NeuroImage xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study

Frédéric Joassin [a,*], Pierre Maurage [a], Salvatore Campanella [b]

[a] *Université catholique de Louvain, IPSY/NEUROCS, Louvain-la-Neuve, Belgium*

[b] *Université Libre de Bruxelles, CHU Brugmann, Laboratory of Psychological Medecine, Brussels, Belgium*

ARTICLE INFO

Article history:
Received 1 July 2010
Revised 30 August 2010
Accepted 31 August 2010
Available online xxxx

Keywords:
Crossmodal
Gender
Faces
Voices
Audiovisual
fMRI

ABSTRACT

The aim of this fMRI study was to investigate the cerebral crossmodal interactions between human faces and voices during a gender categorization task.

Twelve healthy male participants took part to the study. They were scanned in 4 runs that contained 3 conditions consisting in the presentation of faces, voices or congruent face–voice pairs. The task consisted in categorizing each trial (visual, auditory or associations) according to its gender (male or female).

The subtraction between the bimodal condition and the sum of the unimodal ones showed that categorizing face/voice associations according to their gender produced unimodal activations of the visual (right calcarine sulcus) and auditory regions (bilateral superior temporal gyri), and specific supramodal activations of the left superior parietal gyrus and the right inferior frontal gyrus. Moreover, psychophysiological interaction analyses (PPI) revealed that both unimodal regions were inter-connected and connected to the prefrontal gyrus and the putamen, and that the left parietal gyrus had an enhanced connectivity with a parieto-premotor circuit involved in the crossmodal control of attention.

This fMRI study showed that the crossmodal auditory–visual categorization of human gender is sustained by a network of cerebral regions highly similar to those observed in our previous studies examining the crossmodal interactions involved in face/voice recognition (Joassin et al., in press). This suggests that the crossmodal processing of human stimuli requires the activation of a network of cortical regions, including both unimodal visual and auditory regions and supramodal parietal and frontal regions involved in the integration of both faces and voices and in the crossmodal attentional processes, and activated independently from the task to perform or the cognitive level of processing.

Introduction

In daily life, our social interactions are guided by our ability to integrate distinct sensory inputs into a coherent multimodal representation of our interlocutors. For instance, we are able to integrate the auditory information of what is said and the visual information of who is saying it, so that we can attribute a particular speech to a particular person (Kerlin et al., 2010) and thus take part to a conversation. Numerous studies have examined the cerebral correlates of the auditory–visual speech perception (Calvert et al., 2000; Von Kriegstein et al., 2008; Stevenson et al., 2010), underlining the role of the right superior temporal sulcus (STS) in such processes.

Nevertheless, crossmodal interactions occur not only during speech perception but also during the memory processes allowing the identification of familiar people (Campanella and Belin, 2007). For encoding, Sheffert and Olson (2004) have shown that the learning of voice identities was easier when the voices to learn were associated with a face, revealing crossmodal connections similar to those observed in audiovisual speech processing. For recognition, we have recently shown that the recognition of previously learned face–voice associations, compared to the recognition of faces or voices presented alone, activated a cerebral network including unimodal face and voice areas (respectively the right Face Fusiform Area, FFA, Kanwisher et al., 1997; and the right STS, Belin et al., 2004), but also regions whose activation was only observed in the bimodal condition, such as the right hippocampus and the left angular gyrus (Joassin et al., in press).

There are two main hypotheses that have emerged to explain the crossmodal cerebral integration process. The first one postulates direct links between the unimodal regions processing the distinct sensory stimuli (Von Kriegstein et al., 2005, 2006). For instance, the authors showed that the right FFA had an enhanced connectivity with the right STS during speaker recognition, suggesting that multimodal person recognition does not necessarily engage supramodal cortical substrates but can result from the direct sharing of information between the unimodal auditory and visual regions (Von Kriegstein

* Corresponding author. Université catholique de Louvain, Faculté de Psychologie et des Sciences de l'Education – IPSY/NEUROCS, Place Cardinal Mercier, 10, 1348 Louvain-la-Neuve, Belgium. Fax: +32 10 47 37 74.
E-mail address: frederic.joassin@uclouvain.be (F. Joassin).

1053-8119/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2010.08.073

and Giraud, 2006). One possible neural mechanism for such direct links between unimodal regions could be the synchronization of the oscillatory activities of assemblies of neurons, especially in the gamma-band frequency range (>30 Hz, for a review, see Senkowski et al., 2008).

On the other hand, the alternative hypothesis proposes that the crossmodal integration of faces and voices relies on the activation of a neural network including supramodal convergence regions (Driver and Spence, 2000; Bushara et al., 2003). Our previous experiments support this second hypothesis as they revealed a specific activation of supramodal regions such as the right hippocampus and the left inferior parietal regions during the bimodal recognition of previously learned face–name (Campanella et al., 2001; Joassin et al., 2004a) and face–voice associations (Joassin et al., 2004b; in press). This last region, as a part of the associative cortex, is known to be involved in the binding of distinct sensory features (Damasio, 1989; Booth et al., 2002, 2003). Bernstein et al. (2008), using Event-Related Potentials (ERP), observed a specific cerebral activity of the left angular gyrus during audiovisual speech perception, suggesting that this region plays a role in the multimodal integration of visual and auditory speech perception. The precise role of this region in face–voice integration could be related to the multimodal control of attention. Indeed, in our own experiment, a psychophysiological interaction analysis (PPI, Friston, 2004) revealed that the left angular gyrus had an enhanced connectivity with the cerebellum and motor and pre-motor regions including the supplementary motor area and the middle and superior frontal gyri (Joassin et al., in press). This parieto-premotor cortical network is important for the control of attention (Driver & Spence, 1998) and has been reported in several studies using crossmodal stimuli (O'Leary et al., 1997; Bushara et al., 1999, Shomstein & Yantis, 2004). It is thus possible that the parieto-premotor network observed in the present study acts to simultaneously direct attention to targets from distinct sensory modalities (Lewis et al., 2000).

Nevertheless, the results of our previous experiments raised several questions, notably about the specificity of the neural network involved in the multimodal recognition of familiar people. The classical cognitive models of face identification have postulated that recognition, i.e. the access to the biographical information and the name of a familiar person, is independent from the processing of the other facial features such as the ethnicity, the age or the gender (Bruce and Young, 1986; Burton et al., 1990). However, several recent studies have challenged this idea and proposed that gender and identity are processed by a single route. Ganel and Goshen-Gottstein (2002) showed that participants could not selectively attend to either sex or identity without being influenced by the other feature, suggesting that both information are processed by a single route. Moreover, Smith et al. (2007) have recently shown that auditory and visual information interact during face gender processing. In their experiment, participants had to categorize androgynous faces according to their gender. These faces were coupled with pure tones in the male or female fundamental-speaking-frequency range. They found that faces were judged as male faces when coupled with a pure male tone while they were judged as female ones when coupled with a pure female tone.

The aim of the present experiment was thus to investigate the crossmodal audiovisual interactions during gender processing with real faces and voices, in a more ecological approach of face–voice integration processes. We used an experimental paradigm similar to those used in our previous studies (Campanella et al., 2001; Joassin et al., 2004a; 2004b; 2007: in press), enabling the direct comparison between a bimodal condition (FV) in which both faces and voices were presented synchronously and two unimodal conditions in which faces and voices were presented separately (F and V). This paradigm allowed us to perform the main contrast $[FV-(F+V)]$ in order to isolate the specific activations elicited by the integration of faces and voices during gender categorization. This method uses a super-additive criterion to detect these specific activations, requiring multisensory responses larger than the sum of the unisensory responses (Calvert et al., 2001; Beauchamp, 2005). This criterion has often been considered as overly strict in the 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, as the activations observed in our previous experiments have been obtained by this way (Campanella et al., 2001; Joassin et al., 2004a; 2004b; 2007; in press), we decided to continue to apply the same super-additive criterion. In the same way, we used static faces identical to those used in our previous experiments (Joassin et al., 2004b; in press) in order to keep the same general methods and to be able to compare the results of these distinct experiments between each other.

We predicted that if gender and identification processing share a single cognitive route, audiovisual gender categorization should activate the same cerebral network than the recognition of face–voice associations, i.e. a network of cerebral regions composed of the unimodal face and voice areas and supramodal integration regions including left parietal and prefrontal regions.

Methods

Participants

Twelve healthy undergraduate participants performed this fMRI experiment (7 females, mean age: 25.75, SD: 5.01). All were right-handed, French native speakers, had a normal-to-corrected vision and a normal audition, and gave their written informed consent. The experimental protocol was approved by the Biomedical Ethical Committee of the Catholic University of Louvain.

Stimuli

Twelve face–voice associations (6 males) were used in the experiment. Each face–voice association was composed of a static picture of a face (black and white photo, front view, neutral expression, picked from the Stirling Face Database: http://pics.psych.stir.ac.uk) and a voice recorded in our laboratory and saying the French word «bonjour» with a neutral prosody. These voices were selected from a validated battery of vocal prosodies recorded in our laboratory (Maurage et al., 2007a). We used a word rather than a simple syllable to increase the ecological value of the face–voice pairs. All visual stimuli were controlled for contrast and brightness and had an approximative size of 350×350 pixels. All auditory stimuli (presented in Mono, 44100 Hz, 32 bit) were controlled for duration (mean duration of 700 msec) and normalized for amplitude (in dB).

Procedure

Three conditions were presented during the fMRI sessions, and Blood Oxygenation Level-Dependent (BOLD) signal changes were measured while participants had to categorize faces (F), voices (V) and face–voice associations (FV) according to their gender. Participants had to judge as quickly as possible the sex (male or female) of each trial by pressing one of two buttons of a response pad with 2 fingers of the right hand.

Each participant underwent 4 block-designed acquisition runs. Each run comprised 6 experimental blocks of 30 sec (3 conditions repeated once) interleaved with 15-sec fixation periods (white cross on a black background). Each block was composed of 12 trials and each trial was composed of a fixation cross (300 msec), followed by the stimulus for 700 msec and an empty interval of 1500 msec.

Please cite this article as: Joassin, F., et al., The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.08.073

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.

Imaging procedure

Functional images were acquired with a 3.0 T magnetic resonance imager and an 8-channel phased array head coil (Achieva, Philips Medical Systems) as a series of blood-oxygen-sensitive T2*-weighted echo-planar image volumes (GRE-EPI). Acquisition parameters were as follows: TE = 32 ms, TR = 2500 ms, flip angle = 90°, field of view = 220 × 220 mm, slice thickness = 3.5 mm with no interslice gap, and SENSE factor (parallel imaging) = 2.5. Each image volume comprised 36 axial slices acquired in an ascending interleaved sequence. Each functional run comprised 108 volumes, 36 corresponding to the fixation periods and the remaining 72 corresponding to the experimental blocks (24 volumes per condition per run). High-resolution anatomical images were also acquired for each participant using a T1-weighted 3D turbo fast field echo sequence with an inversion recovery prepulse (150 contiguous axial slices of 1 mm, TE = 4.6 ms, TR = 9.1 ms, flip angle = 8°, FOV = 220 × 197 mm, voxel size = 0.81 × 0.95 × 1 mm^3, and SENSE factor = 1.4). Head movement was limited by foam padding within the head coil and a restraining band across the forehead.

fMRI data analysis

Data were processed and analyzed using Statistical Parametric Mapping (SPM2, 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.). Functional images were (1) corrected for slice acquisition delays, (2) realigned to the first scan of the first run (closest to the anatomical scan) to correct for within- and between-run motion, (3) coregistered with the anatomical scan, (4) normalized to the MNI template using an affine fourth degree ß-spline interpolation transformation and a voxel size of 2 × 2 × 2 mm^3 after the skull and bones had been removed with a mask based on the individual anatomical images, and (5) spatially smoothed using a 10-mm FWHM Gaussian kernel.

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 face–voice associations ([FV-fix]) relative to the fixation periods used as a general baseline. The contrast [FV − (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 several 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 analysis focused on one particular region observed in the group analysis. For each participant, we performed a small volume correction (a sphere of 5 mm centered on the maximum peak of activity of each region in the group analysis) before extracting the deconvolved time course of activity in a ROI (a 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 $p < 0.001$, as in Ethofer et al., 2006).

Results

Behavioral data

Reaction times

The mean reaction times of the visual, auditory and audiovisual conditions were respectively 588.1 ms (SD: 87.4), 708.6 ms (SD: 121.5) and 551 ms (SD: 85.7, Fig. 1).

An ANOVA with the modality (audiovisual, auditory and visual) as within-subjects factors was performed on the reaction times. It revealed significant main effects of the modality ($F(2,22) = 24.478$, $p < .0001$). Subsequent one-tailed paired Student *t*-tests, using a Bonferoni correction for multiple comparisons showed that the bimodal condition was performed faster than the auditory ($t(11) = -5{,}3$, $p < .001$) and the visual ($t(11) = -2{,}6$, $p < .03$) conditions. The visual condition was also performed faster than the auditory condition ($t(11) = -4{,}9$, $p < .001$) (Fig. 2).

Percentages of correct responses

The mean percentages of correct answers for the visual, auditory and audiovisual conditions were respectively 97.4% (SD: 2.2), 98.4% (SD: 1.3) and 98.2% (SD: 1.2).

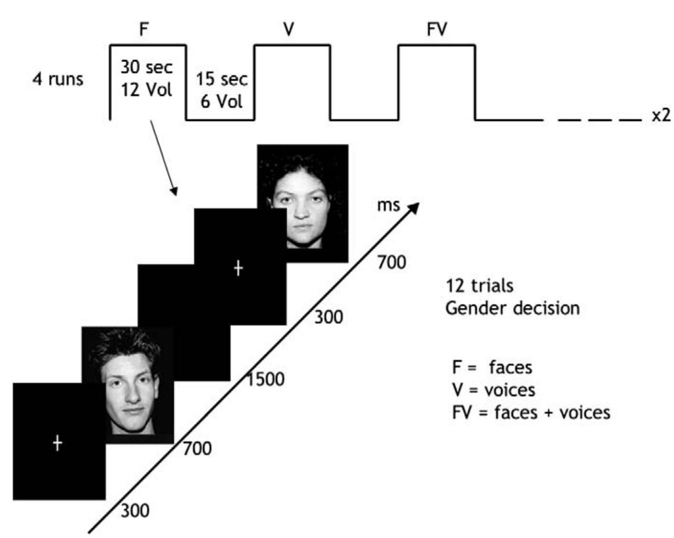

Fig. 1. a) fMRI design: each run consisted in 6 alternances of a 15-sec fixation period (white cross on black background) and a 30-sec activation period. Each activation period corresponded to a different condition (F, V, FV), presented twice in a pseudo-random order. Participants were presented with 12 trials in each condition. Each trial comprised a fixation cross for 300 ms, a stimulus — faces (F), voices (V), or face/voice associations (FV) — for 700 ms and a black intertrial interval for 1500 ms.

Please cite this article as: Joassin, F., et al., The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.08.073

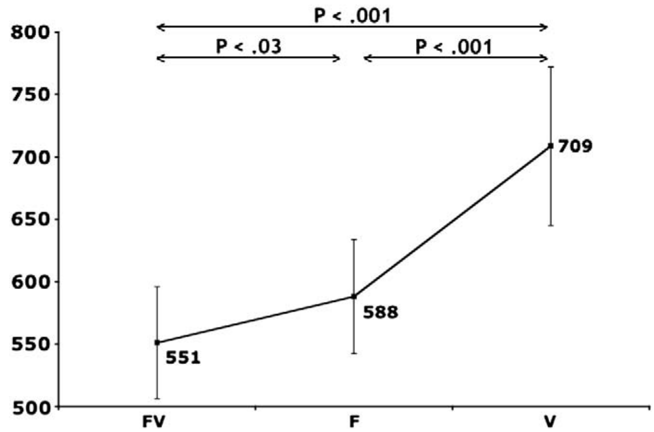

Fig. 2. Mean reaction times (in ms) for the audiovisual (FV), visual (F) and auditory (V) conditions.

The same ANOVA was carried out but failed to reveal any significant effect of the modality ($F(1,11) = 1.47$, ns).

To summarize, the analysis of the behavioral data showed that the face–voice associations produced a clear crossmodal facilitation effect as they were judged significantly faster than the faces and the voice presented alone. This effect was not observed for the performances but it is probably due to a ceiling effect, all the percentages of correct responses being above 97% (Table 1).

Brain imaging results

Gender categorization of the unimodal stimuli

Judging the sex of human faces ([F-fix]) activated the bilateral fusiform gyri, the right inferior frontal gyrus, the left calcarine sulcus, the left thalamus, the left and right inferior parietal gyri and the left putamen (Table 2a).

Judging the sex of human voices ([V-fix]) activated the left and right superior temporal gyri, the right inferior frontal gyrus and the bilateral regions of the cerebellum (Table 2b).

Gender categorization of the associations

Judging the sex of face–voice associations ([FV-fix]) activated the left and right superior and middle temporal gyri, including the left supramarginal and angular gyri, the right inferior occipital gyrus, the left putamen, the left precuneus and the right inferior parietal gyrus (Table 2a).

Table 1
Brain regions showing significant activation compared to baseline (fix) for faces (a) and voices (b).

	Brain regions	x	y	z	L/R	k	t-statistic
a	F-fix						
	Fusiform gyrus	−38	−50	−22	L	8877	5.69
	Fusiform gyrus	40	−56	−20	R		
	Inferior frontal gyrus	42	22	24	R	134	4.75
	Calcarine sulcus	−18	−78	4	L	96	4.34
	Thalamus	−14	−14	10	L	209	4.15
	Inferior parietal gyrus	−32	−46	54	L	197	4.13
	Inferior parietal gyrus	30	−52	52	R	110	4.10
	Putamen	−20	8	8	L	92	3.89
b	V-fix						
	Superior temporal gyrus	−48	−28	0	L	2996	5.41
	Superior temporal gyrus	42	−28	12	R	3414	5.39
	Inferior frontal gyrus	52	16	30	R	175	4.29
	Cerebellum	50	−60	−32	R	1090	4.98

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.

Table 2
Brain regions showing significant activation for face–voice associations (a) compared to baseline (fix), and for subtractions between face–voice associations and unimodal faces and voices (b).

	Brain regions	x	y	z	L/R	k	t-statistic
a	FV−fix						
	Middle temporal gyrus	58	−22	−2	R	4411	5.51
	Putamen	−22	8	−2	L	502	5.45
	Inferior occipital gyrus	44	−82	−8	R	8351	5.11
	Superior temporal gyrus	−52	−38	18	L	6485	5.18
	Supramarginal gyrus	−6	−42	24	L		5.06
	Inferior parietal gyrus	34	−52	46	R	148	4.27
	Precuneus	−54	10	38	L	72	3.87
b	FV − (F + V)						
	Calcarine sulcus	12	−92	−8	R	5188	5.96
	Fusiform gyrus	−42	−46	−22	L		5.86
	Superior temporal gyrus	66	−14	4	R	1295	5.59
	Superior temporal gyrus	−56	−38	8	L	1342	4.70
	Superior parietal gyrus	−38	−54	58	L	242	4.31
	Frontal inferior gyrus	36	24	26	R	249	4.29
	Middle occipital gyrus	−32	−76	40	L	122	4.22

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.

Subtractions between the unimodal and bimodal conditions

The main contrasts of this experiment consisted in subtracting the cerebral activities elicited by the gender categorization of unimodal visual and auditory stimuli from the cerebral activities elicited by the gender categorization of audiovisual stimuli, in order to isolate the specific activations involved in the integration of visual and auditory information during gender processing.

The contrast [FV − (F + V)] revealed an extensive activation of the visual and auditory regions including respectively the right calcarine sulcus and the left fusiform and middle occipital gyri, and the left and right superior temporal gyri (Fig. 3). We also observed specific integrative activations in the left superior parietal gyrus including the angular gyrus and the right inferior frontal gyrus (Table 2b and Fig. 4).

Functional connectivity analyses

The psychophysiological interactions analyses were performed to examine the functional connectivity of the cerebral regions observed in the subtraction. It showed that the left inferior parietal gyrus had an enhanced connectivity with the right fusiform gyrus, the left

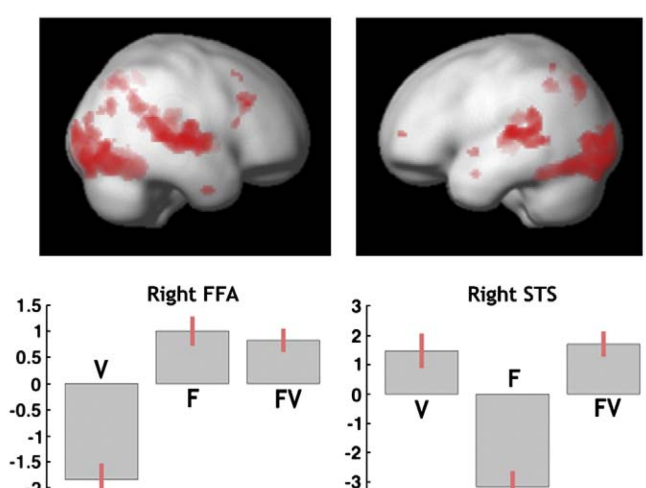

Fig. 3. Brain regions activated in the contrast [FV − (V + F)]. Upper part: statistical parametric maps superimposed on MRI surface renders (left and right views); lower part: activation changes for each condition in the right calcarine sulcus (left histogram) and the right STS (right histogram). $p<.05$ corrected for multiple comparisons at cluster size. V = voices, F = faces, VF = face/voice associations.

Please cite this article as: Joassin, F., et al., The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.08.073

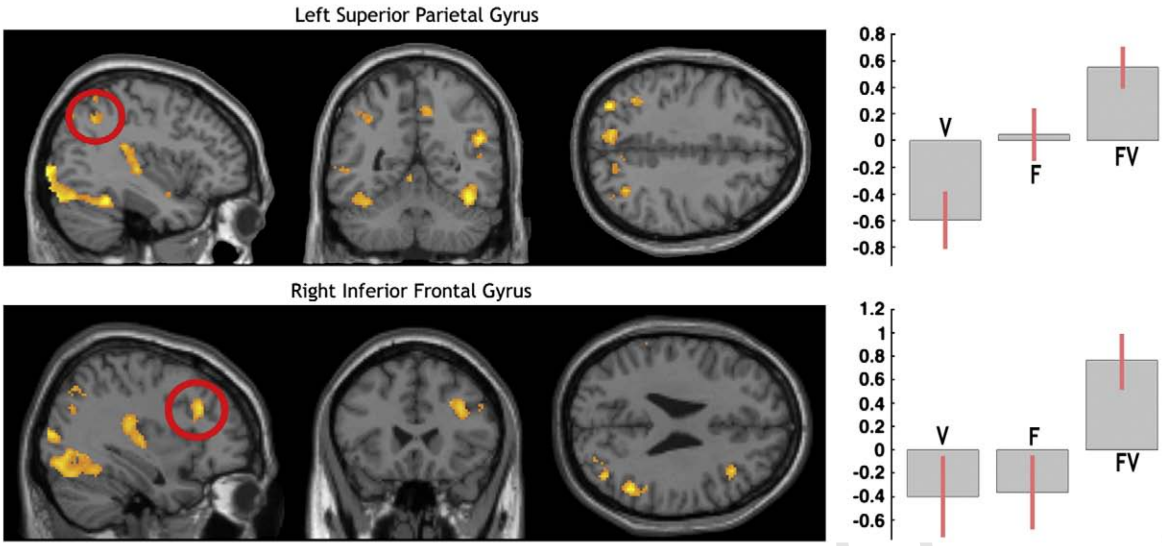

Fig. 4. Left side: brain sections of the contrast [FV − (V + F)] centered on the left superior parietal gyrus (upper part) and the right inferior frontal gyrus (lower part). Right side: activation changes for each condition in the left superior parietal gyrus (upper part) and the right inferior frontal gyrus (lower part). $p < .05$ corrected for multiple comparisons at cluster size. V = voices, F = faces, FV = face/voice associations.

Q4 supplementary motor area and the right cerebellum (Table 3a). The right STS had an enhanced connectivity with the left auditory STS, the left visual fusiform gyrus and the left and right putamen (Table 3b). The right calcarine sulcus had an enhanced connectivity with the right STS and the left putamen (Table 3c). Finally, the right inferior frontal gyrus had an enhanced connectivity with the right supramarginal gyrus, the left inferior occipital gyrus and the left and right STS (Table 3d).

Discussion

This experiment was aimed at examining the specific cerebral activations elicited by the gender categorization of face–voice associations, by comparing them with the cerebral activations produced by the gender processing of unimodal faces and voices.

The analysis of the behavioral data clearly showed a crossmodal facilitation effect in the case of the face–voice associations. The congruent simultaneous presentation of faces and voices helped the participants to process the gender, as these associations were categorized as male or female significantly faster than the faces and the voices presented alone. This facilitation was not observed on the percentages of correct responses, probably due to a ceiling effect on this measure. In our previous experiments, in which we tested the recognition (i.e. the access to the identity) of face–voice associations, we observed that voices were more difficult to recognize than faces and that their simultaneous presentation hampered rather than facilitated recognition (Joassin et al., 2004b; in press). Several factors are known to influence the behavioral crossmodal effects, such as the spatio-temporal proximity (Meredith and Stein, 1986; Robertson and Q5 Schweinberger, 2010) or the semantic congruency (Calvert et al., 2001). On the basis of our previous results, we proposed that the perceptive complexity of the stimuli could be a supplementary factor that could determine the potential gain of bimodal stimulations (Joassin et al., 2008). Nevertheless, it seems that the differences of perceptual complexity (or of expertise) between faces and voices are not the only factor influencing the behavioral crossmodal effects, as in the present case, although the voices were categorized more slowly than the faces, their simultaneous presentation fastened the responses. The level of processing (low such as gender perception, or high such as access to the identity) seems thus to play also an important role in the way in which faces and voices interact.

These crossmodal interactions were examined by a contrast consisting of subtracting the two unimodal conditions from the bimodal one. This contrast showed that the crossmodal processing of faces and voices was sustained by a neural network composed of the unimodal visual and auditory regions but also of two regions, the left superior parietal cortex and the right inferior frontal gyrus, whose activations was specific to the bimodal condition. It is important to note that this network is highly similar to the network of cerebral regions observed in our previous experiment testing the crossmodal recognition of face–voice associations (Joassin et al., in press). In this experiment, the subtraction between unimodal and bimodal conditions also revealed an activation of the unimodal visual and auditory regions and of the left angular gyrus. It seems thus that the involvement of this network does not depend on the level of processing of faces and voices or the task to perform, but is rather specific to the human stimuli.

The activation of the unimodal regions during bimodal stimulations has already been showed (Ghazanfar et al., 2005; Calvert et al., 1999) and it has been hypothesized that direct connections between the unimodal regions could be sufficient to ensure the crossmodal

Table 3
Brain regions showing an enhanced connectivity with the regions activated in the contrast [FV − (F + V)].

	Brain regions	x	y	z	L/R	k	t-statistic
a	Left inferior parietal gyrus						
	Supplementary motor area	−8	20	52	L	29	3.68*
	Fusiform gyrus	44	−50	26	R	102	3.67
	Cerebellum	46	−72	−20	R	20	3.26*
b	Right STS						
	Superior temporal gyrus	−44	−28	18	L	616	4.35
	Fusiform gyrus	−34	−48	−24	L	81	4.25
	Putamen	22	2	12	R	70	4.02
	Putamen	−22	8	4	L	91	3.95
c	Right calcarine sulcus						
	Superior temporal gyrus	44	−18	6	R	91	4.21
	Putamen	−22	6	−4	L	76	4.47
d	Right inferior frontal gyrus						
	Supramarginal gyrus	52	−40	28	R	119	4.46
	Superior temporal gyrus	−52	−16	8	L	159	4.16
	Superior temporal gyrus	50	−12	2	R	27	3.79*
	Inferior occipital gyrus	−38	−80	−2	L	1	3.14*

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. * = Threshold set at $p < .001$ uncorrected.

integration (Von Kriegstein et al., 2005). Nevertheless, the face–voice associations elicited a specific activation of the left superior parietal gyrus (Fig. 5). The activation of this particular region has already been observed in several of our previous experiments (Campanella et al., 2001; Joassin et al., 2004a; 2004b; in press). This region is known to be a part of the associative cortex and receives multiple inputs from modality specific sensory regions (Damasio, 1989; Rämä and Courtney, 2005; Niznikiewicz et al., 2000; Bernstein et al., 2008). More specifically, it could be involved in processes of divided attention, allowing to direct attention simultaneously to targets from distinct sensory modalities (Lewis et al., 2000).

The PPI analysis centered on the left parietal cortex showed that this region had an enhanced connectivity with the cerebellum and the supplementary motor area. This cerebello-parieto-motor network is important in the crossmodal control of attention (Bushara et al., 1999; Driver & Spence, 2000; Shomstein & Yantis, 2004), and it could sustain the integration of faces and voices by allowing an optimal dispatching of the attentional resources between the visual and the auditory modalities. The behavioral data reinforce this interpretation as they showed a clear crossmodal facilitation effect for face–voice pairs. The categorization was thus based on the processing of both stimuli, which needed the attentional resources to be dispatched between both sensory modalities.

The PPI analyses also showed that the unimodal visual and auditory regions were inter-connected, but had also an enhanced connectivity with several other cerebral regions. At first, the left putamen, as a part of the striatum, is known to play the role of a subcortical integration relay allowing to access and regulate multimodal information by means of dopaminergic channels (Haruno & Kawato, 2006). Secondly, we observed that the unimodal regions were also connected to the right inferior frontal gyrus (Brodmann area 44). This region was activated in both unimodal conditions and is known to receive inputs from face (Rolls, 2000; Leube et al., 2001) and voice sensitive areas (Hesling et al., 2005; Rämä and Courtney, 2005). While it has been suggested that its activation could reflect some memory encoding processes of new faces (Leube et al., 2001), the right BA44 was also found to be specifically involved in the processing of F0 modulation, the main acoustic correlate of prosody (Hesling et al., 2005). These differential interpretations suggest that the right BA44 is a heterogeneous region that could be segregated in different parts sustaining distinct cognitive processes. Its more central activation in the bimodal condition reinforces this idea. Indeed, the activation of this region has already been observed in auditory–visual integration experiments (Taylor et al., 2006) and its role in the integration of faces and voices could be related to the processing of the semantic congruency (Hein et al., 2007). These authors, using fMRI, investigated the cerebral regions involved in the integration of congruent/incongruent familiar animal sounds and pictures and in the integration of unfamiliar arbitrary associations between sounds and object images. They found that processing both familiar incongruent associations and unfamiliar visuo-auditory associations elicited a specific activation of the right inferior frontal cortex. This activation was interpreted as reflecting a sensibility of this region to semantic congruency but also an involvement in the learning of novel visuo-auditory associations, as suggested by Gonzalo et al. (2000). Supporting this interpretation, McNamara et al. (2008) showed that the right BA44 was activated by the learning of new associations between an arbitrary sound and a gesture. Further experiments, investigating the encoding of such face–voice associations would be helpful to better understand the role of the frontal regions in the crossmodal processing of human stimuli.

The study of the crossmodal processes between sensory modalities is particularly important for a better understanding of the neural networks operating in the healthy brain, but is also important to better understand the neuro-functional impairments in several psychopathological and developmental disorders. For instance, we have recently shown that chronic alcoholism is associated with a specific impairment of the visuo-auditory recognition of emotions (Maurage et al., 2007b), and that it is linked to a hypo activation of the prefrontal regions (Maurage et al., 2008).

Moreover, it appears that the impairments in the recognition of the emotions might be due to distinct neuro-functional impairments such as a deficit of the connectivity between several brain regions in autism (the amygdala and the associative temporal and prefrontal gryi, Monk et al., 2010), or a hypo activation of the visual regions in other pathological conditions such as schizophrenia (Seiferth et al., 2009). It seems thus that a common symptom — the impairment of the processing of emotions — might be due to different neuro-functional deficits. Exploring the multimodal integration in the growing brain is also particularly important, as it seems that difficulties in information integration may lead to some developmental disorders, such as autism (Melillo and Leisman, 2009) or the Pervasive Developmental Disorders (PDD, Magnée et al., 2008).These authors observed a specific decrease of the electrical cerebral activity when PDD patients were confronted to emotional face/voice pairs. It suggests that the processing difficulties of emotional information during the development could be linked to abnormal patterns of the multimodal cerebral activity. These studies focused on the processing of emotions and it will be interesting to further explore the impairments of the multimodal processing of faces and voices in psychopathology, with the hypothesis of an impairment of the multimodal processing of the human stimuli at other cognitive levels (perceptual processing, gender, recognition, Norton et al., 2009). This new approach could lead to multimodal therapy that would include a combination of somatosensory, cognitive, behavioral, and biochemical interventions (Melillo and Leisman, 2009).

A possible limitation of the present study lies in the fact that we used static faces. Indeed, Schweinberger and his collaborators have

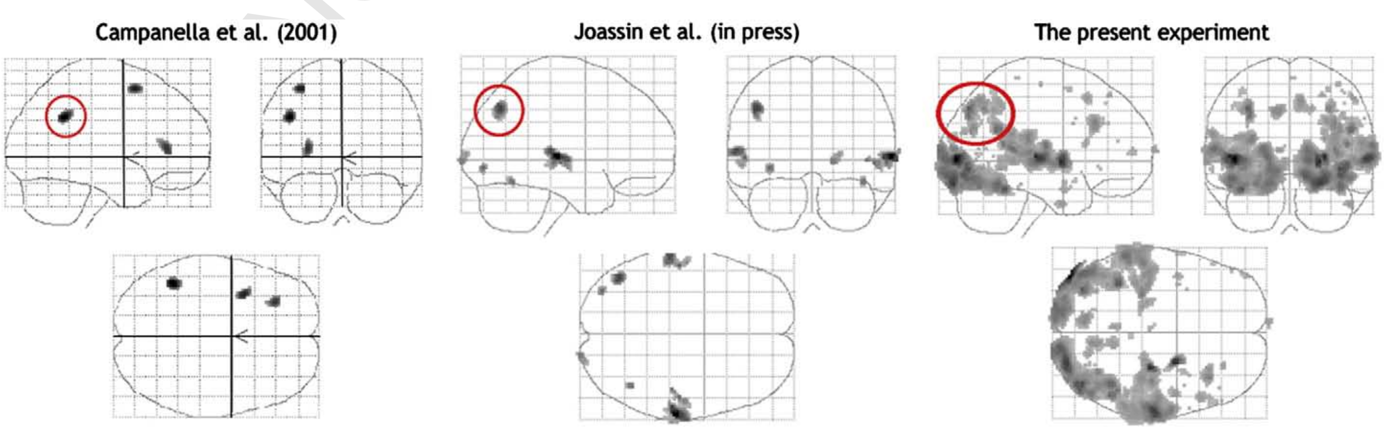

Fig. 5. Illustration of the activations of the left parietal gyrus observed by Campanella et al. (2001), Joassin et al. (in press) and in the present experiment.

Please cite this article as: Joassin, F., et al., The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.08.073

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 audiovisual 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. However, even if the use of dynamic stimuli should be envisaged in our further studies, the fact that we observed a clear behavioral crossmodal facilitation and some specific activations in the bimodal condition ensured that both faces and voices were explicitly processed and that participants took benefit from the bimodal situations.

Further studies should also take benefit from new experimental paradigms and techniques to test the crossmodal gender processing in complex situations in which faces and/or voices are difficult to process. Using morphing techniques (Campanella et al., 2002; Bruckert et al., 2010) would allow us to create ambiguous face–voice pairs and to test, in a controlled experimental manipulation, the respective influence of faces and voices in the crossmodal processing of gender. Such experiments would be highly relevant for the study of the crossmodal interactions between faces and voices in the growing brain. While faces evolve in a continuous way during the development, voices change brutally in pitch and fundamental frequency at puberty (Harries et al., 1997).We can thus assume that faces and voices do not participate to social cognition with an identical weight during childhood, adolescence and adulthood.

Conclusions

In conclusion, the auditory–visual integration of human faces and voices during the multimodal processing of the gender was associated with the activation of a specific network of cortical and subcortical regions. This network included several regions devoted to the different cognitive processing implied in the gender categorization task — the unimodal visual and auditory regions processing the perceived faces and voices and inter-connected via a subcortical relay located in the striatum, the left superior parietal gyrus, part of a larger parieto-motor network dispatching the attentional resources to the visual and auditory modalities, and the right inferior frontal gyrus sustaining the integration of the semantically congruent information into a coherent multimodal representation.

The similarity between the present results and the activations observed in our previous experiment supports the hypothesis that the integration of human faces and voices is sustained by a network of cerebral regions activated independently of the task to perform or the cognitive level of processing (gender processing or recognition). These results raise several new questions that further experiments will help to answer, notably about the possible specificity of the observed network for the processing of the human stimuli relative to other kinds of visuo-auditory associations or the explicit/controlled *vs*. implicit/automatic aspects in the integration of highly ecological social stimuli such as the human faces and voices.

Acknowledgments

Frédéric Joassin and Pierre Maurage are Postdoctoral Researchers, and Salvatore Campanella a Research Associate at the National Fund for Scientific Research (F.N.R.S., Belgium). We thank Dr. Cécile Grandin and the Radiodiagnosis Unit at the Cliniques St. Luc (Brussels) for their support, Ms. Valérie Dormal and Ms. Sue Hamilton for their helpful suggestions during the redaction of this article.

References

Beauchamp, M.S., 2005. Statistical criteria in fMRI studies of multisensory integration. Neuroinformatics 3, 93–113.
Belin, P., Fecteau, S., Bédart, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. 8 (3), 129–135.
Bernstein, L.E., Auer Jr, E.T., Wagner, M., Ponton, C.W., 2008. Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 39, 423–435.
Booth, J.R., Burman, D.D., Meyer, J.R., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., 2002. Functional anatomy of intra- and cross-modal lexical tasks. Neuroimage 16, 7–22.
Booth, J.R., Burman, D.D., Meyer, J.R., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., 2003. Relation between brain activation and lexical performance. Hum. Brain Mapp. 19, 155–169.
Bruce, V., Young, A., 1986. Understanding face recognition. Br. J. Psychol. 77 (3), 305–327.
Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Curr. Biol. 20, 116–120.
Burton, A.M., Bruce, V., Johnston, R.A., 1990. Understanding face recognition with an interactive model. Br. J. Psychol. 81, 361–380.
Bushara, K.O., Weeks, R.A., Ishii, K., Catalan, M.J., Tian, B., Rauschecker, J.P., Hallett, M., 1999. Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. Nat. Neurosci. 2, 759–766.
Bushara, K.O., Hanakawa, T., Immish, I., Toma, K., Kansaku, K., Hallett, M., 2003. Neural correlates of cross-modal binding. Nat. Neurosci. 6 (2), 190–195.
Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S., 1999. Response amplification in sensory-specific cortices during crossmodal binding. NeuroReport 10, 2619–2623.
Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in human heteromodal cortex. Curr. Biol. 10, 649–657.
Calvert, G.A., Hansen, P.C., Iversen, S.D., Brammer, M.J., 2001. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage 14, 427–438.
Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends Cogn. Sci. 11 (12), 535–543.
Campanella, S., Joassin, F., Rossion, B., De Volder, A.G., Bruyer, R., Crommelinck, M., 2001. Associations of the distinct visual representations of faces and names: a PET activation study. Neuroimage 14, 873–882.
Campanella, S., Quinet, P., Bruyer, R., Crommelinck, M., Guérit, J.M., 2002. Categorical perception of happiness and fear facial expressions: an ERP study. J. Cogn. Neurosci. 14 (2), 210–227.
Damasio, A.R., 1989. Time-locked multiregional retroactivation: a system-level proposal for neural substrates of recall and recognition. Cognition 33, 25–62.
Driver, J., Spence, C., 1998. Attention and the crossmodal construction of space. Trends Cogn. Sci. 2, 254–262.
Driver, J., Spence, C., 2000. Multisensory perception: beyond modularity and convergence. Curr. Biol. 10, 731–735.
Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., Reiterer, S., Grodd, W., Wildgruber, D., 2006. Impact of voice on emotional judgment of faces: an event-related fMRI study. Hum. Brain Mapp. 27 (9), 707–814.
Friston, K.J., 2004. Functional and effective connectivity in neuroimaging: a synthesis. Hum. Brain Mapp. 2, 56–78.
Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6, 218–229.
Ganel, T., Goshen-Gottstein, Y., 2002. Perceptual integrity of sex and identity of faces: further evidence for the single-route hypothesis. J. Exp. Psychol. Hum. Percept. Perform. 28, 854–867.
Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K., 2005. Multi-sensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J. Neurosci. 25 (20), 5004–5012.
Gonzalo, D., Shallice, T., Dolan, R., 2000. Time-dependent changes in learning audiovisual associations: a single-trial fMRI study. Neuroimage 11, 243–255.
Harries, M.L.L., Walker, J.M., Williams, D.M., Hawkins, S., Hughes, I.A., 1997. Changes in the male voice at puberty. Arch. Dis. Child. 77, 445–447.
Haruno, M., Kawato, M., 2006. Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus–action–reward association learning. Neural Netw. 19, 1242–1254.
Hein, G., Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L., Naumer, M.J., 2007. Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J. Neurosci. 27 (30), 7881–7887.
Hesling, I., Clément, S., Bordessoules, M., Allard, M., 2005. Cerebral mechanisms of prosodic integration: evidence from connected speech. Neuroimage 24, 937–947.
Joassin, F., Campanella, S., Debatisse, D., Guérit, J.M., Bruyer, R., Crommelinck, M., 2004a. The electrophysiological correlates sustaining the retrieval of face–name associations: an ERP study. Psychophysiology 41, 625–635.
Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., Campanella, S., 2004b. When audition alters vision: an event-related potential study of the cross-modal interactions between faces and voices. Neurosci. Lett. 369, 132–137.
Joassin, F., Meert, G., Campanella, S., Bruyer, R., 2007. The associative processes involved in faces–proper names vs. objects–common names binding: a comparative ERP study. Biol. Psychol. 75 (3), 286–299.

Please cite this article as: Joassin, F., et al., The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.08.073

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

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R.,and Campanella, S. (in press). Cross-modal interactions between human faces and voices involved in person recognition. *Cortex*.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 9, 462–475.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional grain control of ongoing cortical speech representations in a «cocktail party». J. Neurosci. 30 (2), 620–628.

Laurienti, P.J., Perrault, T.J., Stanford, T.R., Wallace, M.T., Stein, B.E., 2005. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. Exp. Brain Res. 166, 289–297.

Leube, D.T., Erb, M., Grodd, W., Bartels, M., Kircher, T.T.J., 2001. Differential activation in parahippocampal and prefrontal cortex during word and face encoding tasks. NeuroReport 12 (12), 2773–2777.

Lewis, J.W., Beauchamp, M.S., Deyoe, E.A.A., 2000. Comparison of visual and auditory motion processing in human cerebral cortex. Cereb. Cortex 10, 873–888.

Magnée, M., de Gelder, B., van Engeland, H., Kemner, C., 2008. Atypical processing of fearful face–voice pairs in Pervasive Developmental Disorder: an ERP study. Clin. Neurophysiol. 119, 2004–2010.

Maurage, P., Joassin, F., Philippot, P., Campanella, S., 2007a. A validated battery of vocal emotional expressions. Neuropsychol. Trends 2, 63–74.

Maurage, P., Campanella, S., Philippot, P., Pham, T., Joassin, F., 2007b. The crossmodal facilitation effect is disrupted in alcoholism: a study with emotional stimuli. Alcohol Alcohol. 42, 552–559.

Maurage, P., Philippot, P., Joassin, F., Alonso Prieto, E., Palmero Soler, E., Zanow, F., Campanella, S., 2008. The auditory–visual integration of anger is disrupted in alcoholism: an ERP study. J. Psychiatry Neursci. 33 (2), 111–122.

McNamara, A., Buccino, G., Menz, M.M., Gläsher, J., Wolbers, T., Baumgärtner, A., Binkofski, F., 2008. Neural dynamics of learning sound–action associations. PLoS ONE 3 (12), 1–10.

Melillo, R., Leisman, G., 2009. Autistic spectrum disorders as functional disconnection syndrome. Rev. Neurosci. 20 (2), 111–131.Meredith, M.A., Stein, B.E., 1986. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res. 365 (2), 350–354.

Monk, C., Weng, S.J., Wiggins, J., Kurapati, N., Louro, H., Carrasco, M., Maslowsky, J., Risi, S., Lord, C., 2010. Neural circuitry of emotional face processing in autism spectrum disorders. J. Psychiatry Neurosci. 35 (2), 105–114.

Niznikiewicz, M., Donnino, R., McCarley, R.W., Nestor, P.G., Losifescu, D.V., O'Donnell, B., Levitt, J., Shenton, M.E., 2000. Abnormal angular gyrus asymmetry in schizophrenia. Am. J. Psychiatry 157 (3), 428–437.

Norton, D., McBain, R., Holt, D.J., Ongur, D., Chen, Y., 2009. Association of impaired facial affect recognition with basic facial and visual processing deficits in schizophrenia. Biol. Psychiatry 65, 1094–1098.

O'Leary, D.S., Andreasen, N.C., Hurtig, R.R., Torres, I.J., Flashman, L.A., Desler, M.L., Arndt, S.V., Cizadlo, T.J., Ponto, L.L.B., Watkins, G.L., Hichwa, R.D., 1997. Auditory and visual attention assessed with PET. Hum. Brain Mapp. 5, 422–436.

Rämä, P., Courtney, S.M., 2005. Functional topography of working memory for face or voice identity. Neuroimage 24, 224–234.

Robertson, D.M.C., Schweinberger, S.R., 2010. The role of audiovisual asynchrony in person recognition. Q. J. Exp. Psychol. 63 (1), 23–30.

Rolls, E.T., 2000. The orbitofrontal cortex and reward. Cereb. Cortex 10, 284–294.

Schneider, W., Eschman, A., Zuccolotto, A., 2002. E-prime User's Guide. Psychology Software Tools Inc., Pittsburgh.

Schweinberger, S.R., Robertson, D., Kaufmann, J.M., 2007. Hearing facial identities. Q. J. Exp. Psychol. 60 (10), 1446–1456.

Seiferth, N., Pauly, K., Kellermann, T., Shah, N., Ott, G., Herpertz-Dahlmann, B., Kircher, T., Schneider, F., Habel, U., 2009. Neuronal correlates of facial emotion discrimination in early onset schizophrenia. Neuropsychopharmacology 34, 477–487.

Senkowski, D., Schneider, T.R., Foxe, J.J., Engel, A.K., 2008. Crossmodal binding through neural coherence: implications for multisensory processing. Trends Cogn. Sci. 31 (8), 401–409.

Sheffert, S.M., Olson, E., 2004. Audiovisual speech facilitates voice learning. Percept. Psychophys. 66 (2), 352–362.

Shomstein, S., Yantis, S., 2004. Control of attention shifts between vision and audition in human cortex. J. Neurosci. 24 (47), 10702–10706.

Smith, E.L., Grabowecky, M., Suzuki, S., 2007. Auditory–visual crossmodal integration in perception of face gender. Curr. Biol. 17, 1680–1685.

Stevenson, R.A., Altieri, N.A., Kim, S., Pisoni, D.B., James, T.W., 2010. Neural processing of asynchronous audiovisual speech perception. Neuroimage 49, 3308–3318.

Taylor, K.I., Moss, H.E., Stamatakis, E.A., Tyler, L.K., 2006. Binding crossmodal object features in perirhinal cortex. Proc. Natl Acad. Sci. USA 21, 8239–8244.

Von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biol. 4, e326.

Von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L., 2005. Interaction of face and voice areas during speaker recognition. J. Cogn. Neurosci. 17 (3), 367–376.

Von Kriegstein, K., Dogan, O., Gruter, M., Giraud, A.L., Kell, C.A., Gruter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. Proc. Natl Acad. Sci. USA 105 (18), 6747–6752.