CORTEX XXX (2010) 1–10

available at www.sciencedirect.com

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

Research report

Cross-modal interactions between human faces and voices involved in person recognition

Frédéric Joassin [a,*], *Mauro Pesenti* [a], *Pierre Maurage* [a], *Emilie Verreckt* [a], *Raymond Bruyer* [a] *and Salvatore Campanella* [b]

[a] *Université catholique de Louvain, Unité de Neurosciences Cognitives (NESC), Louvain-la-Neuve, Belgium*
[b] *Université Libre de Bruxelles Laboratory of Psychological Medecine, Brussels, Belgium*

ARTICLE INFO

Article history:
Received 29 April 2009
Reviewed 23 July 2009
Revised 6 October 2009
Accepted 24 February 2010
Action editor Stefan Schweinberger
Published online xxx

Keywords:
Cross-modal interactions
Faces
Voices
fMRI
Recognition

ABSTRACT

Faces and voices are key features of human recognition but the way the brain links them together is still unknown. In this study, we measured brain activity using functional magnetic resonance imaging (fMRI) while participants were recognizing previously learned static faces, voices and voice–static face associations. Using a subtraction method between bimodal and unimodal conditions, we observed that voice–face associations activated both unimodal visual and auditory areas, and specific multimodal regions located in the left angular gyrus and the right hippocampus. Moreover, a functional connectivity analysis confirmed the connectivity of the right hippocampus with the unimodal areas. These findings demonstrate that binding faces and voices rely on a cerebral network sustaining different aspects of integration such as sensory inputs processing, attention and memory.

1. Introduction

Human social interactions are shaped by our ability to identify individuals, a process to which face and voice recognition contributes both separately and jointly. Much research has been devoted to unimodal face recognition. Neuroimaging studies have shown that human faces are mainly processed by temporo-occipital regions of the brain with a right hemispheric dominance, and in particular in the fusiform gyrus (the so-called Fusiform Face Area – FFA, Sergent et al., 1992; Kanwisher et al., 1997; Rhodes et al., 2004). Studies with brain-damaged patients have revealed a selective impairment of face recognition, called prosopagnosia, associated with lesions of the right fusiform gyrus (De Renzi et al., 1994; Takahashi et al., 1995). Fewer studies have focused on voice recognition. Voice recognition takes place bilaterally in the superior temporal cortex, with a particular recruitment of the anterior part of the right superior temporal sulcus (STS, Belin et al., 2000, 2002; Von Kriegstein et al., 2003). Phonagnosia, the selective impairment of voice recognition, is predominantly associated with lesions of the right hemisphere (Neuner and Schweinberger, 2000), and Van Lancker et al. (1989) showed that an impairment in voice recognition was significantly correlated with right parietal lobe damages.

* *Corresponding author*. U.C.L. – IPSY, Place du Cardinal Mercier, 10, 1348 Louvain-la-Neuve, Belgium.
E-mail address: frederic.joassin@uclouvain.be (F. Joassin).
0010-9452/$ – see front matter
doi:10.1016/j.cortex.2010.03.003

Although neuroanatomically segregated, faces and voices interact, not only at a perceptual level (Calvert et al., 1999; Olson et al., 2002; Sekiyama et al., 2003), but also during person recognition (Burton et al., 1990; Ellis et al., 1997). Such integration skills emerge very early in life (Bahrick et al., 2005), but only few studies investigated cross-modal interactions between faces and voices in person identification. Schweinberger et al. (2007) showed that voice recognition was easier when simultaneously presented with an associated face, whereas it was hampered when presented with a face that did not share the same identity. This demonstrates that listeners cannot ignore a face as soon as it is presented in time synchrony with a voice. This effect was not observed with unfamiliar voices, which suggests that audio–visual integration in person recognition depends on multimodal representation of people, established through experience (for a review, see Campanella and Belin, 2007). However, the brain processes by which voices and faces, which are processed by distinct cerebral regions, are integrated into a unique and coherent representation of a person are still largely unknown.

Here we performed a functional magnetic resonance imaging (fMRI) study to investigate the cerebral correlates of voice–face interactions in a recognition task. We expected that voices alone would elicit a bilateral activation of the temporal cortex and in particular the anterior part of the right STS, and that faces alone would elicit an activation of the right FFA, i.e., the classical areas dedicated to the processing of voices and faces respectively. On the basis of neuroimaging studies showing an involvement of unimodal and multimodal areas in cross-modal binding (Wada et al., 2003; Bushara et al., 2003), we also predicted that bimodal stimulations would activate both unimodal visual and auditory areas (as previously observed for auditory speech perception, Calvert et al., 1999), and multimodal areas such as the anterior part of the temporal lobes, the hippocampus (Brown and Aggleton, 2001; Kirwan and Stark, 2004) and the parietal cortex (Saito et al., 2005; Bernstein et al., 2008). The anterior part of the temporal regions are known to be involved in the cross-modal processing of personal identity (Gorno-Tempini et al., 1998; Gainotti et al., 2003; Tsukiura et al., 2005; Calder and Young, 2005), and the hippocampus is known to be involved in the conjunction of features (Brown and Aggleton, 2001), in particular in the associative processes devoted to the recognition of faces (Kirwan and Stark, 2004). We also expected an activation of the left parietal cortex as (1) we have already observed its specific activation in a Positron Emission Tomography (PET) study investigating the associative processes between faces and written names (Campanella et al., 2001), and (2) the left parietal cortex is known to be a part of the heteromodal associative cortex (Niznikiewicz et al., 2000; Booth et al., 2002, 2003) involved in the binding of visual and auditory speech (Saito et al., 2005).

2. Methods

2.1. Participants

Fourteen healthy volunteers [7 females, mean age: 23.5, standard deviation (SD): 3.99] participated in the fMRI study. All were right handed, native French speakers, had normal vision and audition, and gave their written informed consent. The experimental protocol was approved by the Biomedical Ethical Committee of the Catholic University of Louvain.

2.2. Stimuli

The stimuli consisted of four associations. Each association formed a schematic person (or identity) and was composed of a female face (black and white photos, front view, neutral expression, picked from the Stirling Face Database: http://pics.psych.stir.ac.uk), a Belgian family name, and a female voice recorded in our laboratory and saying the French word «bonjour» (mean duration: 685.75 msec, SD: 142.34). Two learning sessions, each lasting about half an hour, familiarized participants with these four associations. The training sessions were carried out by a computer presentation using Superlab 6.1 software (Cedrus Corporation). Participants were shown each association (name, face and voice) and asked to learn them so that they would be able to retrieve the correct name when confronted to a face or a voice. This learning phase was performed with no time pressure and as many time as asked by each participant. Then, to ensure that the associations were correctly encoded, several recognition tests (face–name matching, voice–face matching, voice–name matching) and identification tests (recall of the name of each face and voice) were performed. Each error was immediately corrected and a new encoding was performed. Each learning session ended only when accuracy reached 100%.

2.3. Procedure

During the fMRI session, 3 different conditions were presented: faces (F), voices (V), and voice–face associations (VF, Fig. 1). Only two of the four identities were included in each

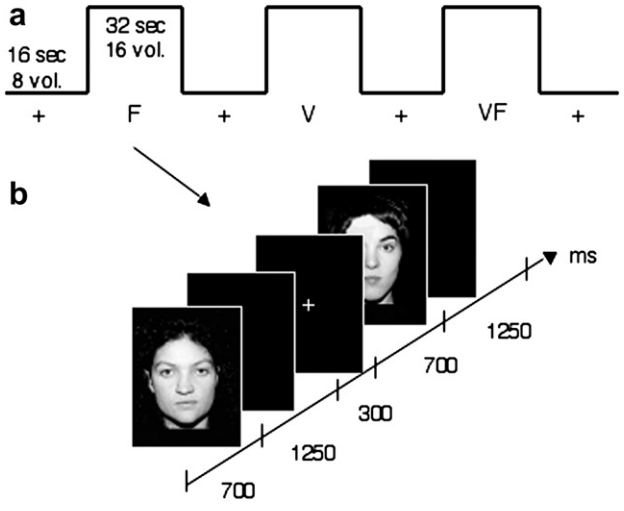

Fig. 1 – a) fMRI design: each run consisted in 3 alternances of a 16-sec fixation period (white cross on black background) and a 32-sec activation period. Each activation period corresponded to a different condition, presented in a pseudo-random order. b) Examples of behavioral task: participants were presented with 12 trials in each condition. Each trial comprised a fixation cross for 300 msec, a stimulus – faces (F), voices (V), or face/voice associations (VF) – for 700 msec and a black intertrial interval for 1500 msec.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

Table 1 – Associations used in each run.

Runs	Identity 1	Identity 2	Written instructions
1	Detiez	Goffin	Detiez (1) or Goffin (2)?
2	Detiez	Gillet	Detiez (1) or Gillet (2)?
3	Detiez	Masson	Detiez (1) or Masson (2)?
4	Goffin	Gillet	Goffin (1) or Gillet (2)?
5	Goffin	Masson	Goffin (1) or Masson (2)?
6	Gillet	Masson	Gillet (1) or Masson (2)?

Identities 1 and 2 = the 2 schematic persons (name, face and voice) used in each run. The written instructions are the names appearing on the screen at the beginning of each run and informing the participants of the 2 identities used in each run (response buttons between parentheses).

run (for instance the identities "Detiez" and "Goffin" in the first run, "Detiez" and "Gillet" in the second run and so on, see Table 1) and these were varied across runs. Participants were informed of the two identities used in each run by a written instruction ("Detiez or Goffin?") appearing on the screen before the beginning of each run. The task consisted of categorizing each trial (face, voice or association) according to its identity (i.e., its name) by pressing one of two keys on a stimpad with two fingers of the right hand (left button for the first identity, right button for the second identity).[1]

Each volunteer participated in 6 runs each consisting of six experimental blocks of 32 sec (2 blocks per condition), interleaved with 16-sec fixation periods (white cross on black background, Fig. 1). The order of the various conditions within the run was pseudo-randomly balanced across runs and subjects. Each experimental block comprised 12 trials. Each trial was composed of a fixation cross (300 msec), followed by the stimulus for 700 msec and an empty interval of 1500 msec. The importance of both speed and accuracy was emphasized.

2.4. *Apparatus and experimental set-up*

Stimulus presentation and response recording were controlled with ePrime (Schneider et al., 2002). Back-projected images were viewed through a tilted mirror (Silent Vision™ System, Avotec, Inc., <http://www.avotec.org>) mounted on the head coil. Auditory stimuli were delivered through headphones and the sound volume was adjusted for each participant so as to be clearly audible above the scanner noise.

2.5. *Imaging procedure*

Functional images were acquired with a 1.5 Tesla magnetic resonance imager and a standard head coil (Gyroscan, Philips Medical Systems) as series of blood-oxygen-sensitive T2*-weighted echo-planar image volumes (GRE-EPI). Acquisition parameters were: TE = 50 msec, TR = 2000 msec, Flip angle = 90°, Field of view = 210 × 210 mm, slice thickness = 6 mm with no interslice gap. Each image volume comprised 20 axial slices acquired in an ascending interleaved sequence. Each functional run comprised 160 volumes, the first 8 volumes being discarded to allow for T1 equilibration effects, which resulted in 16 volumes per condition per run. High-resolution T1-weighted 3D fast field echo anatomical images with 110 1.5-mm contiguous axial slices were also acquired for each participant (TE = 3 msec, TR = 30 msec, Flip angle = 30°, FOV = 220 × 175 mm; in-plane voxel size .859 × .859 × 1.5 mm^3). Head movement was limited by a restraining band across the forehead.

2.6. *Data processing and analysis*

Latencies and percentages of correct responses were analyzed separately, each using an analysis of variance (ANOVA) with the condition (V, F, VF) as within-subject factor, followed by paired *t*-tests comparing each condition to the other two when appropriate.

Neuroimaging data were processed (slice-time correction, realignment, coregistration, normalization to the MNI template , using an affine fourth degree ß-spline interpolation transformation and a voxel size of 2 × 2 × 2 mm^3, smoothing with a Gaussian kernel of 8 mm FWHM) and analyzed using SPM2 (Statistical Parametric Mapping, Welcome Department of Cognitive Neurology, London, UK, <http://www.fil.ion.ac.uk/spm>), implemented in a Matlab 6.5.0 environment (The Mathworks, Inc.). Condition-related changes in regional brain activity were estimated for each participant by a general linear model in which the responses evoked by each condition of interest were modeled by a standard hemodynamic response function. The contrasts of interest were computed at the individual level to identify the cerebral regions significantly activated by voices ([V – fix]), faces ([F – fix]) and bimodal stimuli ([VF – fix]) relative to the fixation periods used as a general baseline. The contrast [VF – (V + F)] was computed to isolate the cerebral regions involved in the associative processes between faces and voices.

Significant cerebral activations were then examined at the group level in random-effect analyses using one-sample *t* tests, with statistical threshold set to $p < .05$ corrected for multiple comparisons using cluster size and extending to at least 10 contiguous voxels. For the cerebral regions for which we had an a-priori hypothesis, the statistical threshold was set at $p < .001$ uncorrected.

We explored the connectivity of the regions activated in the contrast [VF – (V + F)] by computing 4 psychophysiological interaction analyses (PPI, Friston et al., 1997; Friston, 2004). Each PPI analysis employed 3 regressors: one regressor representing the deconvolved activation time course in a given volume of interest (the physiological variable), one regressor representing the psychological variable of interest, and a third regressor representing their cross-product (the psychophysiological interaction term). Each of the four analyses focused on one particular region observed in the group analysis, i.e., the left angular gyrus, the right hippocampus, the right FFA and the

[1] Three other conditions were also included in each run, separately and independently from the 3 other ones. They consisted in a voice condition identical to the auditory condition described in the procedure, and a visual and auditory–visual conditions in which the faces were modified by morphing to become more difficult to recognize. The aim of this modification was to examine the face–voice interactions when both stimuli were equally difficult to recognize. However, as these conditions gave rise to hardly explainable behavioral results, it was decided to exclude them from the analyses.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

right STS. For each participant, we performed a small volume correction (sphere of 5 mm centered on the maximum peak of activity of the region in the group analysis) before extracting the deconvolved time course of activity in a ROI (5-mm radius sphere centered at the voxels displaying maximum peak activity in the group analysis). The time course of activity was corrected for the effect of interest. We then calculated the product of this activation time course with a condition-specific regressor probing the integration of faces and voices [VF − (V + F)] to create the psychophysiological interaction terms. PPI analyses were carried out for each ROI in each subject, and then entered into a random-effects group analysis (uncorrected threshold at $p < .001$, as in Ethofer et al., 2006).

3. Results

3.1. Behavioral data

We observed significant latencies differences between V, F and VF [$F(2,26) = 31.27$, $p < .001$]. Subsequent paired t-tests revealed that (1) voices were identified more slowly than faces [$t(13) = 6.47$, $p < .001$] or voice–face associations [$t(13) = 5.55$, $p < .001$], although these two latter conditions did not significantly differ [$t(13) = 1.36$, ns, Table 2 and Fig. 2].

The percentages of correct responses showed the same patterns of results. Significant differences between V, F and VF [$F(2,26) = 22.93$, $p < .001$] were due to the fact that voices were less correctly identified than faces [$t(13) = -5.14$, $p < .001$] and associations [$t(13) = -6.55$, $p < .001$], these two conditions not differing [$t(13) = 1.66$, ns, Table 2].

3.2. Brain imaging results

3.2.1. Processing of faces, voices and associations

Compared to fixation, voices elicited an activation of the right middle and left superior temporal gyri (Table 3a); faces activated the left calcarine sulcus, the left and right fusiform gyri and the left supramarginal gyrus (Table 3b); and associations between voices and faces activated mainly the auditory temporal regions bilaterally, the right fusiform gyrus and the left supramarginal gyrus (Table 3c).

3.2.2. Cerebral correlates of face and voice integration

To isolate the cerebral regions specifically involved in the associative processes linking faces and voices, the unimodal auditory and visual conditions were subtracted from the bimodal auditory–visual ones. This revealed the activation of visual and auditory regions, including respectively the left calcarine sulcus and the fusiform gyri bilaterally, and the middle temporal gyri bilaterally (Table 4 and Fig. 3).

We also observed the specific activation of 2 bimodal convergence regions which were not activated in the unimodal conditions: the right hippocampus (Fig. 4) and the left angular gyrus (Fig. 5).

3.2.3. Functional connectivity analyses

We assessed the hypothesis that the left angular gyrus and the right hippocampus are involved in the integration of faces and voices, by computing several PPI analyses (Friston et al., 1997; Friston, 2004).

These analyses, performed on the contrast [VF − (V + F)] revealed that the left angular gyrus had an enhanced connectivity with the left middle frontal and post-central gyri and supplementary motor area, and with the cerebellum and vermis (Table 5a). The right hippocampus showed an enhanced connectivity with the left inferior occipital gyrus, the right fusiform gyrus, the left and right middle temporal gyri, the right putamen and the left thalamus (Table 5b).

In order to investigate the connectivity of the unimodal regions (right FFA and right STS) in the bimodal situation, 2 supplementary PPI analyses were performed with these 2 regions used as ROIs. They revealed that the right FFA had an enhanced connectivity with the right and left superior temporal gyri, the right Heschl gyrus and the right hippocampus (Table 5c). The right STS showed an enhanced connectivity with the right middle and superior occipital gyri, the left and right fusiform gyri and the right hippocampus and parahippocampal gyrus (Table 5d).

Table 2 – Mean latencies (in msec) and percentages of correct responses (SD in parentheses).

	V	F	VF
Latencies	592 (86.27)	502 (57.78)	513 (53.42)
%	79 (8.80)	87.3 (5.97)	90.5 (7.37)

V = voices, F = faces, VF = associations between voices and faces.

4. Discussion

The aim of the present study was to investigate the cerebral correlates of voice–face interactions involved in person recognition. By using a subtraction method between bimodal and unimodal conditions, we isolated the cerebral regions sustaining face–voice integration. We observed the activation of a cortical network including unimodal visual and auditory regions along with multimodal regions such as the hippocampus and the left angular gyrus.

4.1. Behavioral data

The behavioral results confirmed our previous findings (Joassin et al., 2004, 2008) that voices are harder to recognize than faces. Such differences have already been observed (Schweinberger et al., 1997; Ellis et al., 1997). Hanley and Turner (2000) postulated that voices are associated with lower overall levels of familiarity than faces. A simulation with the Interactive Activation Model (IAC) model of Burton et al. (1990) confirmed this interpretation in showing that their results could be explained by weaker connections between Voice Recognition Units (VRU) and Person Identity Nodes (PIN) than between Face Recognition Units (FRU) and PIN. On the whole, this body of data leads us to think that human adults are more expert in face recognition than in voice recognition. Nevertheless, functional data showed that faces and voices were not processed independently but

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

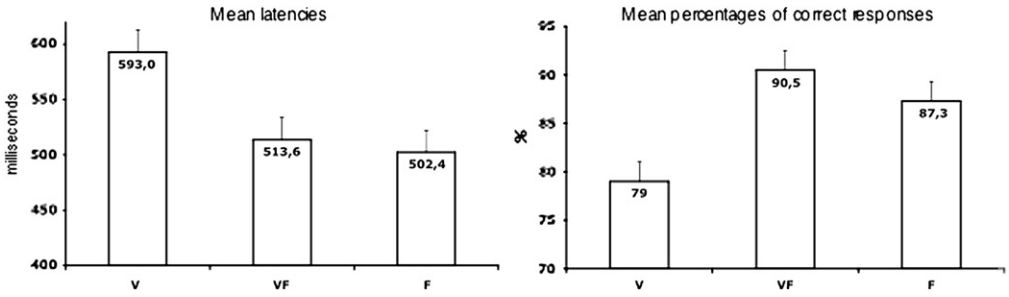

Fig. 2 – Mean latencies (left side) and percentages of correct responses (right side) for the categorization of voices (V), face/voice associations (VF) and faces (F).

interacted when presented together, which is in accordance with the hypothesis that faces and voices could be bound together automatically (Amedi et al., 2005).

It is important here to note that we used static faces in the present experiment. However, Schweinberger and his collaborators have recently shown in two studies that dynamic visual information plays an important role in person recognition. In a first experiment, they showed that (1) the recognition of familiar voices was easier when the voices were combined with corresponding synchronously articulating faces, compared to static faces, and (2) that combining a voice with a non-corresponding face (i.e., of a different identity) hampered voice recognition, but only when the face was dynamic (Schweinberger et al., 2007). Moreover, in a more recent study, Robertson and Schweinberger (2010) showed that there is a precise temporal window for the audio–visual face–voice integration in the recognition of speaker identity. Indeed, voice recognition was significantly easier when the corresponding articulating face was presented in approximate synchrony with the voice, the largest benefit being observed when the voice was presented with a delay of 100 msec after the onset of the face.

We do not think that using static faces weakened our results. Nevertheless the fact that dynamic visual information acts on voice recognition makes obvious the use of moving faces in our future researches.

Table 3 – Brain regions showing significant activation compared to baseline (fix) for voices (a), faces (b), and associations between voices and faces (c).

Brain regions	x	y	z	L/R	k	t-statistic
a) V – fix						
MTG	62	−26	0	R	11875	5.91
Superior temporal gyrus	−52	−32	6	L	12684	5.67
Middle frontal gyrus	36	42	26	R	329	5.64
Middle frontal gyrus	−34	36	20	L	123	3.81
Calcarine sulcus	−14	−98	4	L	195	4.44
b) F – fix						
Calcarine sulcus	−16	−92	−4	L	560	5.30
Fusiform gyrus	−32	−56	−18	L	127	4.25
Fusiform gyrus	38	−58	−12	R	2935	4.89
Inferior frontal oper.	46	10	22	R	703	5.06
Supplementary motor area	6	6	60	R	918	5.03
Pre-central gyrus	46	0	52	R	223	4.33
Pre-central gyrus	−50	−2	46	L	97	4.60
Supramarginal gyrus	−26	−8	48	L	255	4.29
Pallidum	−22	−8	6	L	241	3.75
c) VF – fix						
MTG	64	−26	2	R	21225	5.87
Superior temporal gyrus	−48	44	−8	L	91	4.14
Fusiform gyrus	38	−48	−12	R	342	4.76
Frontal inferior orb.	50	38	−12	R	83	3.88
Frontal inferior tri.	52	20	18	L	83	3.75
Pre-central gyrus	48	4	54	R	443	4.81
Supplementary motor area	0	−2	60	R	1110	5.11
Supramarginal gyrus	−42	−36	36	L	2377	4.88
Caudate	−10	20	−4	L	355	4.49

x, y, z are stereotactic coordinates of peak-height voxels. L = left hemisphere, R = right hemisphere. k = clusters size. Threshold set at $p < .05$ corrected for multiple comparisons using cluster size.

4.2. *Unimodal face and voice areas*

The functional data showed that voice–face associations rely at least in part on the activation of unimodal visual and auditory areas (Von Kriegstein et al., 2005), such as the FFA (Sergent et al., 1992; Kanwisher et al., 1997; Rhodes et al., 2004) and the left and right middle temporal regions, known to be involved in voice identification processes (Giraud et al., 2004; Belin et al., 2004; Beaucousin et al., 2007). Unimodal

Table 4 – Brain regions showing significant activation in the subtraction between unimodal conditions (V and F) and the bimodal one (VF).

Brain regions	x	y	z	L/R	k	t-statistic
VF – (V + F)						
Calcarine sulcus	−16	−96	−6	L	418	5.34
Fusiform gyrus	38	−50	−18	R	2348	4.82
Fusiform gyrus	−34	−54	−18	L	134	3.97
Superior temporal gyrus	62	−26	2	R	901	4.93
Superior temporal gyrus	−62	−34	2	L	890	4.62
Hippocampus	16	−32	−6	R	204	4.33
Angular gyrus	−46	−64	40	L	69	3.44*

x, y, z are stereotactic coordinates of peak-height voxels. L = left hemisphere, R = right hemisphere. k = clusters size. Threshold set at $p < .05$ corrected for multiple comparisons using cluster size.
* p-values < .001 uncorrected.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

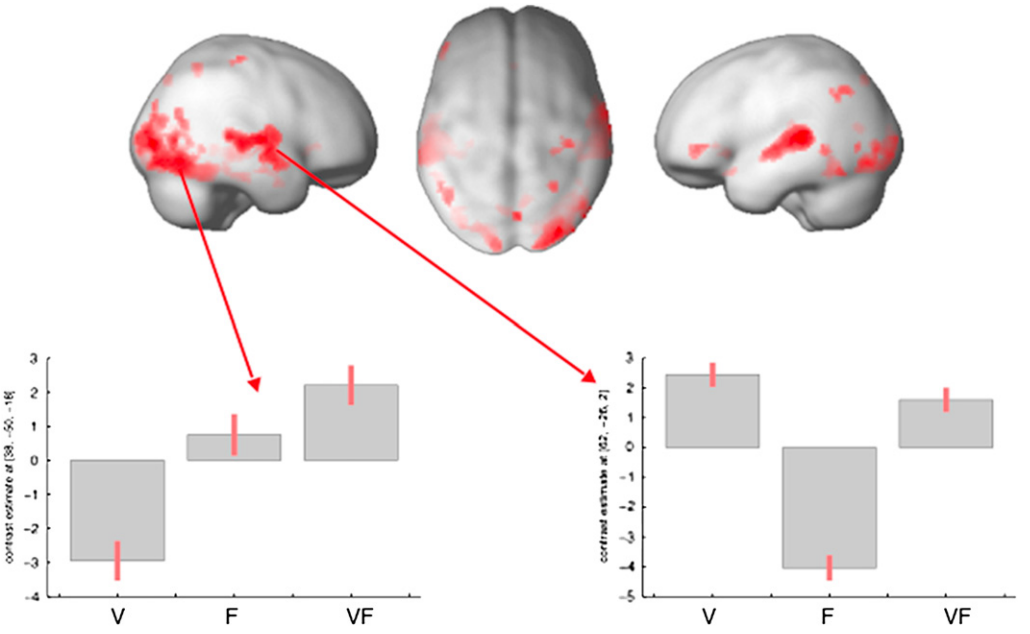

Fig. 3 – Brain regions activated in the contrast [VF – (V + F)]. a) Statistical parametric maps superimposed on MRI surface renders (left, top and right views); b) activation changes for each condition in the right middle fusiform gyrus; c) activation changes for each condition in the right MTG. $p < .05$ corrected for multiple comparisons at cluster size. V = voices, F = faces, VF = face/voice associations.

activations during bimodal stimulations have been described in non-human primates (Ghazanfar et al., 2005) and humans (Calvert et al., 1999; Von Kriegstein et al., 2005), highlighting the importance of sensory brain regions within the auditory–visual integrative processes. Moreover, Von Kriegstein and Giraud (2006) have shown that auditory and visual regions can be jointly activated even when a voice from a previously learned face–voice association was presented in isolation. Our PPI analyses confirm these results in showing that the right FFA and STS were interconnected during the recognition of the face–voice associations. These results confirm the proposition of Von Kriegstein and Giraud (2006) that an exposure to ecological redundant signals, such as faces and voices, induce specific multisensory associations and an optimized connectivity between the visual and auditory sensory regions.

The activation of the posterior part of the middle temporal gyrus (MTG) was observed during both auditory and bimodal conditions. The temporal regions, and in particular the STS is known to show multisensory integration for audio–visual

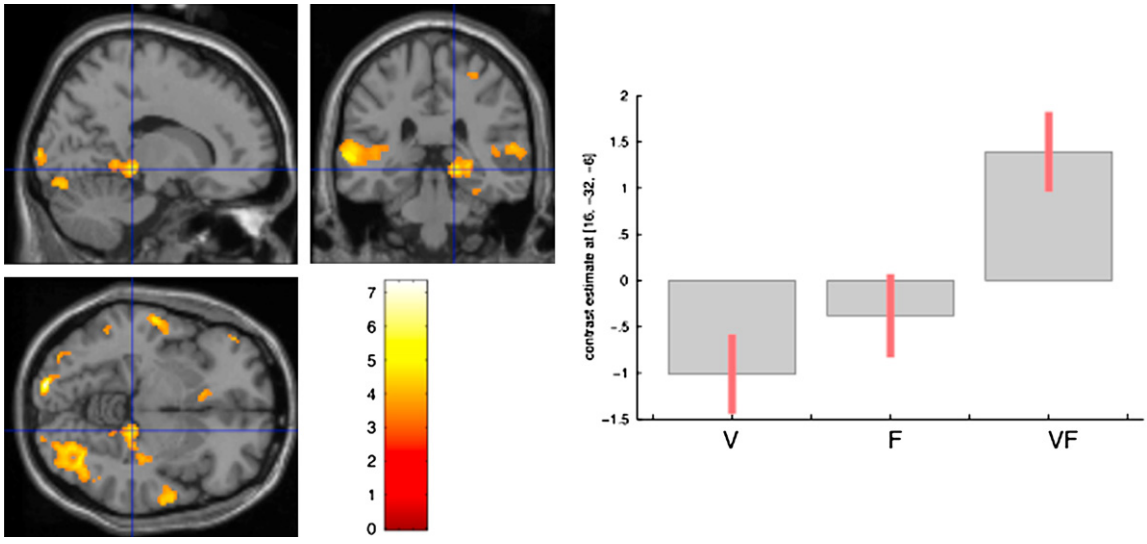

Fig. 4 – Brain sections of the contrast [VF – (V + F)] centered on the right hippocampus (left side). Activation changes for each condition in the right hippocampus (right side). $p < .05$ corrected for multiple comparisons at cluster size. V = voices, F = faces, VF = face/voice associations.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

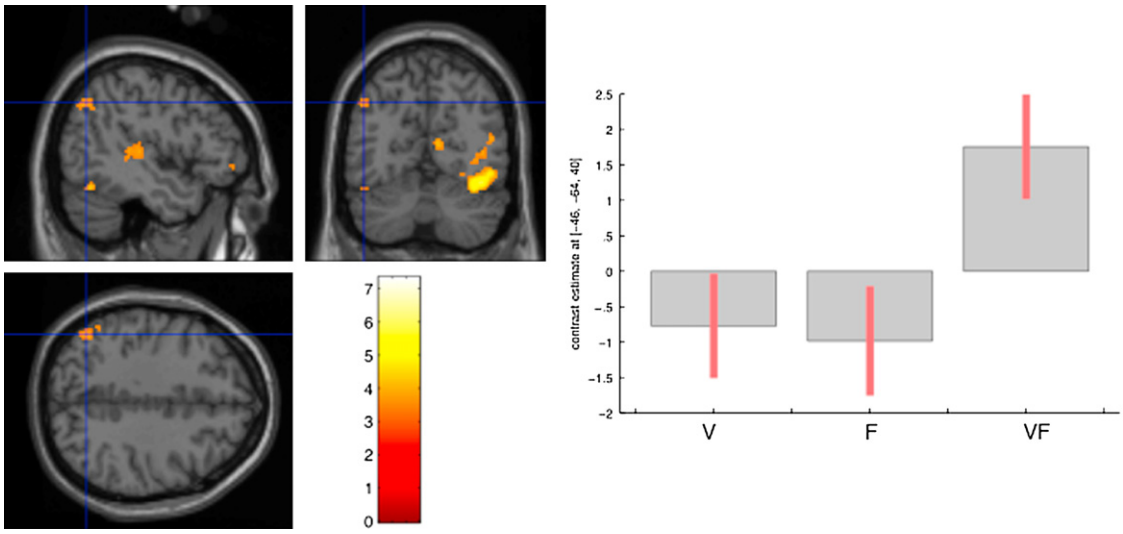

Fig. 5 – Brain sections of the contrast [VF − (V + F)] centered on the left angular gyrus (left side). Activation changes for each condition in the left angular gyrus (right side). $p < .05$ corrected for multiple comparisons at cluster size. V = voices, F = faces, VF = face/voice associations.

stimulations (Calvert, 2001; Beauchamp, 2005; Stevenson et al., 2007; Stevenson and James, 2009). However, the fact that these regions were activated in both bimodal and unimodal auditory conditions can be explained in two ways. At first, we have to remember that the STS is a large region containing contiguous populations of neurons responding to unimodal or multimodal stimulations (Beauchamp et al., 2004; Hein and Knight, 2008). It is possible to find within a single voxel neurons responding specifically to auditory and to face–voice stimuli. The activation of the MTG in the bimodal condition could thus reflect genuine multisensory interactions between faces and voices that overlapped the cerebral activity evoked by the auditory processing in a voxel-based analysis. In the other way, we cannot exclude the hypothesis that the activation of the MTG in the face–voice condition reflect the unimodal processing of the voices during the bimodal trials. Unfortunately, the present data do not allow us to favour one or the other explanation. Further experiments, notably varying the signal-to-noise ratio (SNR, Stevenson and James, 2009), will help to disentangle these two hypotheses.

We did not observe any significant activation of the anterior parts of the temporal lobes, known to be involved in the cross-modal processing of personal identity (Gorno-Tempini et al., 1998; Gainotti et al., 2003; Tsukiura et al., 2005; Calder and Young, 2005). It could be possible that this lack of activation is due to the associations used in the present experiment *per se*. Actually, our associations were only composed of a face, a voice and a name, without any biographical or semantic specific information.

However, it has been showed that the more anterior part of the temporal lobe, especially in the right hemisphere, sustains the amodal retrieval of person specific semantic information (Gainotti et al., 2003). It is thus possible that the absence of activation in these regions in the present experiment is due to the schematic aspect of our associations for which there were no specific semantic information to access. This interpretation is reinforced by the study of Tsukiura et al. (2005) who showed that the anterior temporal lobes were activated by the retrieval of people's names from faces and conversely, but only when a specific semantic knowledge (the occupation) was attached to the face–name associations during encoding.

Von Kriegstein and Giraud (2006), in an fMRI experiment exploring how implicit voice–name or voice–face associations influenced voice recognition, observed an activation of the right anterior temporal cortex during the learning of the associations. Their experiment differed from the present one in its goal, methods and analyses. Nevertheless, the fact that Von Kriegstein and Giraud (2006) found an activation of the anterior temporal lobe during the encoding of associations containing no more semantics than ours open new prospects for future behavioral and neuroimaging studies directly comparing, with our paradigm, the learning versus retrieval of face–voice associations, the presence versus absence of specific biographical information, and the implicit versus explicit nature of the associations.

4.3. *The left angular gyrus*

The contrast between bimodal and unimodal conditions revealed supplementary regions than those involved in the processing of unimodal sensory inputs and whose activation was specific to the bimodal condition. Indeed, we used a super-additive criterion to detect these regions, requiring multisensory responses larger than the sum of the unisensory responses (Calvert et al., 2001; Beauchamp, 2005). This criterion has often be considered as overly strict in that sense that it can introduce type II errors (false negative), due to the fact that, in a single voxel, the activity of super- and sub-additive neurons is measured (Laurienti et al., 2005). Nevertheless, it showed a super-additive BOLD activation of the left angular gyrus. This region is known to form part of the associative

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

Table 5 – Brain regions showing an enhanced connectivity with the left angular gyrus, the right hippocampus, the right FFA and the right STS in the contrasts [VF – (V + F)].

Brain regions	x	y	z	L/R	t-statistic
a) Left angular gyrus (−46, −64, 40)					
Supplementary motor area	−6	0	60	L	4.67
Middle frontal gyrus	−36	−4	54	L	4.46
Post-central gyrus	−44	−28	48	L	3.73
Vermis	0	−72	−18	L/R	3.83
Cerebellum	4	−42	−16	R	4.23
Putamen	−30	−6	12	L	3.66
b) Right hippocampus (16, −32, −6)					
Inferior occipital gyrus	−38	−84	−12	L	3.23
	−32	−86	−8	L	3.20
Fusiform gyrus	40	−56	−22	R	3.28*
MTG	−40	−48	2	L	3.10
MTG	36	−46	14	R	3.07*
	38	−48	20	R	3.07*
Putamen	22	18	0	R	3.49
Thalamus	−4	−8	−4	L	3.42
c) Right fusiform gyrus (38, −50, −18)					
Superior temporal gyrus	58	−14	6	R	4.50
	60	6	−8	R	4.11
Superior temporal gyrus	−42	−30	2	L	4.28
	−64	−28	2	L	4.25
Heschl gyrus	32	−22	6	R	4.10
Hippocampus	30	−32	−4	R	4.30
d) Right superior temporal gyrus (62, −26, 2)					
Superior occipital gyrus	−22	−98	10	L	5.58
Middle occipital gyrus	22	−92	0	R	5.38
MTG	16	−102	2	R	4.10
Lingual gyrus	−22	−90	−14	L	4.65
Calcarine sulcus	−24	−90	0	L	4.28
Fusiform gyrus	32	−56	−12	R	4.07
Parahippocampal gyrus	30	−8	−28	R	4.21
Hippocampus	16	−34	−4	R	2.56**

x, y, z are stereotactic coordinates of peak-height voxels. L = left hemisphere, R = right hemisphere. Threshold set at $p < .001$ uncorrected; * $p < .005$ uncorrected; ** $p < .05$ uncorrected.

cortex and receives multiple inputs from modality specific sensory regions and provides a unique representation of the combined sensory features (Damasio, 1989; Clark et al., 2000; Rämä and Courtney, 2005; Niznikiewicz et al., 2000; Booth et al., 2002, 2003). More specifically, Bernstein et al. (2008) have recently put in light with event-related potentials (ERPs) a sustained activity of the left supramarginal/angular gyrus (from 160 to 220 msec after stimuli onset) during the perception of congruent audio–visual speech. The authors interpreted this activity in the left parietal cortex as reflecting the integration of visual and auditory speech stimulus information.

The present results underline the importance of the left parietal cortex, and in particular of the left angular gyrus, in cross-modal binding of visual and auditory information. It is worth noting that the activation of the left angular gyrus observed in this study is very similar to the activation of the left inferior parietal lobule we observed in a previous PET study (Campanella et al., 2001) examining the cerebral activations elicited by the retrieval of face–name associations presented in the visual modality. The left parietal cortex thus seems to be closely implicated in the associative processes binding the various pieces of information related to individuals, whatever this information is visual, verbal or auditory.

This left parietal activation may be related to some processes of divided attention. Indeed, Saito et al. (2005) found an activation of the left parietal regions during the perception of audio–visual speech in a paradigm of divided attention. It has also been implicated in tasks requiring cross-modal spatial attention (Bushara et al., 2003). Moreover, our PPI analysis on the left angular gyrus revealed that it had an enhanced connectivity with the cerebellum and motor and pre-motor cortical regions including the supplementary motor area, the pre-central gyrus and the middle and superior frontal gyri. This parieto-premotor cortical network is important for the control of attention (Driver and Spence, 2000) and has been reported in several studies using visual (LaBar et al., 1999), auditory (Binder et al., 1997) and cross-modal (O'Leary et al., 1997; Bushara et al., 1999; Shomstein and Yantis, 2004) stimuli. It is thus possible that the parieto-premotor network observed in the present study act to direct attention simultaneously to targets from distinct sensory modalities (Lewis et al., 2000).

4.4. *The hippocampus*

We also observed a super-additive BOLD activation of the right hippocampus during the bimodal condition. This region being deactivated in both unimodal conditions, its activation during the recognition of the face–voice activations cannot be explained by general memory retrieval processes. Rather, it seems involved in the associative recognition of the faces and voices *per se*. Indeed, this particular region of the medial temporal lobe is well known to be involved in declarative memory (Milner et al., 1998), and in particular in the conjunction of features (Eichenbaum, 2000; Brown and Aggleton, 2001). For instance, Stark and Squire (2001) have shown that the hippocampal region was activated in an associative recognition memory task of words and objects. More recently, Kirwan and Stark (2004) observed that the activity of the hippocampus was greater for the retrieval of face–name pairs than for the retrieval of non-associative information (faces or names alone). These data suggest that the hippocampal region is a cross-modal structure involved in the encoding and retrieval of associated information in memory. Moreover, it has recently been found that visual and auditory cortical regions of monkeys project to the hippocampus through the entorhinal cortex, suggesting that unimodal cortical inputs converge in the hippocampus, providing the integration of complex stimuli for internal representations in memory (Mohedano-Moriano et al., 2008). The results of our PPI analysis are in line with this interpretation as (1) we observed that the hippocampus had an enhanced connectivity with both unimodal visual (the fusiform gyrus) and auditory (the superior temporal gyrus) regions, and (2) the right FFA and STS had also an enhanced connectivity with the right hippocampal and parahippocampal regions. We thus propose that the hippocampus is a key region where the representations of faces and voices are integrated into a multimodal representation to be compared with the face–voice representations stored in memory.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

Taken together, the present results demonstrate that cross-modal person recognition relies on the activation of a cerebral network including unimodal face and voices areas along with multimodal regions such as the left angular gyrus, involved in cross-modal attentional processing, and the hippocampus, sustaining the forming and retrieval of auditory–visual representations of people in memory. They also support a dynamic vision of cross-modal interactions in which heteromodal areas are not simply the final stage of a hierarchical unimodal-to-multimodal processing model (Bushara et al., 2003), but rather, they may work in parallel and influence each other.

Acknowledgments

This work was supported by grant No. 1.5.130.05F from the National Fund for Scientific Research (Belgium), and grant No. 01/06-267 from the Communauté Française de Belgique – Actions de Recherche Concertées (Belgium).

Frédéric Joassin and Pierre Maurage are Postdoctoral Researchers, and Salvatore Campanella and Mauro Pesenti Research Associates at the National Fund for Scientific Research (Belgium). We thank the Radiodiagnosis Unit at the Cliniques St. Luc (Brussels) for its support, and Ms. Sue Hamilton for her helpful suggestions during the redaction of this article.

REFERENCES

Amedi A, Von Kriegstein K, van Atteveldt NM, Beauchamp MS, and Naumer MJ. Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, 166: 559–571, 2005.

Bahrick LE, Hernandez-Reif M, and Flom R. The development of infant learning about specific face–voice relations. *Developmental Psychology*, 41(3): 541–552, 2005.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, and Martin A. Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience*, 7(11): 1190–1192, 2004.

Beauchamp MS. Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics*, 3: 93–113, 2005.

Beaucousin V, Lacheret A, Turbelin MR, Morel M, Mazoyer B, and Tzourio-Mazoyer N. FMRI study of emotional speech comprehension. *Cerebral Cortex*, 17: 339–352, 2007.

Belin P, Zatorre RJ, Lafaille P, Ahad P, and Pike B. Voice-selective areas in human auditory cortex. *Nature*, 403: 309–312, 2000.

Belin P, Zatorre RJ, and Ahad P. Human temporal-lobe response to vocal sounds. *Brain Research Cognitive Brain Research*, 13: 17–26, 2002.

Belin P, Fecteau S, and Bédart C. Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3): 129–135, 2004.

Bernstein LE, Auer Jr ET, Wagner M, and Ponton CW. Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage*, 39: 423–435, 2008.

Binder JR, Frost JA, Hammecke TA, Cox RW, Rao SM, and Prieto T. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17: 353–362, 1997.

Booth JR, Burman DD, Meyer JR, Gitelman DR, Parrish TB, and Mesulam MM. Functional anatomy of intra- and cross-modal lexical tasks. *NeuroImage*, 16: 7–22, 2002.

Booth JR, Burman DD, Meyer JR, Gitelman DR, Parrish TB, and Mesulam MM. Relation between brain activation and lexical performance. *Human Brain Mapping*, 19: 155–169, 2003.

Brown MW and Aggleton JP. Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Review Neuroscience*, 2: 51–61, 2001.

Burton AM, Bruce V, and Johnston RA. Understanding face recognition with an interactive model. *British Journal of Psychology*, 81: 361–380, 1990.

Bushara KO, Weeks RA, Ishii K, Catalan MJ, Tian B, Rauschecker JP, et al. Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. *Nature Neuroscience*, 2: 759–766, 1999.

Bushara KO, Hanakawa T, Immish I, Toma K, Kansaku K, and Hallett M. Neural correlates of cross-modal binding. *Nature Neuroscience*, 6(2): 190–195, 2003.

Calder AJ and Young AW. Understanding the recognition of facial identity and facial expression. *Nature Review Neuroscience*, 6: 641–651, 2005.

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, and David AS. Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport*, 10: 2619–2623, 1999.

Calvert GA. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, 11: 1110–1123, 2001.

Calvert GA, Hansen PC, Iversen SD, and Brammer MJ. Detection of audio–visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage*, 14: 427–438, 2001.

Campanella S, Joassin F, Rossion B, De Volder AG, Bruyer R, and Crommelinck M. Associations of the distinct visual representations of faces and names: A PET activation study. *NeuroImage*, 14: 873–882, 2001.

Campanella S and Belin P. Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12): 535–543, 2007.

Clark CR, Egan GF, McFarlane AC, Morris P, Weber D, Sonkkilla C, et al. Updating working memory for words: A PET activation study. *Human Brain Mapping*, 9(1): 42–54, 2000.

Damasio AR. Time-locked multiregional retroactivation: A system-level proposal for neural substrates of recall and recognition. *Cognition*, 33: 25–62, 1989.

De Renzi E, Peroni D, Carlesimo GA, Silveri MC, and Fazio F. Prosopagnosia can be dissociated with damage confined to the right hemisphere – An MRI and PET study and a review of the literature. *Neuropsychologia*, 32(8): 893–902, 1994.

Driver J and Spence C. Multisensory perception: Beyond modularity and convergence. *Current Biology*, 10: 731–735, 2000.

Eichenbaum H. A cortical–hippocampal system for declarative memory. *Nature Review Neuroscience*, 1: 41–50, 2000.

Ellis HD, Jones DM, and Mosdell N. Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, 88: 143–156, 1997.

Ethofer T, Anders S, Erb M, Droll C, Royen L, Saur R, et al. Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping*, 27: 707–714, 2006.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, and Dolan RJ. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6: 218–229, 1997.

Friston KJ. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2: 56–78, 2004.

Gainotti G, Barbier A, and Marra C. Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. *Brain*, 126(4): 792–803, 2003.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003

Ghazanfar AA, Maier JX, Hoffman KL, and Logothetis NK. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, 25(20): 5004–5012, 2005.

Giraud AL, Kell C, Thierfelder C, Sterzer P, Russ MO, Preibisch C, et al. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex*, 14(3): 247–255, 2004.

Gorno-Tempini ML, Price CJ, Josephs O, Vandenberghe R, Cappa SF, Kapur N, et al. The neural systems sustaining face and proper-name processing. *Brain*, 121: 2103–2118, 1998.

Hanley JR and Turner JM. Why are familiar-only experiences more frequent for voices than for faces? *Quarterly Journal of Experimental Psychology*, 53A(4): 1105–1116, 2000.

Hein G and Knight RT. Superior temporal sulcus – It's my area: Or is it? *Journal of Cognitive Neuroscience*, 20(12): 2125–2136, 2008.

Joassin F, Maurage P, Bruyer R, Crommelinck M, and Campanella S. When audition alters vision: An event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience Letters*, 369: 132–137, 2004.

Joassin F, Maurage P, and Campanella S. Perceptual complexity of faces and voices modulates cross-modal behavioral facilitation effects. *Neuropsychological Trends*, 3: 29–44, 2008.

Kanwisher N, McDermott J, and Chun MM. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 9: 462–475, 1997.

Kirwan CB and Stark CEL. Medial temporal lobe activation during encoding and retrieval of novel face–name pairs. *Hippocampus*, 14: 919–930, 2004.

LaBar KS, Gitelman DR, Parrish TB, and Mesulam MM. Neuroanatomic overlap of working memory and spatial attention networks: A functional MRI comparison within subjects. *NeuroImage*, 10: 695–704, 1999.

Laurienti PJ, Perrault TJ, Stanford TR, Wallace MT, and Stein BE. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166: 289–297, 2005.

Lewis JW, Beauchamp MS, and Deyoe EA. A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex*, 10: 873–888, 2000.

Milner B, Squire LR, and Kandel ER. Cognitive neuroscience and the study of memory. *Neuron*, 2: 445–468, 1998.

Mohedano-Moriano A, Martinez-Marcos A, Pro-Sistiaga P, Blaizot X, Arroyo-Jimenez M, Marcos P, et al. Convergence of unimodal and polymodal sensory input to the entorhinal cortex in the fascicularis monkey. *Neuroscience*, 151: 255–271, 2008.

Neuner F and Schweinberger SR. Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition*, 44(3): 342–366, 2000.

Niznikiewicz M, Donnino R, McCarley RW, Nestor PG, Losifescu DV, O'Donnell B, et al. Abnormal angular gyrus asymmetry in schizophrenia. *American Journal of Psychiatry*, 157(3): 428–437, 2000.

O'Leary DS, Andreasen NC, Hurtig RR, Torres IJ, Flashman LA, Desler ML, et al. Auditory and visual attention assessed with PET. *Human Brain Mapping*, 5: 422–436, 1997.

Olson IR, Gatenby JC, and Gore JC. A comparison of bound and unbound audio–visual information processing in the human cerebral cortex. *Cognitive Brain Research*, 14: 129–138, 2002.

Rämä P and Courtney SM. Functional topography of working memory for face or voice identity. *NeuroImage*, 24: 224–234, 2005.

Rhodes G, Byatt G, Michie PT, and Puce A. Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience*, 16(2): 189–203, 2004.

Robertson DMC and Schweinberger SR. The role of audiovisual asynchrony in person recognition. *The Quarterly Journal of Experimental Psychology*, 63(1): 23–30, 2010.

Saito DN, Yoshimura K, Kochiyama T, Okada T, Honda M, and Sadato N. Cross-modal binding and activated attentional networks during audio–visual speech integration: A functional MRI study. *Cerebral Cortex*, 15(11): 1750–1760, 2005.

Schneider W, Eschman A, and Zuccolotto A. *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc., 2002.

Schweinberger SR, Herholz A, and Stief V. Auditory long-term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology*, 50(A): 498–517, 1997.

Schweinberger SR, Robertson D, and Kaufmann JM. Hearing facial identities. *The Quarterly Journal of Experimental Psychology*, 60 (10): 1446–1456, 2007.

Sekiyama K, Kanno I, Miura S, and Sugita Y. Auditory–visual speech perception examined by fMRI and PET. *Neuroscience Research*, 47: 277–287, 2003.

Sergent J, Ohta S, and McDonald B. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115(1): 15–36, 1992.

Shomstein S and Yantis S. Control of attention shifts between vision and audition in human cortex. *The Journal of Neuroscience*, 24(47): 10702–10706, 2004.

Stark CEL and Squire LR. Simple and associative recognition memory in the hippocampal region. *Learning and Memory*, 8: 190–197, 2001.

Stevenson RA, Geoghegan ML, and James TW. Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. *Experimental Brain Research*, 179: 85–95, 2007.

Stevenson RA and James TW. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage*, 44: 1210–1223, 2009.

Takahashi N, Kawamura M, Hirayama K, Shiota J, and Tsono O. Prosopagnosia: A clinical and anatomical study of four patients. *Cortex*, 31: 317–329, 1995.

Tsukiura T, Mochizuki-Kawai H, and Fujii T. Dissociable roles of the bilateral anterior temporal lobe in face–name associations: An event-related fMRI study. *NeuroImage*, 30(2): 617–626, 2005.

Van Lancker DR, Kreiman J, and Cummings J. Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, 11(5): 665–674, 1989.

Von Kriegstein K, Eger E, Kleinschmidt A, and Giraud AL. Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research Cognitive Brain Research*, 17: 48–55, 2003.

Von Kriegstein K, Kleinschmidt A, Sterzer P, and Giraud AL. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, 17(3): 367–376, 2005.

Von Kriegstein K and Giraud AL. Implicit multisensory associations influence voice recognition. *PLoS Biology*, 4: e326, 2006.

Wada Y, Kitagawa N, and Noguchi K. Audio–visual integration in temporal perception. *International Journal of Psychophysiology*, 50: 117–124, 2003.

Please cite this article in press as: Joassin Fré, et al., Cross-modal interactions between human faces and voices involved in person recognition, Cortex (2010), doi:10.1016/j.cortex.2010.03.003