017.03.008)

128. Perrodin C, Kayser C, Abel TJ, Logothetis NK, Petkov CI. 2015 Who is that? Brain networks and mechanisms for identifying individuals. *Trends Cogn. Sci.* **19**, 783–796. (doi:10.1016/j.tics.2015.09.002)

129. Petkov CI, Kikuchi Y, Milne AE, Mishkin M, Rauschecker JP, Logothetis NK. 2015 Different forms of effective connectivity in primate frontotemporal pathways. *Nat. Commun.* **6**, 6000. (doi:10.1038/ncomms7000)

130. Russ BE, Ackelson AL, Baker AE, Cohen YE. 2008 Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *J. Neurophysiol.* **99**, 87–95. (doi:10.1152/jn.01069.2007)

131. Rauschecker JP. 2012 Ventral and dorsal streams in the evolution of speech and language. *Front. Evol. Neurosci.* **4**, 7. (doi:10.3389/fnevo.2012.00007)

132. Belin P, Bodin C, Aglieri V. 2018 A 'voice patch' system in the primate brain for processing vocal information? *Hear. Res.* **366**, 65–74. (doi:10.1016/j.heares.2018.04.010)

133. Hung C-C, Yen CC, Ciuchta JL, Papoti D, Bock NA, Leopold DA, Silva AC. 2015 Functional mapping of face-selective regions in the extrastriate visual cortex of the marmoset. *J. Neurosci.* **35**, 1160–1172. (doi:10.1523/JNEUROSCI.2659-14.2015)

134. Yovel G, Freiwald WA. 2013 Face recognition systems in monkey and human: are they the same thing? *F1000Prime Rep.* **5**, 10. See http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3619156/ (doi:10.12703/P5-10)

135. Formisano E, De Martino F, Bonte M, Goebel R. 2008 'Who' is saying 'what'? brain-based decoding of human voice and speech. *Science* **322**, 970–973. (doi:10.1126/science.1164318)

136. Parr LA, Taubert J, Little AC, Hancock PJB. 2012 The organization of conspecific face space in nonhuman primates. *Q. J. Exp. Psychol.* **65**, 2411–2434. (doi:10.1080/17470218.2012.693110)

137. Leopold DA, Bondar IV, Giese MA. 2006 Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575. (doi:10.1038/nature04951)

138. Parr LA. 2011 The evolution of face processing in primates. *Phil. Trans. R. Soc. B* **366**, 1764–1777. (doi:10.1098/rstb.2010.0358)

139. Parr LA, Hecht E, Barks SK, Preuss TM, Votaw JR. 2009 Face processing in the chimpanzee brain. *Curr. Biol.* **19**, 50–53. (doi:10.1016/j.cub.2008.11.048)

140. von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L. 2005 Interaction of face and voice areas during speaker recognition. *J. Cogn. Neurosci.* **17**, 367–376. (doi:10.1162/0898929053279577)

141. Watson R, Latinus M, Charest I, Crabbe F, Belin P. 2014 People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex* **50**, 125–136. (doi:10.1016/j.cortex.2013.07.011)

142. Hasan BAS, Valdes-Sosa M, Gross J, Belin P. 2016 'Hearing faces and seeing voices': a modal coding of person identity in the human brain. *Sci. Rep.* **6**, 37494. (doi:10.1038/srep37494)

143. Tsantani M, Kriegeskorte N, McGettigan C, Garrido L. 2019 Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. *Neuroimage* **201**, 116004. (doi:10.1016/j.neuroimage.2019.07.017)

144. Davies-Thompson J, Elli GV, Rezk M, Benetti S, van Ackeren M, Collignon O. 2018 Hierarchical brain network for face and voice integration of emotion expression. *Cereb. Cortex* **1**, 16. (doi:10.1093/cercor/bhy240)

145. Lahnakoski JM, Glerean E, Salmi J, Jaaskelainen LP, Sams M, Hari R, Nummenmaa L. 2012 Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Front. Hum. Neurosci.* **6**, 233. (doi:10.3389/fnhum.2012.00233)

146. Isik L, Koldewyn K, Beeler D, Kanwisher N. 2017 Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl Acad. Sci. USA* **114**, E9145–E9152. (doi:10.1073/pnas.1714471114)

147. Chaplin TA, Yu H-H, Soares JGM, Gattass R, Rosa MGP. 2013 A conserved pattern of differential expansion of cortical areas in simian primates. *J. Neurosci.* **33**, 15 120–15 125. (doi:10.1523/JNEUROSCI.2909-13.2013)

148. Patel GH, Sestieri C, Corbetta M. 2019 The evolution of the temporoparietal junction and posterior superior temporal sulcus. *Cortex* **118**, 38–50. (doi:10.1016/j.cortex.2019.01.026)

149. Liu ST, Montes-Lourido P, Wang X, Sadagopan S. 2019 Optimal features for auditory categorization. *Nat. Commun.* **10**, 1302. (doi:10.1038/s41467-019-09115-y)

150. Shepherd SV, Freiwald WA. 2018 Functional networks for social communication in the macaque monkey. *Neuron* **99**, 413–420. (doi:10.1016/j.neuron.2018.06.027)

151. Arcaro MJ, Schade PF, Vincent JL, Ponce CR, Livingstone MS. 2017 Seeing faces is necessary for face-domain formation. *Nat. Neurosci.* **20**, 1404–1412. (doi:10.1038/nn.4635)

152. Gultekin YB, Hage SR. 2018 Limiting parental interaction during vocal development affects acoustic call structure in marmoset monkeys. *Sci. Adv.* **4**, eaar4012. (doi:10.1126/sciadv.aar4012)

153. Grossmann T, Oberecker R, Koch SP, Friederici AD. 2010 The developmental origins of voice processing in the human brain. *Neuron* **65**, 852–858. (doi:10.1016/j.neuron.2010.03.001)

154. Vouloumanos A, Hauser MD, Werker JF, Martin A. 2010 The tuning of human neonates' preference for speech. *Child Dev.* **81**, 517–527. (doi:10.1111/j.1467-8624.2009.01412.x)

155. Nygaard LC, Sommers MS, Pisoni DB. 1994 Speech perception as a talker-contingent process. *Psychol. Sci.* **5**, 42–46. (doi:10.1111/j.1467-9280.1994.tb00612.x)

156. Goggin JP, Thompson CP, Strube G, Simental LR. 1991 The role of language familiarity in voice identification. *Mem. Cognit.* **19**, 448–458. (doi:10.3758/BF03199567)

157. von Kriegstein K, Eger E, Kleinschmidt A, Giraud A. 2003 Modulation of neural responses to speech by directing attention to voices or verbal content. *Cogn. Brain Res.* **17**, 48–55. (doi:10.1016/S0926-6410(03)00079-X)

158. Myers EB, Theodore RM. 2017 Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain Lang.* **165**, 33–44. (doi:10.1016/j.bandl.2016.11.001)

159. Hickok G, Poeppel D. 2007 The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402. (doi:10.1038/nrn2113)

160. Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, Behrens TE. 2008 The evolution of the arcuate fasciculus revealed with comparative DTI. *Nat. Neurosci.* **11**, 426–428. (doi:10.1038/nn2072)