Cerebral correlates and statistical criteria of cross-modal face and voice integration

Scott Love, Frank Pollick, Marianne Latinus

Cerebral Correlates and Statistical Criteria of Cross-Modal Face and Voice Integration *

Scott A. Love [1,**], **Frank E. Pollick** [1] **and Marianne Latinus** [1,2]

[1] School of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK

[2] Centre for Cognitive Neuroimaging (CCNi), Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK

Abstract

Perception of faces and voices plays a prominent role in human social interaction, making multisensory integration of cross-modal speech a topic of great interest in cognitive neuroscience. How to define potential sites of multisensory integration using functional magnetic resonance imaging (fMRI) is currently under debate, with three statistical criteria frequently used (e.g., super-additive, max and mean criteria). In the present fMRI study, 20 participants were scanned in a block design under three stimulus conditions: dynamic unimodal face, unimodal voice and bimodal face–voice. Using this single dataset, we examine all these statistical criteria in an attempt to define loci of face–voice integration. While the super-additive and mean criteria essentially revealed regions in which one of the unimodal responses was a deactivation, the max criterion appeared stringent and only highlighted the left hippocampus as a potential site of face– voice integration. Psychophysiological interaction analysis showed that connectivity between occipital and temporal cortices increased during bimodal compared to unimodal conditions. We concluded that, when investigating multisensory integration with fMRI, all these criteria should be used in conjunction with manipulation of stimulus signal-to-noise ratio and/or cross-modal congruency.

Keywords

Multisensory, audiovisual, fMRI, speech, connectivity, super-additive

1. Introduction

Integrating information provided by the face and the voice plays an important role in human social interaction. Hence, it is a topic of great interest in both psychophysical and neuroimaging investigations of multisensory integration (Campanella and Belin, 2007). Here we present a functional magnetic resonance imaging (fMRI) study designed to investigate the cerebral mechanisms involved in the integration of face and voice information and the statistical criteria used to classify such integration.

Considerable knowledge has already been accumulated about the unimodal processing of faces at both the behavioural and cerebral levels (for an extensive review of both, see Hole and Bourne, 2010). Our understanding of how humans process voices is less advanced yet clearly an important topic of research and consequently, a growing body of knowledge has begun to develop (for reviews, see Belin, Fecteau and Bédard, 2004; Latinus and Belin, 2011). In relation to cross-

modal perception of speech, there is extensive behavioural evidence that the auditory and visual cues to speech interact with each other. As early as 1954, it was shown that speech perception is enhanced in noisy environments when congruent visual information is available (Sumby and Pollack, 1954); a finding subsequently supported and extended (Grant, Walden and Seitz, 1998; Ma *et al.*, 2009; Ross *et al.*, 2007). For various tasks, decreased reaction times have been observed for congruent face and voice information, while increased reaction times are found for incongruent stimuli (e.g., Besle *et al.*, 2004; Latinus, VanRullen and Taylor, 2010). A clear example of face–voice interaction at the behavioural level is the McGurk-effect, which is generally cited as an example of visual speech interfering with auditory speech to produce an illusory percept (McGurk and MacDonald, 1976; Tiippana *et al.*, 2011; van Wassenhove and Nagarajan, 2007); recent work however shows that this interference is actually bidirectional (Baart and Vroomen, 2010). It is clear from the behavioural evidence that during perception there is interaction between the auditory and visual information provided by the face and voice. Such insights, along with considerable physiological evidence of multisensory integration at the neuronal level (for a recent review, see Stein and Stanford, 2008), have led many to investigate the cerebral mechanisms of face–voice integration (reviewed in Campanella and Belin, 2007).

There has been much discussion however around the pros and cons of the statistical criteria used to classify multisensory integration when comparing bimodal to unimodal conditions using fMRI (Beauchamp, 2005; Calvert, 2001; Goebel and van Atteveldt, 2009; Laurienti *et al.*, 2005; Stein *et al.*, 2009). The three main criteria used in fMRI research are: (1) the super-additive criterion, which requires the bimodal response to be greater than the sum of both unimodal responses; (2) the max criterion that requires the bimodal response to be greater than the largest unimodal response; and (3) the mean criterion requiring the bimodal response to be greater than the mean of the unimodal responses. Under the super-additive criterion, portions of the, temporal, occipital, parietal and frontal lobes have all been proposed as part of a face–voice integration network. Two recent fMRI studies, for example, report responses located in subregions of all these lobes to be higher for cross-modal speech than the sum of both unimodal responses (Joassin, Maurage and Campanella, 2011a; Joassin *et al.*, 2011b). Similarly, Calvert *et al.* (1999) reported enhanced activity in regions of the temporal and occipital lobes for audiovisual speech perception relative to perceiving each cue in isolation. In a follow up study, this group also reported super-additive responses in the temporal, occipital, parietal and frontal lobes, whilst focusing their discussion on left posterior superior temporal sulcus, as it also displayed a congruency effect (Calvert, Campbell and Brammer, 2000). Using the max criterion, others (Kreifelts *et al.*, 2007; Szycik, Tausche and Münte, 2008; Wright *et al.*, 2003) have found bilateral superior temporal cortex

(STC) to be loci of face–voice integration. We are not aware of any fMRI studies that have used the mean criterion to implicate brain regions as sites of integration for face and voice; however, for non-speech stimuli it has been used to classify areas of STC as multisensory (e.g., Beauchamp *et al.*, 2004).

The three criteria outlined above are frequently used to identify loci of multisensory integration in fMRI; yet, few studies directly compare those criteria within the same experiment (Beauchamp, 2005; Brefczynski-Lewis *et al.*, 2009). However, it is also noteworthy that they are not the only statistical criteria used, particularly in neurophysiological studies. For example, sub-additivity, in which the audiovisual response is less than the sum of the unimodal responses, reflects multisensory integration in single-neuron recordings (e.g., Perrault *et al.*, 2005; Stanford, Quessy and Stein, 2005). In regard to audiovisual speech research using fMRI, sub-additivity has been interpreted as representative of multisensory inhibition produced by incongruent stimuli (Calvert, Campbell and Brammer, 2000). As we do not manipulate stimulus congruence we do not test for sub-additivity in this experiment; nevertheless, using a single dataset, we do test three statistical criteria (super-additivity, max and mean criteria) frequently used in fMRI research.

Experimental methods also exist that can either circumvent the need for these statistical criteria of integration or can be used in conjunction with them. For example, manipulations of the congruency (e.g., Calvert *et al.*, 2001; Szycik, Tausche and Münte, 2008) and signal strength of stimulus cues (Stevenson and James, 2009) have implicated STC as a site of audiovisual speech integration. Analysing the connectivity between regions found involved in the integration of face and voice has also helped to understand the cerebral mechanisms involved (e.g., Nath and Beauchamp, 2011; Noppeney *et al.*, 2008).

In the current fMRI study, we further investigate the cerebral mechanisms of face–voice integration by presenting participants with either unimodal or bimodal speech stimuli. Using the same data set, we examine the influence of using different statistical criteria on which regions are classified as integrating face and voice information (Beauchamp, 2005). Unlike previous work with similar speech stimuli (e.g., Calvert, Campbell and Brammer, 2000; Joassin, Maurage and Campanella, 2011a; Joassin *et al.*, 2011b), when comparing unimodal to bimodal speech perception we present the results for all of the three main statistical criteria. We also used psychophysiological interaction (PPI) analysis (Friston *et al.*, 1997) to add to the growing evidence on the connectivity between regions involved in audiovisual face–voice perception (e.g., Joassin, Maurage and Campanella, 2011a; Joassin *et al.*, 2011b; Kreifelts *et al.*, 2007; Nath and Beauchamp, 2011; von Kriegstein *et al.*, 2005). Our design enables us to investigate which regions are generally in-

volved in unimodal speech perception and in particular to confirm whether visual speech cues alone, make use of areas in temporal cortex generally regarded as auditory regions (e.g., Puce *et al.*, 1998).

2. Materials and Methods

2.1. Participants

Twenty right-handed native English speakers (10 female, age range = 20–30, mean = 24) participated in the study. All participants had normal or corrected to normal vision and reported having no hearing difficulties or any history of neurological disorders. The experiment was approved by the University of Glasgow ethics committee and participants gave informed written consent and were paid for participation.

2.2. Stimuli

Stimuli were dynamic audiovisual movies (25 frames per second) of either, a native English speaker saying 'tomorrow' or a native Italian speaker saying 'domani' (which is tomorrow in Italian). The visual component contained the full face and

covered a visual angle of 22° in height and 15° in width (Fig. 1). Total duration of each word stimulus was 1.6 s, which included 360 ms of fade-in and fade-out. For baseline a black background with a central white fixation cross was used.

Stimuli were presented using Matlab 2007b (Mathworks Inc., Natick, MA) and the Psychophysics Toolbox (PTB3) extensions (Brainard, 1997; Pelli, 1997) running on a PC. The auditory stimulus cue was presented *via* NordicNeuroLab electrostatic headphones at approximately 90 dB: a compromise between a sound level loud enough to exceed the scanner noise and one relatively comfortable for participants. The visual cue was displayed through NordicNeuroLab VisualSystem goggles.

2.3. Procedure

Speech stimuli were presented in one of three stimulus conditions, audio alone (A), video alone (V) or audiovisual (AV), while blood oxygenation level-dependent (BOLD) signal was measured in the fMRI scanner. Each condition was presented

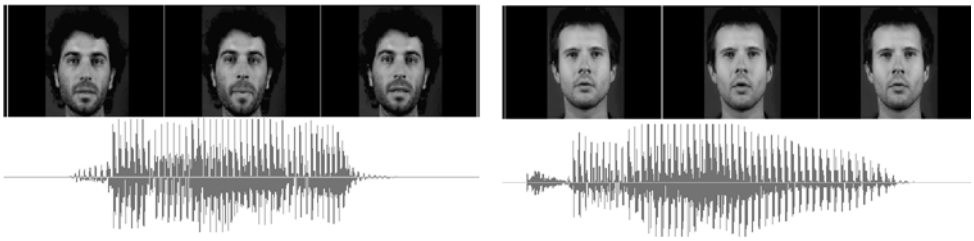

Figure 1. Stimulus illustration. (Left) Italian actor. (Right) British actor. (Top row) Three frames of each movie. (Bottom row) Waveforms for the word ‘domani’ (left) and ‘tomorrow’ (right).

separately within a block-design functional run (~11 min). Stimulation blocks, six for each condition, lasted for 16 s (5 repetitions of each nationality) and after every stimulus block there was an 18 s fixation block. The order of blocks was chosen, separately for each participant, by randomising all six possible orderings of, A, V and AV. At the start of the run there was a 12.5 s fixation period. During stimulus presentation participants had to respond whether the speaker was native Italian or native English, using the index or middle finger of their right hand.

2.4. Imaging Parameters and Analysis

Functional images covering the whole brain (slices = 32, field of view = 210 × 210 mm, voxel size = 3 × 3 × 3 mm) were acquired on a 3T Tim Trio Scanner (Siemens) using an echoplanar imaging (EPI) sequence (interleaved, TR = 2 s, TE = 30 ms, Flip Angle = 80°). At the end of each fMRI session, high resolution T1-weighted images (anatomical scan) were obtained (slices = 192, field of view = 256 mm, voxel size = 1 × 1 × 1 mm, Flip angle = 9°, TR = 1.9 s, TE = 2.52 ms).

SPM8 software (Wellcome Department of Imaging Neuroscience, London, UK) was used to pre-process and analyse the imaging data. First, the anatomical scan was AC-PC centred; this correction was then applied to all the EPI volumes. Functional data were motion corrected using a two-pass six-parameter rigid-body spatial transformation (Friston *et al.*, 1996), which realigned all functional volumes to the first volume of the run and subsequently realigned the volumes to the mean volume. The anatomical scan was co-registered to the mean volume and segmented. The anatomical and functional images were then normalised to the Montréal Neuro-
logical Institute (MNI) template using the parameters issued from the segmentation keeping the voxel resolution of the original scans (1 × 1 × 1 and 3 × 3 × 3, respectively). Functional images were then smoothed with a Gaussian function with a full-width at half maximum of 10 × 10 × 10 mm. Global linear trends were minimised through high-pass filtering the data with a cutoff period of 128 s during
statistical model estimation. All analysis was conducted in a masked skull-stripped search volume, created by combining three matter types (white, grey and CSF) output during the segmentation procedure.

Functional data were analysed in a two-level random-effects design. The firstlevel, fixed effects individual participant analysis involved a design matrix containing a separate regressor for each stimulus condition, which were entered in

the order A, V and AV. These regressors contained boxcar functions representing the onset and offset of stimulation blocks convolved with a canonical hemodynamic response function. To account for residual motion artefacts the realignment parameters were also added as nuisance covariates to the design matrix. Using the modified general linear model parameter estimates for each condition at each voxel were calculated and then used to create contrast images for a condition relative to fixation: A > Fix, V > Fix and AV > Fix. These three contrast images, from each participant, were taken forward into the second-level full factorial ANOVA.

This random-effects (RFX) analysis allows inferences to be made at the population level (Friston *et al.*, 1999). The estimated full factorial model was used to create group-level RFX contrast images for factors of interest. To help clarify exactly what criteria of multisensory integration were used we describe all contrasts
of interest with the contrast vector used to create it in SPM (see also design matrix insets in Fig. 2). Stimulus condition effects were tested with, A > Fix ([1 0 0]) for voices, V > Fix ([0 1 0]) for faces and AV > Fix ([0 0 1]) for cross-modal face–voice. As there is no task in the fixation condition participants did not need to make a response, unlike in the stimulus conditions. Therefore, the cerebral activity related to the response is not subtracted out in these stimulus condition contrasts. To help elucidate and remove regions found significant in these contrasts due to their involvement in planning and execution of the response we also tested for regions displaying more activity to one unimodal condition relative to the other using
A > V ([1 −1 0]) and V > A ([−1 1 0]), which subtracts out the response component. To examine super-additive effects we tested for regions displaying more
activity to the audiovisual face–voice condition than to the sum of the unimodal conditions (AV > A + V, [−1 −1 1]). We used a conjunction analysis to test for regions meeting the max criterion (AV > A ∩ AV > V, [−1 0 1] ∩ [0 −1 1]). Regions meeting the mean criterion (AV > mean[A, V]) were found using the contrast [−1 −1 2] and we also tested for those responding significantly to both unimodal conditions (A > Fix ∩ V > Fix, [1 0 0] ∩ [0 1 0]).

Connectivity between regions during audiovisual speech perception was investigated by modeling psychophysiological interactions (Friston *et al.*, 1997). PPI analysis defines regions that are differentially influenced by the interaction between the response of another (seed) region and a change in experimental factor. We investigated the connectivity of regions found significant in all of the three criteria by conducting a separate PPI analysis for each region/criterion combination. Time-courses of volumes of interest (VOI) were derived by extracting the first eigenvariate of a 6 mm sphere centered on the peak-voxel of a region of interest (ROI), defined at the group level. During extraction the time-courses were adjusted for the effect of interest (omnibus F-test of all conditions). The PPI models contained three regressors: the physiological regressor, which was a deconvolved VOI time-course (Gitelman, 2003); the psychological variable regressor representing the
change in experimental factor (e.g., AV > A + V), and the psychophysiological

interaction regressor which is the product of the first two regressors. Similar to the
GLM analysis described above, PPI was first conducted at the individual level before testing an RFX group analysis. The PPI analysis for the max criterion, which involves a conjunction, was achieved by running separate PPI models for each of AV > A and AV > V before using the results of both in a full factorial RFX group analysis to enable the connectivity conjunction (AV > A ∩ AV > V).

For all contrasts and PPI analysis, unless otherwise stated, we report voxels reaching a significance level of $p < 0.05$ with a family wise error (FWE) correction to control for multiple comparisons. Labelling of significant regions followed the automatic anatomical labelling convention (Tzourio-Mazoyer *et al.*, 2002). In result tables the coordinates of the peak voxel within significant clusters were used to define the anatomical location.

3. Results

3.1. Unimodal and Bimodal Face–Voice Processing

Regions activating more to auditory speech than the fixation condition were bilateral STG, bilateral precentral gyrus, left cerebellum, left supplementary motor area and left pallidum (Table 1(a)). When testing for regions that respond more to auditory than visual speech, again bilateral STG was significant as was right precuneus and left superior parietal gyrus (Table 1(c) and Fig. 2(A)). However, the response profiles (Fig. 2(A, right panel)) of both the precuneus and superior parietal gyrus indicated that the significant result was driven by a deactivation to visual stimulation compared to fixation rather than increased activity to auditory stimulation.

Regions activating more to visual speech than the fixation condition were bilateral occipital cortex, bilateral posterior STG, left postcentral gyrus, right precentral gyrus, bilateral thalamus, right inferior frontal gyrus, left putamen and left supplementary motor area (Table 1(b)). When testing for regions that respond more to visual than auditory speech, again bilateral occipital cortex and thalamus were significant as was an orbital portion of right middle frontal gyrus (Table 1(d) and Fig. 2(A)). The response profiles of all regions found to respond more to visual than auditory speech displayed more activation for visual than fixation conditions (although not significantly more for the orbital part of middle frontal gyrus) and marginally deactivated for auditory compared to fixation conditions (Fig. 2(A, left panel)).

Regions activating more to audiovisual speech than fixation were bilateral occipital and temporal cortex, left postcentral gyrus, right inferior frontal gyrus, left middle frontal gyrus, left supplementary motor area, right superior frontal gyrus and right putamen (Table 1(e)).

3.2. Face–Voice Integration Criteria

Super-additive responses were found in regions of both left and right occipital cortex and in right precentral gyrus (Table 2(a) and Fig. 2(B)). However, the response profiles of the occipital regions indicate that the significant super-additive result was actually driven by a deactivation to auditory stimuli relative to baseline. In the precentral gyrus, there was actually deactivation for all conditions.

Using FWE correction at a significance level of $p < 0.05$, no regions were found to meet the max criterion. At an uncorrected significance level of $p < 0.001$ and a minimum cluster size of 10 voxels only the left hippocampus passed the max criterion (Table 2(b) and Fig. 2(C)) and examination of the response profile supported this finding.

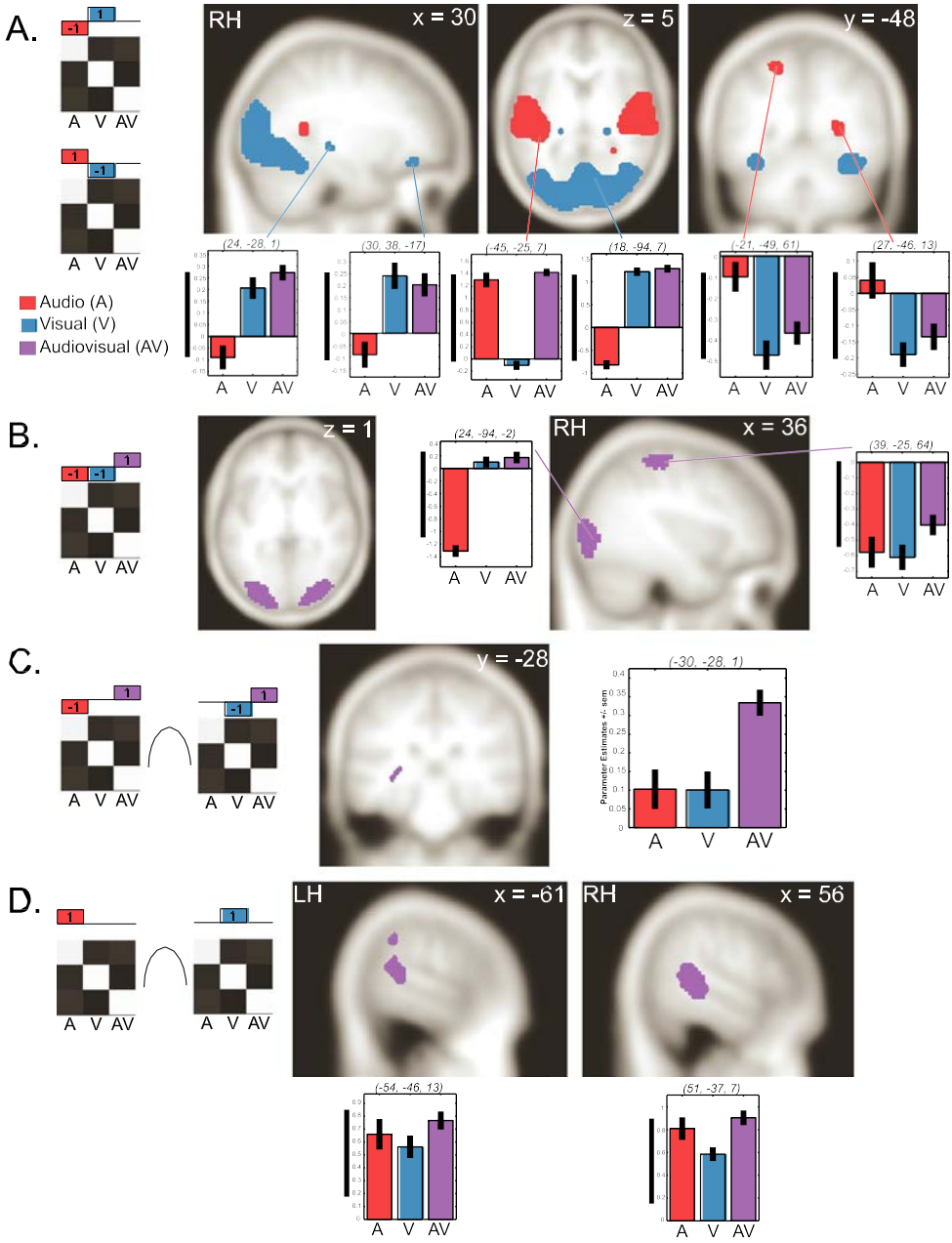

Figure 2. Significant activations from statistical contrasts. (A) Significant activations to unimodal conditions. In blue: response to visual stimuli greater than to auditory. In red: response to auditory stimuli greater than to visual. (B) Significant activations for the super-additive contrast. (C) Significant activations found using the max criterion; max criterion was defined by using a conjunction between audiovisual greater than audio and audiovisual greater than visual (see design matrix on the left; $p < 0.001$ uncorrected). (D) Significant activations for the conjunction of audio greater than fixation and visual greater than fixation. All bar graphs display GLM parameter estimates for each of the 3 stimulus conditions.

Table 1. Results of independently contrasting unimodal (a and b) and cross-modal (e) conditions against fixation and directly contrasting audio and visual unimodal conditions (c and d). Contrasts were height thresholded ($t(57) = 4.682$) to display voxels reaching a significance level of $p < 0.05$ with FWE correction and an additional minimum cluster size of 10 contiguous voxels. MNI coordinates and t-scores are from the peak voxel of a region

Contrast: Region	Hemisphere	MNI coordinates (x, y, z)	Cluster size (voxels)	t
(a) A > Fix:				
Superior temporal gyrus	Left	(−45, −25, 7)	1463	14.96
Superior temporal gyrus	Right	(51, 16, 1)	1638	14.87
Precentral gyrus	Left	(−42, −19, 55)	302	9.23
Cerebellum	Left	(21, −52, −23)	114	8.97
Supplementary motor area	Left	(−6, −1, 58)	138	6.93
Pallidum	Left	(−24, −1, −2)	17	5.20
Precentral gyrus	Right	(48, −1, 46)	12	5.14
(b) V > Fix:				
Superior occipital gyrus	Left	(−9, −94, 4)	3382	22.16
Postcentral gyrus	Left	(−45, −19, 55)	272	8.33
pSuperior temporal gyrus	Right	(51, −37, 7)	323	8.20
Precentral gyrus	Right	(51, 2, 49)	248	6.54
pSuperior temporal gyrus	Left	(−54, −46, 13)	177	6.50
Thalamus	Left	(−21, −25, −2)	40	6.45
Thalamus	Right	(24, −25, −2)	20	6.17
Inferior frontal gyrus, orbital	Right	(42, 44, −14)	72	6.10
Putamen	Left	(−27, −1, −2)	81	5.96
Supplementary motor area	Left	(−6, 2, 58)	40	5.63
(c) A > V:				
Superior temporal gyrus	Right	(51, −13, 1)	1103	15.97
Superior temporal gyrus	Left	(−45, −25, 7)	953	14.45
Precuneus	Right	(27, −46, 13)	27	5.67
Superior parietal gyrus	Left	(−21, −49, 61)	25	5.29
(d) V > A:				
Superior occipital gyrus	Left	(−9, −97, 4)	3880	22.81
Cuneus	Right	(18, −94, 7)		
Thalamus	Left		50	7.68
Thalamus	Right	(21, −28, 1)	49	7.66
Middle frontal gyrus, orbital	Right	(30, 38, −17)	16	5.63
(e) AV > Fix:				
Superior occipital gyrus	Left	(−9, −94, 4)	3678	22.65
Superior temporal gyrus	Left	(−45, −25, 7)	1950	16.53
Superior temporal gyrus	Right	(51, −16, 1)	1537	15.14
Postcentral gyrus	Left	(−45, −19, 58)	329	10.42
Inferior frontal gyrus, orbital	Right	(42, 44, −14)	151	6.4
Middle frontal gyrus	Left	(39, 14, 55)	66	5.86
Supplementary motor area	Left	(−6, 2, 55)	32	5.67
Superior frontal gyrus, medial	Right	(12, 56, 34)	13	5.19
Putamen	Right	(27, 5, −2)	13	4.84

Table 2.
Regions of face–voice integration according to: (a) super-additive criterion, (b) max criterion at a significance level of $p < 0.001$ uncorrected, (c) mean criterion, (d) displays regions found to significantly activate to both unimodal conditions. MNI coordinates and t-scores are from the peak voxel of a region

Contrast: Region	Hemisphere	MNI coordinates (x, y, z)	Cluster size (voxels)	t
(a) AV > (A + V):				
Inferior occipital gyrus	Right	(24, −94, −2)	802	11.15
Middle occipital gyrus	Left	(−30, −91, −5)	747	9.43
Precentral gyrus	Right	(39, −25, 64)	590	7.24
(b) (AV > A) ∩ (AV > V) $p < 0.001$:				
Hippocampus	Left	(−30, −28, 1)	14	4.49
(c) AV > mean(A,V):				
Superior occipital gyrus	Left	(−9, −97, 4)	3162	14.93
Cuneus	Right	(18, −94, 7)		
Superior temporal gyrus	Left		730	11.06
Thalamus	Left	(−21, −28, −2)		
Superior temporal gyrus	Right		558	10.3
Thalamus	Right	(21, −28, −2)		
(d) A > Fix ∩ V > Fix:				
Cerebellum	Right	(21, −52, −23)	114	8.97
Postcentral gyrus	Left	(−45, −19, 55)	222	8.33
Superior temporal gyrus	Right	(51, −37, 7)	317	8.20
Superior temporal gyrus	Left	(−54, −46, 13)	113	6.50
Supplementary motor area	Left	(−6, 2, 58)	39	5.63
Pallidum	Left	(−24, −1, −2)	17	5.20

Bilateral occipital cortex, bilateral STC and bilateral thalamus all passed the mean criterion. The response profiles of the occipital regions and thalamus indicated that the bimodal response was similar to the unimodal visual response while the mean was lower due to the deactivation during the auditory condition. The opposite situation was found in bilateral STC. Note that we did not display the regions or response profiles meeting the mean criterion because all were the same as or largely overlap regions displayed in unimodal contrasts (Fig. 2(A)).

Also, a conjunction analysis revealed, the right cerebellum, left postcentral gyrus, bilateral posterior STC and left pallidum as regions responding significantly more to both unimodal conditions than fixation (Table 2(d) and Fig. 2(D)).

3.3. Connectivity Analysis

PPI analysis using regions found significant in the super-additive contrast as seed regions and this contrast [AV > (A + V)] as the psychological factor of interest highlighted an increased connectivity between both the right inferior occipital gyrus

and left middle occipital gyrus and bilateral STC (Table 3). No regions were found

Table 3. Results of PPI analysis, outlining regions with enhanced connectivity with seed regions from the super-additive and mean criteria contrasts. No regions showed enhanced connectivity in the PPI analysis using the max criterion

PPI seed: Region	Hemisphere	MNI coordinates (*x*, *y*, *z*)	Cluster size (voxels)	*t*
Super-additive contrast as psychological variable				
Right inferior occipital gyrus [24, −94, −2]				
Superior temporal gyrus	Left	(−48, −19, −2)	550	7.95
Superior temporal gyrus	Right	(52, −10, 1)	402	7.84
Left middle occipital gyrus [−30, −91, −5]				
Superior temporal gyrus	Right	(48, −13, 1)	933	10.86
Superior temporal gyrus	Left	(−48, −25, −4)	1046	10.81
Mean criteria contrast as psychological variable				
Left superior occipital gyrus [−9, −97, 4]				
Superior temporal gyrus	Right	(54, −16, 1)	933	8.46
Superior temporal gyrus	Left	(−60, −13, 10)	563	8.21
Left superior temporal gyrus [−45, −25, 7]				
Calcarine sulcus	Right	(21, −91, 4)	1641	11.75
Lingual gyrus	Left	(−12, −85, −5)		
Postcentral gyrus	Left		11	6.77
Right superior temporal gyrus [51, −13, 1]				
Fusiform gyrus	Right	(33, −67, −11)	1418	10.05
Lingual gyrus	Left	(−12, −85, −8)		
Superior temporal gyrus	Right		18	7.78
Hippocampus	Left	(−21, −28, −5)	12	7.41
Postcentral gyrus	Left	(−45, −19, 55)	20	7.06
Superior temporal gyrus	Right	(54, −37, −7)	25	6.94
Superior temporal gyrus	Left	(−57, −37, −7)	11	6.85

to have increased connectivity with the right precentral gyrus in bimodal relative to unimodal conditions. PPI analysis based on the max criterion found no regions with significantly increased connectivity to the left hippocampus, even at a threshold of $p < 0.001$ uncorrected. That is, no regions showed enhanced connectivity to audiovisual conditions compared to both unimodal conditions. Using significant regions from the mean criterion as seeds and this contrast [AV > mean(A, V)] as the psychological factor of interest, mainly highlighted increased connectivity between superior temporal and occipital regions for audiovisual conditions compared to the mean of the unimodal conditions (Table 3).

4. Discussion

Using a single dataset and ecological stimuli, dynamic movies of audiovisual speech, we have shown that the super-additive, max and mean criteria of multisensory integration revealed different loci of audiovisual speech integration.

The

super-additive and mean criteria revealed mostly 'sensory-specific' regions, similar to those observed in unimodal contrasts. The max criterion appeared the most stringent, highlighting only the left hippocampus.

4.1. Unimodal Face and Voice Perception

In line with previous work (reviewed in Belin, Fecteau and Bédard, 2004; Hickok and Poeppel, 2000; Scott and Johnsrude, 2003), perceiving speech from auditory cues of the voice, involved bilateral temporal cortex. While testing for loci of auditory speech perception, the importance of exploring response profiles in fMRI research was further highlighted (Beauchamp, 2005; Goebel and van Atteveldt, 2009). Without examination of the response profiles of the precuneus and superior parietal cortex these regions would also have been classified as voice processing areas. Both regions were found significant in a contrast (A > V) designed to define voice processing areas; however, neither activated more to unimodal auditory speech than baseline, hence it would be wrong to classify them as involved in perceiving auditory speech (Fig. 2(A, right panel)).

Perceiving speech from visual face cues involved bilateral occipital cortex, bilateral thalamus and an orbital part of right middle frontal gyrus; all have previously been implicated in face perception as well as visual perception in general (Hole and Bourne, 2010). Also, and in support of previous findings (Bernstein *et al.*, 2002; Calvert, 1997; Olson, Gatenby and Gore, 2002; Puce *et al.*, 1998; Wright *et al.*, 2003), bilateral STC responded to the articulating mouth movements of speech without any auditory stimulation. There is some debate as to whether activations in STC caused by lipreading extend into primary auditory cortex or not (Calvert, 1997; Bernstein *et al.*, 2002). The STC activation found in the current study is in a posterior portion of STC and thus not believed to be overlapping with primary auditory cortex. However, our experimental design was not optimised to examine this question and our analysis did not make use of defining primary auditory cortex individually (Pekkola *et al.*, 2005) hence, we cannot rule out the possibility that it is activated by lipreading.

4.2. Bimodal Face–Voice Perception

Audiovisual perception of speech mainly involved the occipital and temporal areas that were also activated during unimodal face and voice conditions, respectively. The super-additivity criterion, which is commonly used to highlight loci of multisensory face–voice integration (e.g., Calvert, Campbell and Brammer, 2000; Joassin, Maurage and Campanella, 2011a; Joassin *et al.*, 2011b) was met by bilateral occipital regions and right precentral gyrus in the current study. It was clear that the significant super-additive effect was driven by the audiovisual speech condition being contrasted to the sum of a positive visual response and a large negative auditory response. In Fig. 3 of both Joassin, Maurage and Campanella

(2011a) and Joassin *et al.* (2011b) the authors also highlight that their super-additive effects in occipital and temporal cortex are the result of the bimodal response being compared to the sum of a positive and a negative unimodal response, which they nevertheless interpret as multisensory integration. However, the interpretation of this situation is complicated and it remains an open question whether we can really infer integration from this type of response profile (Calvert *et al.*, 2001; Goebel and van Atteveldt, 2009). The super-additive criterion is often described as the strictest of the multisensory integration criteria. However, this is only true when the implementation of it is restricted to brain regions showing increased activity for both unimodal conditions relative to baseline. Otherwise, 'sensory-specific' cortices, which deactivate to stimulation of other senses, are likely to be categorised as super-additive and multisensory in nature (Goebel and van Atteveldt, 2009). To our knowledge there is only one published result that meets this restricted super-additive response with nondegraded audiovisual speech stimuli. Calvert, Campbell and Brammer (2000) found a region (8 voxels) of the left superior temporal sulcus that was activated by both unimodal visual and auditory speech and satisfied the super-additivity criterion. Here we did not find such a region and can only speculate as to some of the possible reasons. Calvert, Campbell and Brammer (2000) presented the bottom half of the face in their stimuli, while we presented full face. It is possible that due to stimulus effectiveness, full faces produce a stronger signal in this region, which results in unimodal saturation of the BOLD signal. Support for the idea that stimulus factors play a role in response amplitude in this region comes from a study in which a similar area was found to respond more to dynamic than static faces (Campbell *et al.*, 2001). The only region of the brain that met the max criterion was the left hippocampus, albeit only when using a less conservative significance level. This adds some support to the proposal of Joassin *et al.* (2011b), that the hippocampus is a key region in the integration of faces and voices. Also using the max criteria, Szycik, Tausche and Münte (2008) found bilateral superior temporal sulcus to be involved in face–voice integration. Two possible reasons why we do not find temporal cortex to pass the max criterion while they do are stimulus related. First, they present a static face in their unimodal auditory condition and second, they add white noise to their auditory and audiovisual stimuli. We incorporate neither of these factors into our stimuli and the lack of auditory noise in our stimuli, in particular, could have played a crucial role in the difference between the two studies. Lowering the signal-to-noise ratio of stimuli can help to prevent multisensory integration effects being missed due to saturation of the BOLD signal from at least one unimodal conditions (Goebel and van Atteveldt, 2009). Preventing BOLD saturation to enable larger and more detectable multisensory interactions is a very similar concept to the principle of inverse effectiveness described at the neuronal level (Stein and Meredith, 1993). Making use of these concepts, (Stevenson, Geoghegan and James, 2007) highlighted the usefulness of presenting stimuli at

threshold level, in enabling audiovisual super-additive effects to be found in STC. The same group strongly emphasised the advantage of using this technique by parametrically mapping out, using different signal-to-noise ratios, the change from not being able to find super-additive effects to doing so for
both speech and non-speech stimuli (Stevenson and James, 2009).

Using the mean criterion, in the current study, to define regions as integrating faces and voices would implicate the occipital and temporal regions, which were already found to process the unimodal visual and auditory stimulation. Examination of response profiles from these regions shows almost no difference between the response to the combined face–voice and the 'sensory-specific' unimodal response of the region. As pointed out by Goebel and van Atteveldt (2009), the mean criterion, similar to super-additive, is inclined to classify 'sensory-specific' regions of the brain as multisensory due to the reduction of the 'sensory-specific' response in the mean calculation.

It is clear that the choice of statistical criteria has a large impact on which regions are found to be involved in face–voice integration using fMRI (Table 2 and Fig. 2). Goebel and van Atteveldt (2009) provide extensive discussion of the relative merits of each criterion of multisensory integration and conclude that they all have limitations in fMRI research. The fact that our super-additive effects were the result of summing negative and positive unimodal responses and that we may have failed to replicate integration effects based on the max criterion due to BOLD saturation further emphasize these limitations using a single data set. Moreover, it has been argued that there has been an overemphasis on super-additivity as being the litmus test for multisensory integration and that a failure to explore other criteria could have a detrimental effect on our understanding of integration mechanisms (Stanford and Stein, 2007). Therefore, multisensory research using fMRI would benefit from exploring several integration criteria in the same experiment as was done here. Furthermore, combining them with other experimental manipulations (e.g., congruency and signal-to-noise ratio) would be instrumental in enabling strong conclusions about the occurrence of multisensory integration in a particular region.

Our connectivity analysis revealed increased connectivity between occipital and temporal regions for bimodal stimulation relative to unimodal conditions. The existence of such connectivity and its increase in bimodal situations is generally interpreted as providing a mechanism of multisensory integration (e.g., Joassin, Maurage and Campanella, 2011a; Joassin *et al.*, 2011b). However, in these 'sensory-specific' temporal regions the general response to unimodal visual stimulation, in the current study, was a deactivation relative to baseline. Similarly, using non-speech stimuli, Laurienti *et al.* (2002) highlighted deactivations in auditory (temporal) cortex during unimodal visual presentation and also in visual (occipital) cortex during auditory stimulation. These 'cross-modal inhibitory processes' were described, by the authors, as being 'switched off' during audiovisual stimulation, in which the bimodal response was as large as the 'sensory-specific' unimodal response. Hence, another interpretation of increased connectivity, is that it reflects the addition of this 'switching off' process. However, this speculation requires direct empirical testing.

The results presented here and the discussion of the literature suggest that comparing bimodal to unimodal stimulus conditions using the, super-additive, max or

mean criteria of multisensory integration is not the best way to uncover loci of face–voice integration. Although they are all valid approaches and provide important information, much care has to be taken when interpreting the results (Beauchamp, 2005; Calvert and Thesen, 2004; Goebel and van Atteveldt, 2009). As discussed above, using a combination of these criteria alongside manipulations of cross-modal congruency and/or the signal-to-noise ratio of unimodal conditions may prove to be a more cogent method of investigating the cerebral correlates of face–voice integration with fMRI.

Acknowledgements

S. L. was supported by the Economic and Social Research Council. M. L. was supported by an Economic and Social Research Council/Medical Research Council Grant RES-060-25-0010. We also thank two anonymous reviewers for their constructive and helpful comments.

References

Baart, M. and Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading, *Neuroscience Letters* **471**, 100–103.

Beauchamp, M. S. (2005). Statistical criteria in fMRI studies of multisensory integration, *Neuroinformatics* **3**, 93–114.

Beauchamp, M. S., Lee, K. E., Argall, B. D. and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus, *Neuron* **41**, 809–823.

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

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C. W., Don, M. and Singh, M. (2002). Visual speech perception without primary auditory cortex activation, *Neuroreport* **13**, 311–315.

Besle, J., Fort, A., Delpuech, C. and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex, *Eur. J. Neurosci.* **20**, 2225–2234.

Brainard, D. H. (1997). The psychophysics toolbox, *Spatial Vis.* **10**, 433–436.

Brefczynski-Lewis, J., Lowitszch, S., Parsons, M., Lemieux, S. and Puce, A. (2009). Audiovisual nonverbal dynamic faces elicit converging fMRI and ERP responses, *Brain Topography* **21**, 193–206.

Calvert, G. A. (1997). Activation of auditory cortex during silent lipreading, *Science* **276**, 593–596.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies, *Cerebral Cortex* **11**, 1110–1123.

Calvert, G. A. and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain, *J. Physiol. Paris* **98**, 191–205.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D. and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding, *NeuroReport* **10**, 2619–2623.

Calvert, G. A., Campbell, R. and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex, *Current Biology* **10**, 649–657.

Calvert, G. A., Hansen, P. C., Iversen, S. D. and 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. and Belin, P. (2007). Integrating face and voice in person perception, *Trends in Cognitive Sciences* **11**, 535–543.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., McGuire, P., Suckling, J., Brammer, M. J. and David, A. S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning), *Cognitive Brain Res.* **12**, 233–243.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. and Turner, R. (1996). Movement-related effects in fMRI time-series, *Magn. Reson. Med.* **35**, 346–355.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E. and Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging, *NeuroImage* **6**, 218–229.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C. and Worsley, K. J. (1999). Multisubject fMRI studies and conjunction analyses, *NeuroImage* **10**, 385–396.

Gitelman, D. (2003). Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution, *NeuroImage* **19**, 200–207.

Goebel, R. and van Atteveldt, N. (2009). Multisensory functional magnetic resonance imaging: a future perspective, *Exper. Brain Res.* **198**, 153–164.

Grant, K. W., Walden, B. E. and Seitz, P. F. (1998). Auditory-visual speech recognition by hearingimpaired subjects: consonant recognition, sentence recognition, and auditory-visual integration, *J. Acoust. Soc. Amer.* **103**, 2677–2690.

Hickok, G. and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception, *Trends in Cognitive Sciences* **4**, 131–138.

Hole, G. and Bourne, V. (2010). *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*. Oxford University Press, Oxford.

Joassin, F., Maurage, P. and Campanella, S. (2011a). The neural network sustaining the crossmodal processing of human gender from faces and voices: an fMRI study, *NeuroImage* **54**, 1654–1661.

Joassin, F., Pesenti, M., Maurage, P., Verreckt, E., Bruyer, R. and Campanella, S. (2011b). Crossmodal interactions between human faces and voices involved in person recognition, *Cortex* **47**, 367–376.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M. and Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study, *NeuroImage* **37**, 1445–1456.

Latinus, M. and Belin, P. (2011). Human voice perception, *Current Biology* **21**, R1–R3.

Latinus, M., VanRullen, R. and Taylor, M. J. (2010). Top–down and bottom–up modulation in processing bimodal face/voice stimuli, *BMC Neurosci.* **11**, 36.

Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S. and Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli, *J. Cognitive Neurosci.* **14**, 420–429.

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

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J. and Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space, *PloS One* **4**, e4638.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 746–748.

Nath, A. R. and Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech, *J. Neurosci.* **31**, 1704–1714.

Noppeney, U., Josephs, O., Hocking, J., Price, C. J. and Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds, *Cerebral Cortex* **18**, 598–609.

Olson, I. R., Gatenby, J. C. and Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex, *Cognitive Brain Res.* **14**, 129–138.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A. and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T, *Neuroreport* **16**, 125–128.

Pelli, D. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies, *Spatial Vis.* **10**, 437–442.

Perrault, T. J., Vaughan, J. W., Stein, B. E. and Wallace, M. T. (2005). Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli, *J. Neurophysiol.* **93**, 2575–2586.

Puce, A., Allison, T., Bentin, S., Gore, J. C. and McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements, *J. Neurosci.* **18**, 2188–2199.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, *Cerebral Cortex* **17**, 1147–1153.

Scott, S. K. and Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception, *Trends in Neurosciences* **26**, 100–107.

Stanford, T. R., Quessy, S. and Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus, *J. Neurosci.* **25**, 6499–6508.

Stanford, T. R. and Stein, B. E. (2007). Superadditivity in multisensory integration: putting the computation in context, *Neuroreport* **18**, 787–792.

Stein, B. E. and Meredith, M. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA.

Stein, B. E. and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron, *Nature Rev. Neurosci.* **9**, 255–266.

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J. and Rowland, B. A. (2009). Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness, *Exper. Brain Res.* **198**, 113–126.

Stevenson, R. A., Geoghegan, M. L. and James, T. W. (2007). Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects, *Exper. Brain Res.* **179**, 85–95.

Stevenson, R. A. and James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition, *NeuroImage* **44**, 1210–1023.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Amer.* **26**, 212–215.

Szycik, G. R., Tausche, P. and Münte, T. F. (2008). A novel approach to study audiovisual integration in speech perception: localizer fMRI and sparse sampling, *Brain Res.* **1220**, 142–149.

Tiippana, K., Puharinen, H., Möttönen, R. and Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated, *Seeing and Perceiving* **24**, 67–90.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, *NeuroImage* **15**, 273–289.

van Wassenhove, V. and Nagarajan, S. S. (2007). Auditory cortical plasticity in learning to discriminate modulation rate, *J. Neurosci.* **27**, 2663–2672.

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

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J. and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech, *Cerebral Cortex* **13**,

1034–1043.