Neural correlates of audio-visual integration of socially meaningful information in macaque monkeys

Mathilda Froesel¹, Maëva Gacoin¹, Simon Clavagnier¹, Marc Hauser², Quentin Goudard¹, Suliann Ben Hamed¹

1. Institut des Sciences Cognitives Marc Jeannerod, UMR5229 CNRS Université de Lyon, 67 Boulevard Pinel, 69675 Bron Cedex, France
2. Risk-Eraser, LLC, PO Box 376, West Falmouth, MA, 02574

Corresponding authors : mathilda.froesel@isc.cnrs.fr, benhamed@isc.cnrs.fr

Abstract

Social interactions rely on the ability to interpret semantic and emotional information, often from multiple sensory modalities. In human and nonhuman primates, both the auditory and visual modalities are used to generate and interpret communicative signals. In individuals with autism, not only are there deficits in social communication, but in the integration of audio-visual information. At present, we know little about the neural mechanisms that subserve the interpretation of complex social events, including the audio-visual integration that is often required with accompanying communicative signals. Based on heart rate estimates and fMRI in two macaque monkeys (*Macaca mulatta*), we show that individuals systematically associate affiliative facial expressions or social scenes with corresponding affiliative vocalizations, aggressive facial expressions or social scenes with corresponding aggressive vocalizations and escape visual scenes with scream vocalizations. In contrast, vocalizations that are incompatible with the visual information are fully suppressed, suggesting top-down regulation over the processing of sensory input. The process of binding audio-visual semantic and contextual information relies on a core functional network involving the superior temporal sulcus (STS) and lateral sulcus (LS). Peak activations in both sulci co-localize with face or voice patches that have been previously described. While all of these regions of interest (ROIs) respond to both auditory and visual information, LS ROIs have a preference for auditory and audio-visual congruent stimuli while STS ROIs equally respond to auditory, visual and audio-visual

33 congruent stimuli. To further specify the cortical network involved in the control of this semantic
 34 association, we performed a whole brain gPPI functional connectivity analysis on the LS and STS
 35 cumulated ROIs. This gPPI analysis highlights a functional network connected to the LS and STS,
 36 involving the anterior cingulate cortex (ACC), area 46 in the dorsolateral prefrontal cortex (DLPFC),
 37 the orbitofrontal cortex (OFC), the intraparietal sulcus (IPS), the insular cortex and subcortically, the
 38 amygdala and the hippocampus. Comparing human and macaque results, we propose that the
 39 integration of audio-visual information for congruent, meaningful social events involves homologous
 40 neural circuitry, specifically, an emotional network composed of the STS, LS, ACC, OFC, and limbic
 41 areas, including the amygdala, and an attentional network including the STS, LS, IPS and DLPFC. As
 42 such, these networks are critical to the amodal representation of social meaning, thereby providing
 43 an explanation for some of deficits observed in autism.

44

45

46

47

48

49 Introduction

50 Brain structure and function have evolved in response to social relationships, both within and
 51 between groups, in all mammals. For example, across species, brain size and gyrification has been
 52 shown to increase with average social group size (Fox et al., 2017; Shultz & Dunbar, 2010; Van Essen
 53 & Dierker, 2007), as well as meta-cognitive abilities (Devaine et al., 2017). Within a given species,
 54 functional connectivity within the so-called social brain has been shown to be stronger in macaques
 55 living in larger social groups (Mars et al., 2012). In this context, successful social interactions require
 56 the proper interpretation of social signals (Ghazanfar & Hauser, 1999), whether visual (body
 57 postures, facial expressions, inter-individual interactions) or auditory (vocalization).

58 In humans, the core language system is amodal, in the sense that our phonology, semantics and
 59 syntax function in the same way whether the input is auditory (speech) or visual (sign). In monkeys
 60 and apes, vocalizations are often associated with specific facial expressions and body postures (Parr
 61 et al., 2005). This raises the question of whether and how auditory and visual information are
 62 integrated to interpret the meaning of a given situation, including emotional state and functional
 63 behavioral responses. For example, macaque monkeys *scream* as an indication of fear, triggered by
 64 potential danger from conspecifics or heterospecifics. In contrast, macaques *coo* during positive
 65 social interactions, involving approach, feeding and group movement. To what extent does hearing a
 66 scream generate a visual representation of the individual(s) involved in such an antagonistic
 67 situation, as opposed to a positive social situation, and does seeing an antagonistic situation set up
 68 an expectation that screams, but not coos, will be produced?

69 Face, voice, and social scene processing in monkeys have been individually explored, to some extent,
 70 from the behavioural (Gothard et al., 2004, 2009; Rendall et al., 1996; Sliwa et al., 2011) and the
 71 neuronal point of view (Aparicio et al., 2016; Arcaro et al., 2017; Cohen et al., 2007; Eifuku, 2014; Gil-
 72 da-Costa et al., 2004, 2006; Hesse & Tsao, 2020; Issa & DiCarlo, 2012; Joly, Pallier, et al., 2012; Joly,
 73 Ramus, et al., 2012; Moeller et al., 2008; Ortiz-Rios et al., 2015; Petkov et al., 2008; Pinsk et al., 2005,
 74 2009; Poremba et al., 2003, 2004; Romanski et al., 2005; Russ et al., 2008; Schwiedrzik et al., 2015;
 75 Sliwa & Freiwald, 2017; Tsao et al., 2003). Audiovisual integration during naturalistic social stimuli
 76 has recently been shown in specific regions of the monkey face-patch system (Khandhadia et al.,
 77 2021), the voice-patch system (Ghazanfar, 2009; Ghazanfar et al., 2005; Perrodin et al., 2014, 2015),
 78 as well as in the prefrontal voice area (Romanski, 2012). However, beyond combining sensory
 79 information, social perception also involves integrating contextual, behavioural and emotional
 80 information (Freiwald, 2020; Ghazanfar & Santos, 2004). In this context, how macaque monkeys
 81 associate specific vocalizations with specific social visual scenes based on their respective meaning
 82 has scarcely been explored. Our goal is to help fill this gap.

83 This study used video-based heart rate monitoring and functional magnetic resonance in awake
 84 behaving monkeys to show that rhesus monkeys (*Macaca mulatta*) systematically associate the
 85 meaning of a vocalization with the meaning of a visual scene. Specifically, they associate affiliative
 86 facial expressions or social scenes with corresponding affiliative vocalizations, aggressive facial
 87 expressions or social scenes with corresponding aggressive vocalizations, and escape visual scenes
 88 with scream vocalizations. In contrast, vocalizations that are incompatible with the visual information
 89 are fully suppressed, indicating a top-down regulation over the processing of sensory input. Providing
 90 evidence of a homology with humans (Haxby et al., 2002; Haxby & Gobbini, 2011), we further show,
 91 using a functional connectivity analysis, that this audio-visual association involves two functionally
 92 coupled networks, one involved in the emotional processing of social stimuli, and one involved in
 93 their cognitive and attentional assessment.

94 Results

95 In the following, we investigate whether and how macaques associate visual and auditory stimuli
 96 based on their semantic content, and we characterize the neuronal bases underlying this audio-visual
 97 integration. We obtained neural and autonomic data from two macaques using functional magnetic
 98 resonance brain imaging and video-based heart rate tracking. Each task combined visual stimuli of
 99 identical social content with either semantically congruent or incongruent monkey vocalizations. On
 100 each block of trials, the monkeys could be exposed to either visual stimuli only, auditory congruent
 101 stimuli only, auditory incongruent stimuli only, audio-visual congruent stimuli or audio-visual
 102 incongruent stimuli, in a blocked design (Figure 1a). Importantly, paired blocked conditions shared
 103 the same auditory stimuli, but opposite social visual content (Figure 1b). We report group fMRI and
 104 group heart-rate analyses. All reported statistics are based on non-parametric tests.

105

Figure 1: A) Experimental design. Example of an aggressive face (F-) blocked condition. Each run was composed of three randomized repetitions of six different blocks of 16 seconds. The six blocks could be visual (Vi), auditory with sounds congruent with the visual stimuli (AC), auditory with sounds incongruent with the visual stimuli (AI), audio-visual with sounds congruent with the visual stimuli (VAC), audio-visual with sounds incongruent with the visual stimuli (VAI), or fixation with no sensory stimulation (Fx). Each sensory stimulation block contained a rapid succession of 500ms stimuli. Each run started and ended with 10 seconds of fixation. **B) Description of blocked conditions.** Six different blocked conditions were used. Each blocked condition combined visual stimuli of identical social content with either semantically congruent or incongruent monkey vocalizations. Pairs blocked conditions shared the same auditory stimuli, but opposite social visual content (F+ vs. F-; S1+ vs. S1-; S2+ vs. S2-).

Auditory whole brain activations depend on semantic congruence with visual context.

Figure 2: Whole-brain activation FACE blocked condition (F+ & F-): main contrasts. Whole-brain activation maps of the F+ (face affiliative) and F- (face aggressive) runs, cumulated over both monkeys, for the visual (white, Vi vs. Fx), auditory congruent (dark green, AC vs. Fx) and auditory

incongruent (dark red, AI vs. Fx). Note that the AC and AI conditions contain exactly the same sound samples (coos and aggressive calls). Darker shades of red indicate level of significance at $p < 0.001$ uncorrected, t -score 3.09. Lighter shades of yellow and brown outlines indicate level of significance at $p < 0.05$ FWE, t -score 4.6. ios: Inferior Occipital Sulcus; LS: Lateral Sulcus; STS: Superior Temporal Sulcus.

Combining the F+ and F- face blocked conditions (Figure 2), which includes faces expressing lipsmacks or aggressive threats, we find robust bilateral activation ($p < 0.05$ FWE) in the extra-striate cortex, along the superior temporal sulcus (STS) as well as in the prefrontal cortex, as expected from previous studies (Eifuku, 2014; Moeller et al., 2008; Tsao, Schweers, et al., 2008). Activations were also observed in the posterior part of the fundus of the intraparietal sulcus at an uncorrected level ($p < 0.0001$). Please note that receiving coils were placed so as to optimize temporal and prefrontal cortex signal-to-noise ratio. As a result, no activations can be seen in the occipital cortex. The congruent auditory versus fixation contrast, which combined aggressive calls and coos in the two different blocked conditions, leads to activation within the inferior bank of the lateral sulcus, both at corrected ($p < 0.05$ FWE) and uncorrected levels ($p < 0.0001$), as described in previous studies (Joly, Pallier, et al., 2012; Petkov et al., 2008; Poremba et al., 2003). Importantly, this blocked condition also leads to the same robust bilateral activations as the visual contrast: the extra-striate cortex, along the superior temporal sulcus (STS) ($p < 0.05$ FWE), as well as in the prefrontal and intraparietal cortex ($p < 0.0001$ uncorrected). These activations are similar whether the congruent auditory stimuli are coos (Figure 3b) or aggressive calls (Figure 3c). In contrast, when we present the exact same aggressive calls and coos, the incongruent auditory versus fixation contrast leads to minimal activation, if any (Figure 2). Again, this doesn't depend on whether the incongruent sounds are aggressive calls (Figure 3d) or coos (Figure 3e). This pattern of activation therefore confirms that auditory activation does not depend on the nature of the vocalization. Rather, it depends on whether the vocalizations are congruent or not to the semantic content of the visual stimuli.

Figure 3: Auditory activations depend on semantic congruence with visual context. A) Whole-brain activation maps of the F+ (face affiliative) and F- (face aggressive) runs, for the *auditory congruent vs auditory incongruent* (relative to the visual context) contrast. Whole-brain activation map for the F+ (face affiliative) B) auditory congruent (coos, dark green, AC vs. Fx) and D) auditory incongruent (aggressive calls, dark red, AI vs. Fx) conditions. Whole-brain activation map for the F- (face aggressive) C) auditory congruent (aggressive calls, dark green, AC vs. Fx) and E) auditory incongruent (coos, dark red, AI vs. Fx) conditions. Darker shades of red indicate level of significance at $p < 0.001$ uncorrected, t-score 3.09. Lighter shades of yellow and brown outlines indicate level of significance at $p < 0.05$ FWE, t-score 4.6.

These observations are reproduced in a different set of blocked conditions, in which the visual stimuli involve social scenes (grooming, aggression or escape) with either semantically congruent or incongruent vocalizations (Figure 4 for all social blocked conditions and Figure S1 for S+ and S- social blocked conditions independently).

Figure 4: Whole-brain activation Social blocked conditions (S1+, S1-, S2+ & S2-): main contrasts. Whole-brain activation maps of the S1+, S2+ (social affiliative 1 & 2), S1- (social aggressive) and S2- (social escape) runs, cumulated over both monkeys, for the visual (white, Vi vs. Fx), auditory congruent (dark green, AC vs. Fx) and auditory incongruent (dark red, AI vs. Fx). Note that the AC and AI conditions contain exactly the same sound samples (coos, aggressive calls and screams). Darker shades of red indicate level of significance at $p < 0.001$ uncorrected, t-score 3.09. Lighter shades of yellow and brown outlines indicate level of significance at $p < 0.05$ FWE, t-score 4.6.

Taken together, these results indicate that audio-visual semantic associations are implemented in a specific cortical network involved in the processing of both visual face and social stimuli as well as auditory voice stimuli. An important question is thus whether these neuronal computations impact the behaviour or the physiology of the monkeys. In the following section, we investigate how heart rate changes in response to auditory-visual stimuli that are either congruent or incongruent with the social situation.

Heart rate variations depend on semantic congruence with visual context.

In this study, monkeys were required to fixate the centre of the screen while the different auditory and visual stimuli were presented. As a result, it was not possible to analyse whether gaze is spontaneously affected by the different stimulus categories. It was, however, possible to analyse heart-rate variation using a video-based method developed by our team (Froesel et al., 2020). Figure 5 focuses on heart rate variation in response to the auditory sound categories in the different blocked conditions. Although heart rate measures vary from one blocked condition to the other, in all blocked conditions, congruent auditory (Figure 5a, green) is systematically associated with lower heart rates than incongruent blocked condition (Figure 5a, red, Wilcoxon paired non-parametric test, $p < 0.001$ for all blocked conditions except the F- task, $p < 0.05$). This effect is more pronounced for the

social blocked conditions (S1+/S1- and S2+/S2-) than for the face blocked conditions (Figure 5b, F+/F-, Friedman nonparametric test, $p < 0.001$, $N = 127$, Wilcoxon $p < 0.001$), suggesting an intrinsic difference between the processing of faces and social scenes. This effect is also more pronounced for blocked conditions involving affiliative visual stimuli (F+, S1+ and S2+) than for blocked conditions involving aggressive or escape visual stimuli (Figure 5b, F-, S1- and S2-, Wilcoxon non-parametric test, $p < 0.001$). This latter interaction possibly reflects an additive effect between the semantics and emotional valence of the stimuli. Indeed, affiliative auditory stimuli are reported to decrease heart rate relative to aggressive or alarm stimuli (Kreibig, 2010). As a result, emotionally positive stimuli would enhance the semantic congruence effect, while emotionally negative stimuli would suppress the semantic congruence effect. Overall, these observations indicate that semantic congruence is perceptually salient, at least implicitly.

Figure 5: Blocked condition-related heart rate (BMP) variations. A) Absolute heart rate (BMP, beats per minute) during the congruent (green) and incongruent (red) auditory blocks of each task. Dashed lines correspond to the affiliative blocked condition as defined by the visual stimuli, whereas continuous lines refer to aggressive or escape blocked conditions. Blocked conditions are defined by pairs involving the same vocalization categories but different visual stimuli, as defined in Figure 1b. Each blocked condition pair shows significantly higher heart rates for incongruent auditory stimuli compared to congruent auditory stimuli (Friedman nonparametric test, Face: $p < 0.001$, 271.442, $N = 254$; Social 1: $p < 0.001$, 295.34, $N = 254$; Social 2: $p < 0.001$, 174.66, $N = 254$). This is also true for each individual blocked condition (Wilcoxon paired non-parametric test, $p < 0.001$ for all blocked conditions except F-: $p < 0.05$). B) Difference between AC and AI bloc means. All significantly different from zero (Wilcoxon paired non-parametric test, $p < 0.001$ for all blocked conditions except F-: $p < 0.05$).

Visual auditory gradients across the lateral sulcus (LS) and superior temporal sulcus (STS)

While LS demonstrates stronger activation for socially congruent auditory stimuli relative to visual stimuli, the STS appears to be equally activated by both sensory modalities. To better quantify this

effect, we define regions of interest (ROIs, 1.5 mm spheres) at local peak activations in the auditory congruent (AC vs Fx) contrast, in the face (F+ and F-) blocked conditions (Figure 6A, see Figure S2 for a precise localization of each of these local maxima on corresponding brain anatomy). These peaks match peak activations in the social blocked conditions (S1+, S1-, S2+ and S2-) auditory congruent (AC vs Fx) contrast. This latter social blocked condition contrast reveals two additional peaks in the right LS which were used to define two additional ROIs (right LS4 and LS6). Overall, 8 ROIs are thus defined in the right STS, 6 in the left STS, 4 in the left LS and 6 in the right LS. The numbering of these ROIs was adjusted so as to match mirror positions across hemispheres. Figure 6B presents mean percentage signal change (%PSC) for each independent ROI, in the left and right hemispheres, on each of the 5 face and social blocked conditions respectively. Overall, STS ROIs and LS ROIs had similar %PSC profiles across the 5 blocked conditions for each group of blocked conditions (face vs. social). No interhemispheric difference could be noted.

Figure 6: Percentage of signal change (%PSC) for selected left and right hemisphere ROIs in the lateral sulcus (light blue) and in the superior temporal sulci (dark blue). (A) ROIs are 1.5mm spheres located at local peak activations. Left and right hemisphere numbering associate mirror ROIs. ROI location in the each of the left and right STS and LS is described in the bottom flat maps. (B) %PSC (mean) are presented for each ROI (8 in right STS, 6 in left STS, 4 in left and 6 in right lateral sulcus)

and each blocked condition of interest (V: visual, AC: auditory congruent, AI: auditory incongruent, VAC: visuo-auditory congruent, VAI: visuo-auditory incongruent).

In the STS, in both of the face (F+ and F-) and social blocked conditions (S1+, S1-, S2+ and S2-), %PSC in the visual blocked condition relative to fixation across all ROIs is not significantly different from %PSC in the auditory congruent blocked condition relative to fixation, (Figure 7, left, Wilcoxon non-parametric test). The STS thus appears as equally responsive to visual and auditory social stimuli (%PSC of all blocked conditions are significantly different from fixation %PSC, Wilcoxon non-parametric test, $p < 0.01$ or $p < 0.001$). In contrast, in the LS, %PSC in the visual blocked condition relative to fixation across all ROIs is significantly different from %PSC in the auditory congruent blocked condition relative to fixation, (Figure 7, left, Wilcoxon non-parametric test, $p < 0.005$). This result therefore suggests a strong auditory preference for LS (%PSC of all auditory are significantly different from fixation %PSC, Wilcoxon non-parametric test, $p < 0.01$), although LS is also significantly activated by the visual stimuli in the face blocked condition ($p < 0.01$). Overall, therefore, LS appears preferentially sensitive to auditory stimuli while STS appears to be equally responsive to visual and auditory stimuli.

Figure 7: Percentage of signal change (%PSC) across all lateral sulcus (light blue) and superior temporal sulci (dark blue) ROIs of both hemispheres, comparing the auditory and visual blocked conditions. Statistical differences relative to fixation are between blocked conditions and indicated as follows: ***, $p < 0.001$; **, $p < 0.01$ (Wilcoxon non-parametric test).

Visual-auditory integration in the STS during the social blocked conditions

When processed in the brain, sensory stimuli from different modalities are combined such that the neuronal response to their combined processing is different from the sum of the neuronal responses to each one of them. This process is called multisensory integration (Avillac et al., 2007) and is more pronounced when unimodal stimuli are ambiguous or difficult to perceive (Alais & Burr, 2004; Ernst & Banks, 2002). The question here, therefore, is whether and how the LS and the STS combine visual and auditory social stimuli as a function of their semantic congruence. Multisensory integration is not

straightforward to assess based on fMRI signals. A minimal criterion here would be to have significant %PSC signal differences between the bimodal blocked conditions and both of the unimodal blocked conditions. Figure 8 shows the whole brain activation maps obtained for the visual-auditory blocked condition contrasted, with fixation, the visual blocked condition and the auditory blocked condition, for the congruent (Figure 8A) and incongruent (Figure 8B) auditory vocalizations, for the face (Figure 8, left panel) and the social blocked conditions (Figure 8, right panel). Figure 8C presents the contrast between the congruent and incongruent visuo-auditory blocked conditions.

Overall, in the face blocked condition, activations in the audio-visual conditions are not significantly different from the visual and auditory conditions alone (Figure 8A&B, left panel). Likewise, no significant difference can be seen between the congruent and incongruent visuo-auditory conditions (Figure 8C, left panel). Figure S3 compares %PSC for the bimodal and unimodal conditions across all STS selected ROIs and all LS selected ROIs. Neither reach the minimal criteria set for multisensory integration. In the social blocked condition, activation in the audio-visual conditions show local significant differences relative to the visual and auditory conditions alone (Figure 8A&B, left panel), but none in the same regions. However, when comparing the %PSC for the bimodal and unimodal conditions across all STS selected ROIs and all LS selected ROIs, the STS ROIs reach the minimal criteria set for multisensory integration, as their %PSC is significantly different from each of the bimodal conditions and each of the unimodal conditions (Wilcoxon non-parametric test, $p < 0.01$). Thus multisensory integration appears to take place, specifically in the STS, and during the social blocked conditions, possibly due to the higher ambiguity in interpreting social static scenes relative to faces (Figure S4). Importantly, and while most significant activations in the bimodal vs. unimodal auditory condition are located within the audio-visual vs. fixation network, a bilateral activation located in the anterior medial part of the LS deserves attention. Indeed, this activation, encompassing part of the insula and of anterior SII/PV, is identified both in the congruent and incongruent auditory conditions and might be involved in the interpretation of semantic congruence between the visual and auditory stimuli. This possibility is addressed next.

(social aggressive) and S2- (social escape) runs (right panel), for the congruent auditory vocalizations (green). Contrasts from top to bottom: audio-visual vs. fixation, audio-visual vs. auditory and audio-visual vs. visual. B) Same as in A) but for the incongruent auditory vocalizations (red). C) Whole-brain activation maps for the audio-visual incongruent vs audio-visual congruent contrast. All else as in A). Darker shades of red indicate level of significance at $p < 0.001$ uncorrected, t -score 3.09. Lighter shades of yellow and brown outlines indicate level of significance at $p < 0.05$ FWE, t -score 4.6.

Semantic visuo-auditory association network

Both our whole brain fMRI analyses and heart rate measures indicate that the monkeys actively associate the auditory voice stimuli with the social context set by the visual stimuli, whether faces or social scenes. To further specify the cortical network involved in the control of this semantic association, we performed a whole brain gPPI functional connectivity analysis on the LS and STS cumulated ROIs, to identify the cortical and subcortical regions that systematically correlate with either the LS or STS time series, irrespective of the blocked condition. To increase the statistical power of our analysis, we independently performed this gPPI analysis on each individual task. Table 1 of the Supplementary Material summarizes for selected cortical regions, the blocked conditions and ROIs for which a significant correlation is observed in at least one voxel ($p < 0.005$ uncorrected). In the following, we only consider for further discussion the cortical regions that correlate, with either the LS (Figure 8a) or STS time series (Figure 8b) in: (criteria-1) at least three of the six blocked conditions, for each of the left and right hemispheres; (criteria-2) at least three of the six blocked conditions in at least one hemisphere and in at least eight of the blocked conditions across both hemispheres (Figure 8, bold front). This gPPI analysis highlights a functional network connected to the LS and involving the anterior cingulate cortex (ACC), area 46 in the dorsolateral prefrontal cortex (DLPFC), the orbitofrontal cortex (OFC), the intraparietal sulcus (IPS), the superior temporal sulcus (STS), and subcortically, the amygdala and the hippocampus. The same core network is identified for its connectivity with the STS, with the addition of the insula. It is worth noting that the dorsal pulvinar, although at the limit of the inclusion criteria, was detected several times. This nucleus often acts as a attentional modulator and is implicated not only in social perception, but can also be considered as an audio-visual integrator given its connection with face patches (Schwiedrzik et al., 2015), and the auditory and visual cortex (Froese et al., 2021). Further investigation should be carried out to determine its role in this network.

Figure 9: Whole brain gPPI functional connectivity analysis for the lateral sulcus (A) and in the superior temporal sulcus ROIs (B). Reported here are functional connectivity analyses between cortical regions under the following statistical criteria. Areas in bold fonts show a significant correlation at $p < 0.005$ uncorrected in at least 4 out of the 6 individual blocked conditions, in each hemisphere (see supplementary table 1). Areas in regular font show a significant correlation at $p < 0.005$ uncorrected in at least 3 out of the 6 individual blocked conditions, in each hemisphere (see supplementary table 1). ACC: Anterior Cingulate Cortex; DLPFC: dorsolateral prefrontal cortex; OFC: orbitofrontal cortex; IPS: Intraparietal sulcus.

Discussion

Based on heart rate estimates and fMRI, our results show that rhesus monkeys systematically associate affiliative facial expressions or social scenes with corresponding affiliative vocalizations, aggressive facial expressions or social scenes with corresponding aggressive vocalizations, and escape visual scenes with scream vocalizations. In contrast, vocalizations that are incompatible with the visual information are fully suppressed, suggesting a top-down regulation over the processing of sensory input. In other words, rhesus monkeys correctly associate the meaning of a vocalization with the meaning of a visual scene. This audio-visual, semantic binding with contextual information relies on a core functional network involving the superior temporal sulcus (STS) and the lateral sulcus (LS). LS regions of interest (ROIs) have a preference for auditory and audio-visual congruent stimuli while STS ROIs respond equally to auditory, visual and audio-visual congruent stimuli. A functional connectivity analysis (gPPI) identified a functional network connected to the LS and STS, involving the anterior cingulate cortex (ACC), area 46 in the dorsolateral prefrontal cortex (DLPFC), the orbitofrontal cortex (OFC), the intraparietal sulcus (IPS), the insula and subcortically, the amygdala and the hippocampus. Overall, we propose that the integration of congruent social meaning from audio-visual information involves an emotional network composed of the STS, LS, ACC, OFC, and limbic areas, including the amygdala, and an attentional network including the STS, LS, IPS and DLPFC. These observations are highly robust as they are reproduced over six sets of independent behavioral blocked conditions, involving distinct associations of visual and auditory social information.

Interpretation of social scenes and vocalization by macaque monkeys

As is the case for human oral communication, monkey vocalizations are expected to be interpreted as a function of their emotional or contextual meaning. For example, a monkey scream indicates potential danger, is associated with fear and calls for escape and flight from the dangerous context. In contrast, coos are produced during positive social interactions and often elicit approach. Here, we show that when two different types of vocalizations are presented together with a social visual stimulus, the heart rate of the monkeys significantly decreases when the vocalization is congruent with the visual scene as opposed to incongruent. Likewise, we show that the activity of the voice processing network is dramatically suppressed in response to the incongruent vocalization. This pattern of activation provides direct neurobiological evidence that macaques infer meaning from both social auditory and visual information and are able to associate congruent information. In the network of interest, activations are not significantly different between the auditory, visual or audio-visual conditions. Most interestingly, aggressive calls are associated with both aggressive faces and aggressive social scenes, whereas coos are associated with both lipsmacks and inter-individual social

grooming. We thus propose that these networks might represent social meaning irrespective of sensory modality, thereby implying that social meaning is amodally represented. We hypothesize that such representations are ideal candidate precursors to the lexical categories that trigger, when activated, a coherent set of motor, emotional and social repertoires.

Audio-visual social stimuli robustly activate the face and voice patches

Face processing is highly specialized in the primate brain (Hesse & Tsao, 2020). In the macaque brain, it recruits a specific system called the face patch system, composed of interconnected areas, identified by both fMRI (Afraz et al., 2015; Aparicio et al., 2016; Arcaro et al., 2017; Eifuku, 2014; Freiwald & Tsao, 2010; Hadj-Bouziane et al., 2008; Issa & DiCarlo, 2012; Moeller et al., 2008; Pinsk et al., 2005, 2009; Tsao et al., 2003) and single cell recording (Grimaldi et al., 2016; Moeller et al., 2008; Tsao et al., 2006). This system recruits areas in the superior temporal sulcus, as well as in the prefrontal and orbito-frontal cortex. Specific limbic and parietal regions are also recruited together with this core system during, respectively, the emotional and attentional processing of faces (Schwiedrzik et al., 2015). The core face patches are divided into five STS areas (Anterior medial, AM; anterior fundus, AF; anterior lateral, AL; middle fundus, MF and middle lateral ML) and the PL (posterior lateral patch), a posterior face patch in the occipital cortex (Eifuku, 2014; Hesse & Tsao, 2020; Tsao et al., 2003; Tsao, Moeller, et al., 2008). Based on a review of the literature, and anatomical landmark definitions, we associate the activation peaks identified in the present study with these five face patches (Figure 10). Correspondence is unambiguous and the STS 4 ROIs matches ML, STS 7 matches MF, STS 5 matches AL and STS 6 matches AF. The occipital face patch PL is also identified in the general contrast maps as well as the frontal area defined in the literature as PA (prefrontal accurate) (Tsao, Schweers, et al., 2008). It is worth noting that in our experimental design, these face patches are activated both during the purely auditory congruent condition as well as during the visual conditions. Such activations are not reported during purely auditory conditions, indicating that this network is recruited during audio-visual association based on meaning.

Similarly to face patches, voice processing also involves a system composed of voice patches (for review, see Belin, 2017). In macaques, voice specific areas include the anterior superior temporal gyrus (aSTG), the orbitofrontal cortex (OFC) and a part of the STS close to the lateral sulcus (Cohen et al., 2007; Joly, Pallier, et al., 2012; Joly, Ramus, et al., 2012; Perrodin et al., 2015; Petkov et al., 2008; Poremba et al., 2003). The auditory processing circuit is proposed to be organized in two main networks, a ventral and a dorsal network (see for review Kuśmierek & Rauschecker, 2014), such that the auditory ventral stream is activated by species-specific vocalizations whereas the dorsal stream is involved in the spatial location of sounds (Ortiz-Rios et al., 2015; Russ et al., 2008). This functional dissociation is observable as early as in the lateral belt such that its caudal part (CL) is selectively

associated with sound location while the anterior part (AL) is more linked to sound identity such as vocalizations (Kuśmierk & Rauschecker, 2009, 2014; Tian et al., 2001). Again, based on a review of the literature, and anatomical landmark definitions, we associate the activation peaks identified in the present study with these voice patches (Figure 10). Correspondence is unambiguous and the LS 1 ROI can be associated to CL (i.e. dorsal sound processing pathway) and LS2 to core primary auditory area A1. Within the ventral sound processing pathway, LS 4 ROI can be associated to area AL, LS 5 to rostro-temporal lateral area (RTL) and LS 6 to the rostro-temporal polar field (RTp). Last, LS 6 is compatible with the anterior most voice STG area described by Petkov and colleagues (2008). The voice patch system also involves the ventral dorsolateral prefrontal cortex or vlPFC (Romanski et al., 2005), located in the inferior dimple at the boundary between area 45a and 46 (Petrides & Pandya, 2002). This cortical region has been proposed to play a key role in the cognitive control of vocalizations as well as in the interpretation of call meaning (Romanski & Averbek, 2009). Microstimulations further indicate that this prefrontal voice patch is functionally connected with the putative macaque homologue of human's Broca area 44 (Rocchi et al., 2021). In the present study, the ventral prefrontal activation, while matching nicely with the PA face patch, only partially overlaps with the prefrontal voice patch, suggesting a possible functional specialization. Taken together, these results indicate that the association between vocalization meaning and social visual stimuli recruits the face and voice patch system.

In the right hemisphere, two supplementary STS activations are reported, STS 2 and STS 3. They are located posteriorly to the putative ML face patch and possibly coincide with the gaze following patch reported in the dorsal posterior infero-temporal cortex (PITd) (Marciniak et al., 2014). This cortical region is recruited during motion discrimination blocked conditions (Stemann & Freiwald, 2016) and is proposed to play a key role in attentional selection due to its strong connectivity with the dorsal attentional areas such as lateral intraparietal areas (LIP) and frontal eye fields (FEF) (Sani et al., 2019). In spite of the fact that the monkeys did not have to produce any active behaviour, these activations possibly reflect enhanced, automatic, attention to social cues.

Visual fMRI activations have already been described in the LS, in the primary auditory cortex and in the non-primary core (belt) (Kayser et al., 2007). This observation has been confirmed using single cell recording studies (Kayser et al., 2008). In contrast, to our knowledge, no extended auditory responses have yet been described in the STS. This suggests that the STS auditory activations described here arise from the task design and the implicit association between a visual stimulus and two competing auditory stimuli. An important question to be addressed by single unit recording studies is whether these STS auditory activations correspond to neuromodulatory LFP modulations or to actual spiking activity. Quite interestingly, while we identify an audio-visual gradient between the

LS and the STS, the LS showing higher activations for voice as compared to visual social stimuli, and the STS showing equal responses to both, no clear gradient of auditory or visual activations can be identified either within the STS or within the LS. This suggests that voice-social visual associations rely on the activity of the entire network, rather than on some of its subparts.

Figure 10: Correspondence between task-related ROIs and face patches (left panels) and voice areas (right panels). Color-scale runs start at $p < 0.001$ uncorrected levels. Task related ROIs are numbered as in Figure 5. PA (prefrontal accurate); AM (anterior medial); AF (anterior fundus); AL (anterior lateral); MF (middle fundus); ML (middle lateral); PL a posterior face patch in the occipital cortex; CL (Caudal part of the lateral belt), A1 (primary auditory cortex), RTL (Rostrotemporal lateral area), STS (Superior Temporal Sulcus), STG (Superior Temporal Gyrus). *: Sources for face patch localization. **: Sources for voice areas.

Audio-visual association based on meaning and multisensory integration

The strict definition of multisensory integration involves the combination of sensory inputs from different modalities under the assumption of a common source (Lee & Noppeney, 2014; Stein et al., 2014). In this context, it has been shown that multisensory integration speeds up reaction times and enhances perception (Grant & Seitz, 2000; Lehmann & Murray, 2005; Murray et al., 2005; Raab,

1962; Welch et al., 1986), including when processing lip movement during speech (Navarra & Soto-Faraco, 2007; Shahin & Miller, 2009; Van Wassenhove et al., 2005). Multisensory processes are also at play to predict the consequences of one modality onto another, i.e. in the temporal domain (Cléry et al., 2015, 2017; Cléry et al., 2020; Guipponi et al., 2015). At the neuronal level, multisensory integration is defined as a process whereby the neuronal response to two sensory inputs is different from the sum of the neuronal responses to each on its own (Avillac et al., 2007; Stein et al., 2009). In the present study, the auditory and visual stimuli are associated based on their meaning (e.g., coos are associated with grooming) and possible contingency (e.g., screams are associated with escape scenes). In addition, incongruent auditory stimuli are actively suppressed by visual context. As a result, this association based on stimulus meaning does not correspond to the low level association classically understood by multisensory integration. Yet, one could expect that associating meaning might lead to an enhancement of neuronal processes similar to that described during multisensory integration. To probe this hypothesis, we apply the less stringent multisensory integration criteria used in fMRI studies, namely we test for audio-visual responses statistically higher (or lower) than each of the uni-sensory conditions (Beauchamp, 2005; Gentile et al., 2010; Pollick et al., 2011; Tyll et al., 2013; Werner & Noppeney, 2010). Although face-voice integration has been described in the auditory cortex (CL, CM, in awake and anesthetized monkeys; A1 only in awake monkeys) and the STS (Ghazanfar et al., 2008; Perrodin et al., 2015), and to a lesser extent in specific face-patches (Khandhadia et al., 2021), here, enhancement of the audio-visual response can only be seen in the blocked conditions involving visual scenes. The parsimonious interpretation of these observations is that face-vocalization binding was easier than scene-vocalization binding, thereby resulting in enhanced integrative processes, specifically in this latter condition, in agreement with the fact that neuronal multisensory integration is more pronounced for low saliency stimuli.

Cortical and subcortical network for social audio-visual association based on meaning

In the present study, as indexed by the heart rate and hemodynamic brain signal modulation in the LS and the STS, the social visual stimulus used in each blocked condition sets the context and the subsequent distinctive processing of congruent versus incongruent auditory vocalization stimuli. We used a gPPI in order to identify the network contributing to setting this context. Both the LS and the STS are associated with a cortico-cortical network composed of the IPS, ACC (and vmPFC), DLPFC and OFC at the cortical level (and, to a lesser extent, the insula), and the amygdala and the hippocampus at the subcortical level.

The human brain has evolved a functional specialization for processing social information, such that the auditory cortex is involved in peri-lexical speech perception, visual areas in visual perception of speech, STS in faces and lips movement processing, the limbic system (amygdala, insula and ACC) in

emotional processing and the IPS in spatially directed attention (Haxby et al., 2002; Haxby & Gobbini, 2011). A more recent review argues in favour of an interaction between attention and social processes to select information in a social environment (Capozzi & Ristic, 2018). This interaction is at play in three different levels of social processing: perception, interpretation and evaluation. First, attention acts at the perceptual level by facilitating relevant social information perception. Then a link with the emotional state of the individual and the social meaning of the cue is achieved, gating responses as a result of interpretation. Lastly, the valuation of the cue is estimated (Capozzi & Ristic, 2018). We hypothesize that the above-described network is homologous between macaques and humans, consisting of two interacting networks, one involved in the emotional processing of social stimuli, and one involved in their cognitive and attentional assessment.

We propose that the first homologous network involves the LS, the STS, the ACC (and vmPFC), the OFC, the amygdala, the hippocampus and the insula. This is in general agreement with the observation that species-specific vocalisations activate a network recruiting, in addition to the voice patches, visual areas such as V4, MT, STS areas TE and TEO, as well as areas from the limbic and paralimbic system, including the hippocampus, the amygdala and the ventromedial prefrontal cortex (vmPFC) (Gil-da-Costa et al., 2004). Brain stimulations applied to the auditory cortex directly activate vLPFC and indirectly the hippocampus (Rocchi et al., 2021). This is also in agreement with the finding that the observation of visual social interactions recruit vmPFC, vLPFC, ACC and OFC (Cléry et al., 2021; Roberts, 2006; Rudebeck et al., 2006; Rushworth et al., 2007; Sliwa & Freiwald, 2017).

We propose that the second homologous network involves the LS, the STS, the IPS and the DLPFC. DLPFC and area 46 are reciprocally connected with the caudal and rostral auditory cortex (Romanski et al., 1999). DLPFC is generally proposed to play a key role in attentional selection and memory processes (Courtney et al., 1997), and has been specifically associated with working memory during face processing (Rowe & Passingham, 2001). Spontaneous, non-trained responses to non-social (Guipponi et al., 2013; Schlack et al., 2005) and social (Joly, Pallier, et al., 2012; Ortiz-Rios et al., 2015; Poremba et al., 2003) auditory stimuli have been described in the IPS. In addition, the parieto-prefrontal network is classically associated with attentional selection (Buschman & Miller, 2007; Ibos et al., 2013) and forms with STS PITd area a larger attentional network (Sani et al., 2019). In humans, the IPS is described as a major node for the social and affective modulation by attention in naturalistic social visual information (see for review Frank & Sabatinelli, 2017), further interacting with the amygdala (for emotional processing), the hippocampus (for memory retrieval), and the OFC and PFC for top-down control over emotional processes. This is in agreement with our proposal of two homologous networks.

This homology opens the door to clinical research. Indeed, understanding these mechanisms is not only important from a comparative perspective with our own species, but may represent a fundamental contribution to issues concerning mental health. In particular, autistic individuals are often challenged by understanding social scenes, including the integration of auditory and visual information (Feldman et al., 2018; Stevenson, Siemann, Schneider, et al., 2014; Stevenson, Siemann, Woynaroski, et al., 2014a, 2014b). Such deficits may result from not only deficits in face and voice processing on their own, but the ability to integrate each modality in the service of predicting and understanding social interactions. This would implicate the two macaque networks we propose which give the possibility to test novel clinical approaches.

Conclusion

Our experiments demonstrate, using indirect measures (heart rate and hemodynamic brain response), that macaque monkeys are able to associate social auditory and visual information based on their abstract meaning. This supports the idea that non-human primates display advanced social competences, amodally represented, that may have paved the way, evolutionary, for human social cognition. We further show that these processes recruit two functional networks that are, we propose, homologous to those observed in our own species.

Contributions

Conceptualization, S.B.H. M.F.; Stimuli preparation, M.H., M.F, Q.G, M.G; Data Acquisition, M.F. M.G.; Methodology, M.F., S.C., Q.G., and S.B.H; Investigation, M.F. and S.B.H.; Writing – Original Draft, M.F and S.B.H.; Writing – Review & Editing, S.B.H., M.F., M.H.; Funding Acquisition, S.B.H.; Supervision, S.B.H.

Acknowledgements

S.B.H. were funded by the French National Research Agency (ANR)ANR-16-CE37-0009-01 grant and the LABEX CORTEX funding (ANR-11-LABX-0042) from the Université de Lyon, within the program Investissements d’Avenir (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). We thank Fidji Francioly and Laurence Boes for animal care, Julian Amengual and Justine Cléry for their rich scientific exchanges during data collection and analyses, Franck Lamberton and Danièle Ibarrola for their MRI methodological support and Holly Rayson for help on visual stimuli collection.

Ethics declarations

The authors declare no competing interests. Animal experiments were authorized by the French Ministry for Higher Education and Research (project no. 2016120910476056 and 1588-2015090114042892) in accordance with the French transposition texts of Directive 2010/63/UE. This

549 authorization was based on ethical evaluation by the French Committee on the Ethics of Experiments
550 in Animals (C2EA) CELYNE registered at the national level as C2EA number 42.

Material and methods

Subjects and surgical procedures

Two male rhesus monkeys (*Macaca mulatta*) participated in the study (T, 15 years, 10kg and S, 12, 11kg). The animals were implanted with a Peek MRI-compatible headset covered by dental acrylic. The anaesthesia for the surgery was induced by Zoletil (Tiletamine-Zolazepam, Virbac, 5 mg/kg) and maintained by isoflurane (Belamont, 1–2%). Post-surgery analgesia was ensured thanks to Temgesic (buprenorphine, 0.3 mg/ml, 0.01 mg/kg). During recovery, proper analgesic and antibiotic coverage was provided. The surgical procedures conformed to European and National Institutes of Health Guidelines for the Care and Use of Laboratory Animals. The project was authorized by the French Ministry for Higher Education and Research (project no. 2016120910476056 and 1588-2015090114042892) in accordance with the French transposition texts of Directive 2010/63/UE. This authorization was based on ethical evaluation by the French Committee on the Ethics of Experiments in Animals (C2EA) CELYNE registered at the national level as C2EA number 42.

Experimental setup

During the scanning sessions, monkeys sat in a sphinx position in a plastic monkey chair (Vanduffel et al., 2001) facing a translucent screen placed 60 cm from the eyes. Visual stimuli were retro-projected onto this translucent screen. Their head was restrained and the auditory stimuli were displayed by Sensimetrics MRI-compatible S14 insert earphones. The monkey chair was secured in the MRI with safety rubber stoppers to prevent any movement. Eye position (X, Y, right eye) was recorded thanks to a pupil-corneal reflection video-tracking system (EyeLink at 1000 Hz, SR-Research) interfaced with a program for stimulus delivery and experimental control (EventIDE®). Monkeys were rewarded for maintaining fixation into a 2x2° tolerance window around the fixation point.

General run design

On each run, monkeys were required to fixate a central cross on the screen (Figure 1A). Runs followed a block design. Each run started with 10 s of fixation in the absence of sensory stimulation followed by three repetitions of a pseudo-randomized sequence containing six possible 16 s blocks: fixation (Fx), visual (Vi), auditory congruent (AC), auditory incongruent (AI), congruent audio-visual (VAC) and incongruent audio-visual (VAI). Each block (except the fixation block) consisted in an alternation of 500 ms stimuli (except for lip smacks, 1s dynamic stimuli succession) of the same semantic category (see Stimuli section below), in the visual, auditory or audio-visual modalities. Each block ended by 10 s of fixation in the absence of sensory stimulations.

Face and social task design

Six audio-visual blocked conditions were presented to both monkeys, organized in runs as described above (Figure 1B). Six different blocked conditions were presented to both monkeys, organized in runs as described above (Figure 1B). Each task combined visual stimuli of identical social content with either semantically congruent or incongruent monkey vocalizations (Figure 1b). The face affiliative task (F+) combined lipsmacks with coos and aggressive calls. The face aggressive task (F-) combined aggressive faces with coos and aggressive calls. The first social affiliative task (S1+) combined grooming scenes with coos and aggressive calls. The second social affiliative task (S2+) combined grooming scenes with coos and screams. The social aggressive task (S1-) combined aggressive group or individual scenes with coos and aggressive calls. The social escape task (S2-) combined fleeing groups or individual scenes with coos and screams. Importantly, pairs of blocked conditions (F+ & F-; S1+ & S1-; S2+ & S2-) shared the same auditory conditions, but opposite social visual content.

Stimuli

Vocalizations were recorded from semi-free-ranging rhesus monkeys during naturally occurring situations by Marc Hauser. Detailed acoustic and functional analyses of this repertoire has been published elsewhere (e.g., Gouzoules et al., 1984; Hauser & Marler, 1993). Field recordings were then processed, restricting to selection of experimental stimuli to calls that were recorded from known individuals, in clearly identified situations, and that were free of competing noise from the environment. Exemplars from this stimulus set have already been used in several imaging studies (Belin et al., 2007; Cohen et al., 2007; Romanski, 2012; Romanski et al., 2005; Russ et al., 2008). All stimuli were normalized in intensity. The frequency ranges varied between the different types of stimuli as shown in Figure S4. For each of the three vocalization categories, we used 10 unique exemplars coming from matched male and female individuals, thus controlling for possible effects due to gender, social hierarchy or individual specificity. Coos are vocalisations typically produced during affiliative social interactions, including grooming, approach, coordinated movement, and feeding. Aggressive calls are typically used by a dominant animal toward a subordinate, often as a precursor to an actual physical attack. Screams are produced by subordinates who are either being chased or attacked, or as they are witnessing others in the same condition. Face (lipsmacks and aggressive facial expression) and social scene (group grooming, aggressive individual alone or in group / escaping individual or group) stimuli were extracted from videos collected by the Ben Hamed lab, as well as by Marc Hauser on Cayo Santiago, Puerto Rico. Images were normalized for average intensity and size. All stimuli were 4° x 4° in size. However, we decided to keep them in colour to get closer to natural stimuli even if it produced greater luminosity disparity between the different stimuli preventing us to use pupil diameter as a physiological marker. Only unambiguous

facial expressions and social scenes were retained (Figure S4). A 10% blur was applied to all images, in the hope of triggering multisensory integration processes (but see result section). For each visual category, 10 stimuli were used.

Scanning Procedures

The in-vivo MRI scans were performed on a 3T Magnetom Prisma system (Siemens Healthineers, Erlangen, Germany). For the anatomical MRI acquisitions, monkeys were first anesthetized with an intramuscular injection of ketamine (10 mg/kg). Then, the subjects were intubated and maintained under 1-2% of isoflurane. During the scan, animals were placed in a sphinx position in a Kopf MRI-compatible stereotaxic frame (Kopf Instruments, Tujunga, CA). Two L11 coils were placed on each side of the skull and a L7 coil was placed on the top of it. T1-weighted anatomical images were acquired for each subject using a magnetization-prepared rapid gradient-echo (MPRAGE) pulse sequence. Spatial resolution was set to 0.5 mm, with TR= 3000 ms, TE=3.62 ms, Inversion Time (TI)=1100 ms, flip angle=8°, bandwidth=250 Hz/pixel, 144 slices. T2-weighted anatomical images were acquired per monkey, using a Sampling Perfection with Application optimized Contrasts using different flip angle Evolution (SPACE) pulse sequence. Spatial resolution was set to 0.5 mm, with TR= 3000 ms, TE= 366.0 ms, flip angle=120°, bandwidth=710 Hz/pixel, 144 slices. Functional MRI acquisitions were as follows. Before each scanning session, a contrast agent, composed of monocrystalline iron oxide nanoparticles, Molday ION™, was injected into the animal's saphenous vein (9-11 mg/kg) to increase the signal to noise ratio (Leite et al., 2002; Vanduffel et al., 2001). We acquired gradient-echoechoplanar images covering the whole brain (TR=2000 ms; TE=18 ms; 37 sagittal slices; resolution: 1.25x1.25x1.38 mm anisotropic voxels) using an eight-channel phased-array receive coil; and a loop radial transmit-only surface coil (MRI Coil Laboratory, Laboratory for Neuro- and Psychophysiology, Katholieke Universiteit Leuven, Leuven, Belgium, see Kolster et al., 2014). The coils were placed so as to maximise the signal on the temporal lobe.

Data description

In total, 76 runs were collected in 12 sessions for monkey T and 65 runs in 9 sessions for monkey S. Based on the monkey's fixation quality during each run (85% within the eye fixation tolerance window) we selected 60 runs from monkey T and 59 runs for monkey S in total, i.e. 10 runs per task, except for one task of monkey S.

Data analysis

Data were pre-processed and analysed using AFNI (Cox, 1996), FSL (Jenkinson et al., 2012; Smith et al., 2013), SPM software (version SPM12, Wellcome Department of Cognitive Neurology, London, UK, <https://www.fil.ion.ucl.ac.uk/spm/software/>), JIP analysis toolkit (<http://www.nitrc.org/projects/jip>)

and Workbench (<https://www.humanconnectome.org/software/get-connectome-workbench>). The T1-weighted and T2-weighted anatomical images were processed according to the HCP pipeline (Autio et al., 2020; Glasser et al., 2013) and were normalized into the MY19 Atlas (Donahue et al., 2016). Functional volumes were corrected for head motion and slice time and skull-stripped. They were then linearly realigned on the T2-weighted anatomical image with flirt from FSL, the image distortions were corrected using nonlinear warping with JIP. A spatial smoothing was applied with a 3-mm FWHM Gaussian Kernel.

Fixed effect individual analyses were performed for each monkey, with a level of significance set at $p < 0.05$ corrected for multiple comparisons (FWE, t-scores 4.6) and $p < 0.001$ (uncorrected level, t-scores 3.09). Head motion and eye movements were included as covariate of no interest. Because of the contrast agent injection, a specific MION hemodynamic response function (HRF) (Vanduffel et al., 2001) was used instead of the BOLD HRF provided by SPM. The main effects were computed over both monkeys. In most analyses, face blocked conditions and social blocked conditions were independently pooled.

ROI analyses were performed as follows. ROIs were determined from the auditory congruent contrast (AC vs Fx) of face blocked conditions with the exception of two ROIs of the right lateral sulcus (LS4 and LS6) that were defined from the same contrast of social blocked conditions. ROIs were defined as 1.5 mm diameter spheres centred around the local peaks of activation. In total, 8 ROIs were selected in the right STS, 6 from the left STS, 4 in the left LS and 6 in the right LS. Figure S1 shows the peak activations defining each selected ROI; so as to confirm the location of the peak activation on either of the inferior LS bank, the superior STS bank or the inferior STS bank. For each ROI, the activity profiles were extracted with the Marsbar SPM toolbox (marsbar.sourceforge.net) and the mean percent of signal change (+/- standard error of the mean across runs) was calculated for each condition relative to the fixation baseline. %PSC were compared using Wilcoxon non-parametric paired tests.

Generalized Form of Context-Dependent Psychophysiological Interactions was performed as follows (gPPI, <http://www.nitrc.org/projects/gppi>), using the CONN toolbox (www.nitrc.org/projects/conn, RRID:SCR_009550), an open-source Matlab/SPM-based cross-platform software. Specifically, we were interested in identifying the network activated throughout the runs and that could account for the dependence of auditory perception on the visual context of the task. The gPPI analysis was performed independently on each task, over the averaged LS and STS ROIs time series respectively, for each hemisphere. We report, in supplementary table 1 the cortical and subcortical regions the time series of which showed a significant temporal correlation with the seeds ($p = 0.005$ uncorrected). Are considered for discussion only the cortical regions the time series of which correlate, with either

the LS or STS time series, in at least three of the six blocked conditions, for each of the left and right hemispheres (criterion 1), or in at least three of the six blocked conditions, in at least one hemisphere and in at least eight of the blocked conditions across both hemispheres (criterion 2, more stringent).

Behaviour and Heart rate

During each run of acquisition, videos of the faces of monkeys S and T were recorded in order to track heart rate variations (HRV) as a function of blocked conditions and blocks (Froesel et al., 2020). We focus on heart rate variations between auditory congruent and incongruent stimuli. For each task, we extracted HRV during AC and AI blocs. As changes in cardiac rhythm are slow, analyses were performed over the second half (8s of each block). This has been done for each run of each task, grouping both monkeys. Because the data were not normally distributed (Kolmogorov-Smirnov Test of Normality), we carried out Friedman tests and non-parametric post hoc tests.

References

- Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, 112(21), 6730–6735. <https://doi.org/10.1073/pnas.1423328112>
- Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, 14(3), 257–262. <https://doi.org/10.1016/j.cub.2004.01.029>
- Aparicio, P. L., Issa, E. B., & DiCarlo, J. J. (2016). Neurophysiological Organization of the Middle Face Patch in Macaque Inferior Temporal Cortex. *Journal of Neuroscience*, 36(50), 12729–12745. <https://doi.org/10.1523/JNEUROSCI.0237-16.2016>
- Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., & Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience*, 20(10), 1404–1412. <https://doi.org/10.1038/nn.4635>
- Autio, J. A., Glasser, M. F., Ose, T., Donahue, C. J., Bastiani, M., Ohno, M., Kawabata, Y., Urushibata, Y., Murata, K., Nishigori, K., Yamaguchi, M., Hori, Y., Yoshida, A., Go, Y., Coalson, T. S., Jbabdi, S., Sotiropoulos, S. N., Kennedy, H., Smith, S., ... Hayashi, T. (2020). Towards HCP-Style

712 macaque connectomes: 24-Channel 3T multi-array coil, MRI sequences and preprocessing.
713 *NeuroImage*, 215, 116800. <https://doi.org/10.1016/j.neuroimage.2020.116800>

714 Avillac, M., Ben Hamed, S., & Duhamel, J.-R. (2007). Multisensory Integration in the Ventral
715 Intraparietal Area of the Macaque Monkey. *Journal of Neuroscience*, 27(8), 1922–1932.
716 <https://doi.org/10.1523/JNEUROSCI.2646-06.2007>

717 Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-
718 temporal cortex. *Current Opinion in Neurobiology*, 15(2), 145–153.
719 <https://doi.org/10.1016/j.conb.2005.03.011>

720 Belin, P., Fecteau, S., Charest, I., Nicastro, N., Hauser, M. D., & Armony, J. L. (2007). Human cerebral
721 response to animal affective vocalizations. *Proceedings of the Royal Society B: Biological*
722 *Sciences*. <https://doi.org/10.1098/rspb.2007.1460>

723 Buschman, T. J., & Miller, E. K. (2007). Top-Down Versus Bottom-Up Control of Attention in the
724 Prefrontal and Posterior Parietal Cortices. *Science*, 315(5820), 1860–1862.
725 <https://doi.org/10.1126/science.1138071>

726 Capozzi, F., & Ristic, J. (2018). How attention gates social interactions. *Annals of the New York*
727 *Academy of Sciences*. <https://doi.org/10.1111/nyas.13854>

728 Cléry, J., Hori, Y., Schaeffer, D. J., Menon, R. S., & Everling, S. (2021). Neural network of social
729 interaction observation in marmosets. *eLife*, 10, e65012.
730 <https://doi.org/10.7554/eLife.65012>

731 Cléry, J., Schaeffer, D. J., Hori, Y., Gilbert, K. M., Hayrynen, L. K., Gati, J. S., Menon, R. S., & Everling, S.
732 (2020). Looming and receding visual networks in awake marmosets investigated with fMRI.
733 *NeuroImage*, 215, 116815. <https://doi.org/10.1016/j.neuroimage.2020.116815>

734 Cléry, J., Guipponi, O., Odouard, S., Pinède, S., Wardak, C., & Ben Hamed, S. (2017). The Prediction of
735 Impact of a Looming Stimulus onto the Body Is Subserved by Multisensory Integration
736 Mechanisms. *Journal of Neuroscience*, 37(44), 10656–10670.
737 <https://doi.org/10.1523/JNEUROSCI.0610-17.2017>

738 Cléry, J., Guipponi, O., Odouard, S., Wardak, C., & Ben Hamed, S. (2015). Impact Prediction by
739 Looming Visual Stimuli Enhances Tactile Detection. *Journal of Neuroscience*, 35(10),
740 4179–4189. <https://doi.org/10.1523/JNEUROSCI.3031-14.2015>

741 Cohen, Y. E., Theunissen, F., Russ, B. E., & Gill, P. (2007). Acoustic Features of Rhesus Vocalizations
742 and Their Representation in the Ventrolateral Prefrontal Cortex. *Journal of Neurophysiology*,
743 97(2), 1470–1484. <https://doi.org/10.1152/jn.00769.2006>

744 Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a
745 distributed neural system for human working memory. *Nature*, 386(6625), 608–611.
746 <https://doi.org/10.1038/386608a0>

747 Devaine, M., San-Galli, A., Trapanese, C., Bardino, G., Hano, C., Jalme, M. S., Bouret, S., Masi, S., &
748 Daunizeau, J. (2017). Reading wild minds: A computational assay of Theory of Mind
749 sophistication across seven primate species. *PLOS Computational Biology*, 13(11), e1005833.
750 <https://doi.org/10.1371/journal.pcbi.1005833>

751 Donahue, C. J., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Behrens, T. E., Dyrby, T. B.,
752 Coalson, T., Kennedy, H., Knoblauch, K., Essen, D. C. V., & Glasser, M. F. (2016). Using
753 Diffusion Tractography to Predict Cortical Connection Strength and Distance: A Quantitative
754 Comparison with Tracers in the Monkey. *Journal of Neuroscience*, 36(25), 6758–6770.
755 <https://doi.org/10.1523/JNEUROSCI.0493-16.2016>

756 Eifuku, S. (2014). Neural representations of perceptual and semantic identities of individuals in the
757 anterior ventral inferior temporal cortex of monkeys. *Japanese Psychological Research*, 56(1),
758 58–75. <https://doi.org/10.1111/jpr.12026>

759 Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically
760 optimal fashion. *Nature*, 415(6870), 429–433. <https://doi.org/10.1038/415429a>

761 Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., & Woynarowski, T. G. (2018). Audiovisual
762 multisensory integration in individuals with autism spectrum disorder: A systematic review

763 and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 95, 220–234.

764 <https://doi.org/10.1016/j.neubiorev.2018.09.020>

765 Fox, K. C. R., Muthukrishna, M., & Shultz, S. (2017). The social and cultural roots of whale and dolphin

766 brains. *Nature Ecology & Evolution*, 1(11), 1699–1705. [https://doi.org/10.1038/s41559-017-](https://doi.org/10.1038/s41559-017-0336-y)

767 0336-y

768 Frank, D. W., & Sabatinelli, D. (2017). Primate Visual Perception: Motivated Attention in Naturalistic

769 Scenes. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00226>

770 Freiwald, W. A. (2020). Social interaction networks in the primate brain. *Current Opinion in*

771 *Neurobiology*, 65, 49–58. <https://doi.org/10.1016/j.conb.2020.08.012>

772 Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization

773 within the macaque face-processing system. *Science (New York, N.Y.)*, 330(6005), 845–851.

774 <https://doi.org/10.1126/science.1194908>

775 Froesel, M., Cappe, C., & Ben Hamed, S. (2021). A multisensory perspective onto primate pulvinar

776 functions. *Neuroscience & Biobehavioral Reviews*.

777 <https://doi.org/10.1016/j.neubiorev.2021.02.043>

778 Froesel, M., Goudard, Q., Hauser, M., Gacoin, M., & Ben Hamed, S. (2020). Automated video-based

779 heart rate tracking for the anesthetized and behaving monkey. *Scientific Reports*, 10(1),

780 17940. <https://doi.org/10.1038/s41598-020-74954-5>

781 Gentile, G., Petkova, V. I., & Ehrsson, H. H. (2010). Integration of Visual and Tactile Signals From the

782 Hand in the Human Brain: An fMRI Study. *Journal of Neurophysiology*, 105(2), 910–922.

783 <https://doi.org/10.1152/jn.00840.2010>

784 Ghazanfar, A. A. (2009). The multisensory roles for auditory cortex in primate vocal communication.

785 *Hearing Research*, 258(1), 113–120. <https://doi.org/10.1016/j.heares.2009.04.003>

786 Ghazanfar, A. A., Chandrasekaran, C., & Logothetis, N. K. (2008). Interactions between the Superior

787 Temporal Sulcus and Auditory Cortex Mediate Dynamic Face/Voice Integration in Rhesus

788 Monkeys. *Journal of Neuroscience*, 28(17), 4457–4469.

789 <https://doi.org/10.1523/JNEUROSCI.0541-08.2008>

790 Ghazanfar, A. A., & Hauser, M. D. (1999). The neuroethology of primate vocal communication: Substrates for the evolution of speech. *Trends in Cognitive Sciences*, 3(10), 377–384.

791 [https://doi.org/10.1016/S1364-6613\(99\)01379-0](https://doi.org/10.1016/S1364-6613(99)01379-0)

792 Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *Journal of Neuroscience*, 25(20), 5004–5012. <https://doi.org/10.1523/JNEUROSCI.0799-05.2005>

793 Ghazanfar, A. A., & Santos, L. R. (2004). Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience*, 5(8), 603–616. <https://doi.org/10.1038/nrn1473>

794 Gil-da-Costa, R., Braun, A., Lopes, M., Hauser, M. D., Carson, R. E., Herscovitch, P., & Martin, A. (2004). Toward an evolutionary perspective on conceptual representation: Species-specific calls activate visual and affective processing systems in the macaque. *Proceedings of the National Academy of Sciences*, 101(50), 17516–17521.

795 <https://doi.org/10.1073/pnas.0408077101>

796 Gil-da-Costa, R., Martin, A., Lopes, M. A., Muñoz, M., Fritz, J. B., & Braun, A. R. (2006). Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. *Nature Neuroscience*, 9(8), 1064–1070. <https://doi.org/10.1038/nn1741>

797 Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80, 105–124.

798 <https://doi.org/10.1016/j.neuroimage.2013.04.127>

799 Gothard, K. M., Brooks, K. N., & Peterson, M. A. (2009). Multiple perceptual strategies used by macaque monkeys for face recognition. *Animal Cognition*, 12(1), 155–167.

800 <https://doi.org/10.1007/s10071-008-0179-7>

813 Gothard, K. M., Erickson, C. A., & Amaral, D. G. (2004). How do rhesus monkeys (*Macaca mulatta*)
814 scan faces in a visual paired comparison task? *Animal Cognition*, 7(1), 25–36.
815 <https://doi.org/10.1007/s10071-003-0179-6>

816 Gouzoules, S., Gouzoules, H., & Marler, P. (1984). Rhesus monkey (*Macaca mulatta*) screams:
817 Representational signalling in the recruitment of agonistic aid. *Animal Behaviour*, 32(1),
818 182–193. [https://doi.org/10.1016/S0003-3472\(84\)80336-X](https://doi.org/10.1016/S0003-3472(84)80336-X)

819 Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of
820 spoken sentences. *The Journal of the Acoustical Society of America*, 108(3 Pt 1), 1197–1208.
821 <https://doi.org/10.1121/1.1288668>

822 Grimaldi, P., Saleem, K. S., & Tsao, D. (2016). Anatomical Connections of the Functionally Defined
823 « Face Patches » in the Macaque Monkey. *Neuron*, 90(6), 1325–1342.
824 <https://doi.org/10.1016/j.neuron.2016.05.009>

825 Guipponi, O., Odouard, S., Pinède, S., Wardak, C., & Ben Hamed, S. (2015). FMRI Cortical Correlates
826 of Spontaneous Eye Blinks in the Nonhuman Primate. *Cerebral Cortex*, 25(9), 2333–2345.
827 <https://doi.org/10.1093/cercor/bhu038>

828 Guipponi, O., Wardak, C., Ibarrola, D., Comte, J.-C., Sappey-Marinier, D., Pinède, S., & Ben Hamed, S.
829 (2013). Multimodal Convergence within the Intraparietal Sulcus of the Macaque Monkey.
830 *Journal of Neuroscience*, 33(9), 4128–4139. [https://doi.org/10.1523/JNEUROSCI.1421-](https://doi.org/10.1523/JNEUROSCI.1421-12.2013)
831 12.2013

832 Hadj-Bouziane, F., Bell, A. H., Knusten, T. A., Ungerleider, L. G., & Tootell, R. B. H. (2008). Perception
833 of emotional expressions is independent of face selectivity in monkey inferior temporal
834 cortex. *Proceedings of the National Academy of Sciences*, 105(14), 5591–5596.
835 <https://doi.org/10.1073/pnas.0800489105>

836 Hauser, M. D., & Marler, P. (1993). Food-associated calls in rhesus macaques (*Macaca mulatta*): I.
837 Socioecological factors. *Behavioral Ecology*, 4(3), 194–205.
838 <https://doi.org/10.1093/beheco/4.3.194>

839 Haxby, J. V., & Gobbini, M. I. (2011, juillet 28). *Distributed Neural Systems for Face Perception*. Oxford
840 Handbook of Face Perception. <https://doi.org/10.1093/oxfordhb/9780199559053.013.0006>
841 Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and
842 social communication. *Biological Psychiatry*, 51(1), 59–67. <https://doi.org/10.1016/S0006->
843 3223(01)01330-0
844 Hesse, J. K., & Tsao, D. Y. (2020). The macaque face patch system: A turtle’s underbelly for the
845 brain. *Nature Reviews Neuroscience*, 1–22. <https://doi.org/10.1038/s41583-020-00393-w>
846 Ibos, G., Duhamel, J.-R., & Ben Hamed, S. (2013). A Functional Hierarchy within the Parietofrontal
847 Network in Stimulus Selection and Attention Control. *Journal of Neuroscience*, 33(19),
848 8359–8369. <https://doi.org/10.1523/JNEUROSCI.4058-12.2013>
849 Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the Eye Region in Neural Processing of Faces. *Journal*
850 *of Neuroscience*, 32(47), 16666–16682. <https://doi.org/10.1523/JNEUROSCI.2391-12.2012>
851 Joly, O., Pallier, C., Ramus, F., Pressnitzer, D., Vanduffel, W., & Orban, G. A. (2012). Processing of
852 vocalizations in humans and monkeys: A comparative fMRI study. *NeuroImage*, 62(3),
853 1376–1389. <https://doi.org/10.1016/j.neuroimage.2012.05.070>
854 Joly, O., Ramus, F., Pressnitzer, D., Vanduffel, W., & Orban, G. A. (2012). Interhemispheric Differences
855 in Auditory Processing Revealed by fMRI in Awake Rhesus Monkeys. *Cerebral Cortex*, 22(4),
856 838–853. <https://doi.org/10.1093/cercor/bhr150>
857 Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional Imaging Reveals Visual
858 Modulation of Specific Fields in Auditory Cortex. *Journal of Neuroscience*, 27(8), 1824–1835.
859 <https://doi.org/10.1523/JNEUROSCI.4737-06.2007>
860 Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual Modulation of Neurons in Auditory Cortex.
861 *Cerebral Cortex*, 18(7), 1560–1574. <https://doi.org/10.1093/cercor/bhm187>
862 Khandhadia, A. P., Murphy, A. P., Romanski, L. M., Bizley, J. K., & Leopold, D. A. (2021). Audiovisual
863 integration in macaque face patch neurons. *Current Biology*.
864 <https://doi.org/10.1016/j.cub.2021.01.102>

865 Kolster, H., Janssens, T., Orban, G. A., & Vanduffel, W. (2014). The retinotopic organization of
866 macaque occipitotemporal cortex anterior to V4 and caudoventral to the middle temporal
867 (MT) cluster. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*,
868 34(31), 10168–10191. <https://doi.org/10.1523/JNEUROSCI.3288-13.2014>

869 Kreibitz, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological*
870 *Psychology*, 84(3), 394–421. <https://doi.org/10.1016/j.biopsycho.2010.03.010>

871 Kuśmirek, P., & Rauschecker, J. P. (2009). Functional Specialization of Medial Auditory Belt Cortex in
872 the Alert Rhesus Monkey. *Journal of Neurophysiology*, 102(3), 1606–1622.
873 <https://doi.org/10.1152/jn.00167.2009>

874 Kuśmirek, P., & Rauschecker, J. P. (2014). Selectivity for space and time in early areas of the
875 auditory dorsal stream in the rhesus monkey. *Journal of Neurophysiology*, 111(8),
876 1671–1685. <https://doi.org/10.1152/jn.00436.2013>

877 Lee, H., & Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Current*
878 *Biology*, 24(8), R309–R310. <https://doi.org/10.1016/j.cub.2014.02.007>

879 Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object
880 discrimination. *Cognitive Brain Research*, 24(2), 326–334.
881 <https://doi.org/10.1016/j.cogbrainres.2005.02.005>

882 Leite, F. P., Tsao, D., Vanduffel, W., Fize, D., Sasaki, Y., Wald, L. L., Dale, A. M., Kwong, K. K., Orban, G.
883 A., Rosen, B. R., Tootell, R. B. H., & Mandeville, J. B. (2002). Repeated fMRI using iron oxide
884 contrast agent in awake, behaving macaques at 3 Tesla. *NeuroImage*, 16(2), 283–294.
885 <https://doi.org/10.1006/nimg.2002.1110>

886 Marciniak, K., Atabaki, A., Dicke, P. W., & Thier, P. (2014). Disparate substrates for head gaze
887 following and face perception in the monkey superior temporal sulcus. *eLife*, 3, e03222.
888 <https://doi.org/10.7554/eLife.03222>

889 Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. S. (2012). On the
890 relationship between the “default mode network” and the “social brain”. *Frontiers in Human*
891 *Neuroscience*, 6. <https://doi.org/10.3389/fnhum.2012.00189>

892 Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with Links: A Unified System for
893 Processing Faces in the Macaque Temporal Lobe. *Science*, 320(5881), 1355–1359.
894 <https://doi.org/10.1126/science.1157436>

895 Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., Schroeder, C. E.,
896 & Foxe, J. J. (2005). Grabbing Your Ear: Rapid Auditory–Somatosensory Multisensory
897 Interactions in Low-level Sensory Cortices Are Not Constrained by Stimulus Alignment.
898 *Cerebral Cortex*, 15(7), 963–974. <https://doi.org/10.1093/cercor/bhh197>

899 Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory
900 information enables the perception of second language sounds. *Psychological Research*,
901 71(1), 4–12. <https://doi.org/10.1007/s00426-005-0031-5>

902 Ortiz-Rios, M., Kuśmierek, P., DeWitt, I., Archakov, D., Azevedo, F. A. C., Sams, M., Jääskeläinen, I. P.,
903 Keliris, G. A., & Rauschecker, J. P. (2015). Functional MRI of the vocalization-processing
904 network in the macaque brain. *Frontiers in Neuroscience*, 9.
905 <https://doi.org/10.3389/fnins.2015.00113>

906 Parr, L. A., Waller, B. M., & Fugate, J. (2005). Emotional communication in primates: Implications for
907 neurobiology. *Current opinion in neurobiology*, 15(6), 716–720.
908 <https://doi.org/10.1016/j.conb.2005.10.017>

909 Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2014). Auditory and visual modulation of
910 temporal lobe neurons in voice-sensitive and association cortices. *The Journal of*
911 *Neuroscience: The Official Journal of the Society for Neuroscience*, 34(7), 2524–2537.
912 <https://doi.org/10.1523/JNEUROSCI.2805-13.2014>

913 Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2015). Natural asynchronies in audiovisual
914 communication signals regulate neuronal multisensory interactions in voice-sensitive cortex.

915 *Proceedings of the National Academy of Sciences*, 112(1), 273–278.

916 <https://doi.org/10.1073/pnas.1412817112>

917 Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N. K. (2008). A voice

918 region in the monkey brain. *Nature Neuroscience*, 11(3), 367–374.

919 <https://doi.org/10.1038/nn2043>

920 Petrides, M., & Pandya, D. N. (2002). Comparative cytoarchitectonic analysis of the human and the

921 macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the

922 monkey. *The European Journal of Neuroscience*, 16(2), 291–310.

923 <https://doi.org/10.1046/j.1460-9568.2001.02090.x>

924 Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009).

925 Neural Representations of Faces and Body Parts in Macaque and Human Cortex: A

926 Comparative fMRI Study. *Journal of Neurophysiology*, 101(5), 2581–2600.

927 <https://doi.org/10.1152/jn.91198.2008>

928 Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and

929 body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the*

930 *National Academy of Sciences of the United States of America*, 102(19), 6996–7001.

931 <https://doi.org/10.1073/pnas.0502605102>

932 Pollick, F., Love, S., & Latinus, M. (2011). Cerebral Correlates and Statistical Criteria of Cross-Modal

933 Face and Voice Integration. *Seeing and Perceiving*, 24(4), 351–367.

934 <https://doi.org/10.1163/187847511X584452>

935 Poremba, A., Malloy, M., Saunders, R. C., Carson, R. E., Herscovitch, P., & Mishkin, M. (2004). Species-

936 specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature*, 427(6973),

937 448–451. <https://doi.org/10.1038/nature02268>

938 Poremba, A., Saunders, R. C., Crane, A. M., Cook, M., Sokoloff, L., & Mishkin, M. (2003). Functional

939 Mapping of the Primate Auditory System. *Science*, 299(5606), 568–572.

940 <https://doi.org/10.1126/science.1078900>

941 Raab, D. H. (1962). Division of Psychology: Statistical Facilitation of Simple Reaction Times*.
942 *Transactions of the New York Academy of Sciences*, 24(5 Series II), 574-590.
943 <https://doi.org/10.1111/j.2164-0947.1962.tb01433.x>

944 Rendall, D., Rodman, P. S., & Emond, R. E. (1996). Vocal recognition of individuals and kin in free-
945 ranging rhesus monkeys. *Animal Behaviour*, 51(5), 1007-1015.
946 <https://doi.org/10.1006/anbe.1996.0103>

947 Roberts, A. C. (2006). Primate orbitofrontal cortex and adaptive behaviour. *Trends in Cognitive*
948 *Sciences*, 10(2), 83-90. <https://doi.org/10.1016/j.tics.2005.12.002>

949 Rocchi, F., Oya, H., Balezeau, F., Billig, A. J., Kocsis, Z., Jenison, R. L., Nourski, K. V., Kovach, C. K.,
950 Steinschneider, M., Kikuchi, Y., Rhone, A. E., Dlouhy, B. J., Kawasaki, H., Adolphs, R.,
951 Greenlee, J. D. W., Griffiths, T. D., Howard, M. A., & Petkov, C. I. (2021). Common fronto-
952 temporal effective connectivity in humans and monkeys. *Neuron*, 109(5), 852-868.e8.
953 <https://doi.org/10.1016/j.neuron.2020.12.026>

954 Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual
955 streams of auditory afferents target multiple domains in the primate prefrontal cortex.
956 *Nature Neuroscience*, 2(12), 1131-1136. <https://doi.org/10.1038/16056>

957 Romanski, L. M. (2012). Integration of faces and vocalizations in ventral prefrontal cortex:
958 Implications for the evolution of audiovisual speech. *Proceedings of the National Academy of*
959 *Sciences*, 109(Supplement 1), 10717-10724. <https://doi.org/10.1073/pnas.1204335109>

960 Romanski, L. M., & Averbeck, B. B. (2009). The Primate Cortical Auditory System and Neural
961 Representation of Conspecific Vocalizations. *Annual Review of Neuroscience*, 32(1), 315-346.
962 <https://doi.org/10.1146/annurev.neuro.051508.135431>

963 Romanski, L. M., Averbeck, B. B., & Diltz, M. (2005). Neural Representation of Vocalizations in the
964 Primate Ventrolateral Prefrontal Cortex. *Journal of Neurophysiology*, 93(2), 734-747.
965 <https://doi.org/10.1152/jn.00675.2004>

966 Rowe, J. B., & Passingham, R. E. (2001). Working Memory for Location and Time: Activity in
967 Prefrontal Area 46 Relates to Selection Rather than Maintenance in Memory. *NeuroImage*,
968 14(1), 77–86. <https://doi.org/10.1006/nimg.2001.0784>

969 Rudebeck, P. H., Buckley, M. J., Walton, M. E., & Rushworth, M. F. S. (2006). A Role for the Macaque
970 Anterior Cingulate Gyrus in Social Valuation. *Science*, 313(5791), 1310–1312.
971 <https://doi.org/10.1126/science.1128197>

972 Rushworth, M. F. S., Behrens, T. E. J., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for
973 cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends in Cognitive*
974 *Sciences*, 11(4), 168–176. <https://doi.org/10.1016/j.tics.2007.01.004>

975 Russ, B. E., Ackelson, A. L., Baker, A. E., & Cohen, Y. E. (2008). Coding of Auditory-Stimulus Identity in
976 the Auditory Non-Spatial Processing Stream. *Journal of Neurophysiology*, 99(1), 87–95.
977 <https://doi.org/10.1152/jn.01069.2007>

978 Sani, I., McPherson, B. C., Stemmann, H., Pestilli, F., & Freiwald, W. A. (2019). Functionally defined
979 white matter of the macaque monkey brain reveals a dorso-ventral attention network. *eLife*,
980 8, e40520. <https://doi.org/10.7554/eLife.40520>

981 Schlack, A., Sterbing-D'Angelo, S. J., Hartung, K., Hoffmann, K.-P., & Bremmer, F. (2005). Multisensory
982 Space Representations in the Macaque Ventral Intraparietal Area. *Journal of Neuroscience*,
983 25(18), 4616–4625. <https://doi.org/10.1523/JNEUROSCI.0455-05.2005>

984 Schwiedrzik, C. M., Zarco, W., Everling, S., & Freiwald, W. A. (2015). Face Patch Resting State
985 Networks Link Face Processing to Social Cognition. *PLOS Biology*, 13(9), e1002245.
986 <https://doi.org/10.1371/journal.pbio.1002245>

987 Shahin, A. J., & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *The*
988 *Journal of the Acoustical Society of America*, 125(3), 1744–1750.
989 <https://doi.org/10.1121/1.3075576>

990 Shultz, S., & Dunbar, R. (2010). Encephalization is not a universal macroevolutionary phenomenon in
991 mammals but is associated with sociality. *Proceedings of the National Academy of Sciences*,
992 107(50), 21582–21586. <https://doi.org/10.1073/pnas.1005246107>

993 Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the
994 primate brain. *Science*, 356(6339), 745–749. <https://doi.org/10.1126/science.aam6383>

995 Sliwa, J., Duhamel, J.-R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice–face identity matching by
996 rhesus monkeys for familiar conspecifics and humans. *Proceedings of the National Academy*
997 *of Sciences*, 108(4), 1735–1740. <https://doi.org/10.1073/pnas.1008169108>

998 Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., & Rowland, B. A. (2009). Challenges in
999 quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness.
1000 *Experimental Brain Research*, 198(2), 113. <https://doi.org/10.1007/s00221-009-1880-8>

1001 Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from
1002 the perspective of the individual neuron. *Nature Reviews Neuroscience*, 15(8), 520–535.
1003 <https://doi.org/10.1038/nrn3742>

1004 Stemmann, H., & Freiwald, W. A. (2016). Attentive Motion Discrimination Recruits an Area in
1005 Inferotemporal Cortex. *Journal of Neuroscience*, 36(47), 11918–11928.
1006 <https://doi.org/10.1523/JNEUROSCI.1888-16.2016>

1007 Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynarowski, T. G., Camarata, S. M., &
1008 Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *The*
1009 *Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 34(3), 691–697.
1010 <https://doi.org/10.1523/JNEUROSCI.3615-13.2014>

1011 Stevenson, R. A., Siemann, J. K., Woynarowski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., &
1012 Wallace, M. T. (2014a). Brief report: Arrested development of audiovisual speech
1013 perception in autism spectrum disorders. *Journal of Autism and Developmental Disorders*,
1014 44(6), 1470–1477. <https://doi.org/10.1007/s10803-013-1992-7>

1015 Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., &
1016 Wallace, M. T. (2014b). Evidence for diminished multisensory integration in autism spectrum
1017 disorders. *Journal of Autism and Developmental Disorders*, 44(12), 3161–3167.
1018 <https://doi.org/10.1007/s10803-014-2179-6>

1019 Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional Specialization in
1020 Rhesus Monkey Auditory Cortex. *Science*, 292(5515), 290–293.
1021 <https://doi.org/10.1126/science.1058911>

1022 Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. H. (2003). Faces and
1023 objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9), 989.
1024 <https://doi.org/10.1038/nn1111>

1025 Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A Cortical Region Consisting
1026 Entirely of Face-Selective Cells. *Science (New York, N.Y.)*, 311(5761), 670–674.
1027 <https://doi.org/10.1126/science.1119983>

1028 Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and
1029 humans. *Proceedings of the National Academy of Sciences*, 105(49), 19514–19519.
1030 <https://doi.org/10.1073/pnas.0809662105>

1031 Tsao, D. Y., Schweers, N., Moeller, S., & Freiwald, W. A. (2008). Patches of face-selective cortex in the
1032 macaque frontal lobe. *Nature Neuroscience*, 11(8), 877–879.
1033 <https://doi.org/10.1038/nn.2158>

1034 Tyll, S., Bonath, B., Schoenfeld, M. A., Heinze, H.-J., Ohl, F. W., & Noesselt, T. (2013). Neural basis of
1035 multisensory looming signals. *NeuroImage*, 65, 13–22.
1036 <https://doi.org/10.1016/j.neuroimage.2012.09.056>

1037 Van Essen, D. C., & Dierker, D. L. (2007). Surface-Based and Probabilistic Atlases of Primate Cerebral
1038 Cortex. *Neuron*, 56(2), 209–225. <https://doi.org/10.1016/j.neuron.2007.10.015>

1039 Vanduffel, W., Fize, D., Mandeville, J. B., Nelissen, K., Hecke, P. V., Rosen, B. R., Tootell, R. B. H., &
1040 Orban, G. A. (2001). Visual Motion Processing Investigated Using Contrast Agent-Enhanced

1041 fMRI in Awake Behaving Monkeys. *Neuron*, 32(4), 565–577. <https://doi.org/10.1016/S0896->
1042 6273(01)00502-5

1043 Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural
1044 processing of auditory speech. *Proceedings of the National Academy of Sciences*, 102(4),
1045 1181–1186. <https://doi.org/10.1073/pnas.0408949102>

1046 Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to
1047 temporal rate perception. *Perception & Psychophysics*, 39(4), 294–300.
1048 <https://doi.org/10.3758/BF03204939>

1049 Werner, S., & Noppeney, U. (2010). Superadditive Responses in Superior Temporal Sulcus Predict
1050 Audiovisual Benefits in Object Categorization. *Cerebral Cortex*, 20(8), 1829–1842.
1051 <https://doi.org/10.1093/cercor/bhp248>

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

Supplementary Material

Figure S1: Auditory activations depend on semantic congruence with visual context in social blocked conditions. A) Whole-brain activation maps of the S1+, S2+ (social affiliative 1 & 2), S1- (social aggressive) and S2- (social escape) runs, for the *auditory congruent vs auditory incongruent* (relative to the visual context) contrast. B) Whole-brain activation map for the S+ (social affiliative, S1+&S2+) auditory congruent (coos, dark green, AC vs. Fx) and auditory incongruent (aggressive calls and screams, dark red, AI vs. Fx) conditions. C) Whole-brain activation map for the S- (social negative, S1-&S2-) auditory congruent (aggressive calls and screams, dark green, AC vs. Fx) and auditory incongruent (coos, dark red, AI vs. Fx) conditions. Darker shades of red indicate level of significance at $p < 0.001$ uncorrected, t-score 3.09. Lighter shades of yellow and brown outlines indicate level of significance at $p < 0.05$ FWE, t-score 4.6.

Figure S2: Peak activations that were used to define the ROIs of interest, on coronal slices, based on the AC vs. Fx contrast presented in (Figure 2 and Figure 4). Antero-posterior level relative to the intra-aural line indicated at the top right corner of each slice, in millimeters. Bold fonts refer to the location peak. Normal fonts refer to cortical activation extending beyond the peak. Activation thresholds were varied on some sections in order to clearly show the existence of local activation maxima. Activation color-scale was however kept constant across all slices. Down-left panel: ROIs locations on flatmaps (lateral sulcus; light blue; superior temporal sulcus: dark blue), same conventions as in Figure 5.

Figure S3: Percentage of signal change (%PSC) across all lateral sulcus (light blue) and superior temporal sulci (dark blue) ROIs of both hemispheres, comparing unimodal and multimodal congruent and incongruent conditions. Statistical differences relative to fixation are between conditions are indicates as follows: *, $p < 0.01$ (Wilcoxon non-parametric test).

Figure S4: Example of visual and auditory stimuli. At the right are shown examples of visual stimuli used for social and face blocked conditions. On the left, congruent calls spectrograms are associated to the visual stimuli are shown. The affiliative call is a coos, the aggressive congruent auditory stimulus is an aggressive call and the escape call is a scream. Stimuli were not strictly normalized in terms of in low visual and auditory feature properties, thus making their social meaning the dominant cue across the different stimuli of a given category.

Table S1: Whole brain gPPI functional connectivity analysis for the left and right lateral sulcus and the left and right superior temporal sulcus ROIs. Highlighted are the cortical regions showing at least one voxel with significant correlation with seed regions ($p < 0.005$ uncorrected).

Left LS	IPS	Insula	dIPFC (area 46)	STS	ACC	OFC	hippocampus	amygdala	dorsal pulvinar
F+	+	+	+	+	-	-	-	+	-
F-	+	+	+	+	+	+	-	+	-
S1-	+	-	+	+	+	+	+	+	-
S2-	+	-	+	+	+	+	-	-	-
S1+	+	+	+	+	+	+	+	+	+
S2+	-	+	+	+	+	+	+	-	+

Right LS	IPS	Insula	dIPFC (area 46)	STS	ACC	OFC	hippocampus	amygdala	dorsal pulvinar
F+	+	+	+	+	-	+	-	+	-
F-	+	-	+	+	-	+	-	+	-
S1-	+	-	+	+	+	+	+	+	+
S2-	+	-	+	+	+	-	+	+	+
S1+	-	-	+	+	+	+	+	-	-
S2+	+	+	+	+	+	+	+	+	+

Left STS	IPS	Insula	dIPFC (area 46)	LS	ACC	OFC	hippocampus	amygdala	dorsal pulvinar
F+	+	-	+	+	-	+	+	+	-
F-	-	+	+	+	+	+	+	-	+
S1-	+	-	-	+	+	+	+	-	+
S2-	+	+	-	+	+	+	-	+	+
S1+	+	+	+	+	+	+	-	-	-
S2+	+	-	+	+	-	-	+	+	-

Right STS	IPS	Insula	dIPFC (area 46)	LS	ACC	OFC	hippocampus	amygdala	dorsal pulvinar
F+	+	+	-	+	+	+	-	+	+
F-	+	+	+	-	+	+	+	+	-
S1-	+	-	+	+	+	-	-	+	-
S2-	+	+	-	-	+	+	-	-	-
S1+	+	-	-	+	+	+	+	+	-
S2+	+	+	+	+	+	+	+	+	+