Brain & Language 126 (2013) 253–262

Contents lists available at SciVerse ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Neural correlates of audiovisual speech processing in a second language

Alfonso Barrós-Loscertales [a,1], Noelia Ventura-Campos [a,1], Maya Visser [a,b,*,1], Agnès Alsius [c], Christophe Pallier [d], César Ávila Rivera [a], Salvador Soto-Faraco [b,e]

[a] Dept. de Psicología Básica, Clínica y Psicobiología, Universitat Jaume I, Castellón, Spain
[b] Dept. de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain
[c] Dept. of Psychology, Queens University, Ontario, Canada
[d] INSERM-CEA Cognitive Neuroimaging Unit, Neurospin center, Gif-sur-Yvette, France
[e] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ARTICLE INFO

Article history:
Accepted 20 May 2013

Keywords:
Audiovisual speech
Bilingualism
fMRI

ABSTRACT

Neuroimaging studies of audiovisual speech processing have exclusively addressed listeners' native language (L1). Yet, several behavioural studies now show that AV processing plays an important role in non-native (L2) speech perception. The current fMRI study measured brain activity during auditory, visual, audiovisual congruent and audiovisual incongruent utterances in L1 and L2. BOLD responses to congruent AV speech in the pSTS were stronger than in either unimodal condition in both L1 and L2. Yet no differences in AV processing were expressed according to the language background in this area. Instead, the regions in the bilateral occipital lobe had a stronger congruency effect on the BOLD response (congruent higher than incongruent) in L2 as compared to L1. According to these results, language background differences are predominantly expressed in these unimodal regions, whereas the pSTS is similarly involved in AV integration regardless of language dominance.

1. Introduction

Audiovisual (AV) binding is an integral aspect of language processing in natural face-to-face conversations, as well as in modern media such as TV, cinema, or video-conferencing. Visual cues can strongly support the perception of speech when they correlate with auditory cues (especially in noisy environments; e.g., Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2006; Sumby & Pollack, 1954) and they can dramatically alter auditory perception when they do not correspond to acoustic speech (e.g., McGurk & MacDonald, 1976). The neural correlates underlying AV integration of speech have been addressed by several neurogimaging approaches, including fMRI, MEG and EEG (Callan, Callan, Gamez, Sato, & Kawato, 2010; Calvert, 2001; Calvert, Campbell, & Brammer, 2000; Calvert, Hansen, Iversen, & Brammer, 2001; Colin, Radeau, Soquet, Dachy, & Deltenre, 2002; Colin, Radeau, Soquet, & Deltenre, 2004; Colin et al., 2002; Miller & D'Esposito, 2005; Skipper, Nusbaum, & Small, 2005; van Wassenhove, Grant, & Poeppel, 2005). However, these investigations have exclusively addressed native language processing. According to recent behavioural literature, an important, yet barely investigated, aspect of AV speech integration relates to its contribution in second language comprehension. This is the focus of this study.

Previous behavioural studies have shown how the visual correlates of speech alone contain sufficient information for speakers to discriminate between languages (Soto-Faraco et al., 2007), even in pre-linguistic infants (Weikum et al., 2007). These results indicate that some aspects of visual information from facial movements can be decoded to some extent, even in a non-native (or unfamiliar) language. Thus from a theoretical standpoint, we can expect the involvement of visual speech, and therefore of AV integration, in second language perception (when visual cues are available to listeners). The potential gain in overall comprehension arising from the integration of vision with speech sound tends to be larger because the information available from sound is less reliable (Sumby & Pollack, 1954). In fact, this is precisely the situation encountered when attempting to understand the sounds of a second language. In line with this idea, behavioural studies have shown that the addition of visual information (e.g., mouth movements) can enable the phonological discrimination between non-native sounds, which are otherwise undistinguishable on the basis of auditory

Abbreviations: L1, native language; L2, non-native language; pSTS, posterior superior temporal sulcus; STG, superior temporal gyrus; AV, audiovisual; AVc, audiovisual congruent; AVi, audiovisual incongruent; A, auditory; V, visual; B, baseline; MSI, multisensory interaction.

* Corresponding author. Address: Department of Psicología Básica, Clínica y Psicobiología, Universitat Jaume I, Avda. Sos Baynat s/n, 12071, Spain. Fax: +34 964729267.

E-mail address: maya.visser@gmail.com (M. Visser).

[1] These authors have contributed equally to this work.

0093-934X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bandl.2013.05.009

cues alone (Hirata & Kelly, 2010; Navarra & Soto-Faraco, 2007; Reisberg, McLean, & Goldfield, 1987), resulting in an improved overall comprehension of L2 (although the contribution of AV integration in L2 perception is not always effective; Hazan et al., 2006). For example, a recent study showed that available visual mouth movements improve auditory L2 learning (Hirata & Kelly, 2010). Likewise, Wang et al. (2011) demonstrated that (1) adding visual speech information to auditory speech results in improved phoneme perception in L2, but not L1, and in a (2) stronger AV integration, as shown by an increased McGurk[2] effect on L2 as compared to L1. This suggests that the perception of non-native speech is more influenced by visual speech, whereas auditory input is more dominant in native speech.

Although some behavioural studies have investigated the contribution of AV integration in second language processing (see above), its neural correlates are still largely unknown. As far as we know, the neuroimaging literature on second language processing has exclusively focused on unimodal situations, such as auditory speech comprehension and visual reading, but not on multimodal aspects (for reviews, see Abutalebi, 2008; Indefrey, 2006; van Heuven & Dijkstra, 2010). Unisensory literature on bilingualism has generally shown that the brain regions underlying speech processing in the native (L1) and non-native (L2) language often overlap (i.e., Abutalebi, 2008). However, L2 processing frequently expresses over more extended regions and more strongly. For instance, L2 processing has sometimes been found to result in stronger activations in the frontal and temporal regions, suggesting that more neural resources are used to accomplish the same task in L2 as compared to L1. This seems logical as information at the phonetic, syntactic and semantic levels is harder to extract and parse in L2 (Abutalebi, 2008).

In the current study, we address the potential differences and similarities in the multisensory neural network involved in AV speech perception in L1 and L2. Based on previous unimodal literature, we expect a similar network of brain regions to underlie AV speech integration in L1 and L2, and the pattern of activation during multimodal integration to vary depending on the language background (native or not). We expect the multisensory regions to be more involved in L2 compared to L1 as the integration of the added visual information might play a more important role for L2 comprehensions, as suggested by behavioural results (discussed above). In line with this, we expect (visual) occipital regions to play an important role in these multisensory processes, and more so when dealing with a non-native language. Several studies now converge on the idea that unisensory regions can also respond to stimulation across sensory modality. This suggests that multisensory processes may involve the orchestration of a network that engages classical association areas, as well as regions traditionally regarded as unisensory (e.g., Driver & Noesselt, 2008). For example, auditory regions have been found to respond to visual speech stimuli presented in silence (e.g., speech-reading; Calvert et al., 1997; Kayser, Petkov, Lippert, & Logothetis, 2005; Miller & D'Esposito, 2005).

Former literature on AV speech integration (always in the native language) has identified some regions of interest. Most notably, a considerable body of evidence has associated the posterior part of the superior temporal sulcus (pSTS) with AV integration during language processing (for reviews, see Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, 2005; Campbell, 2008). This cortical region responds to both visual and auditory speech stimuli and, more importantly, it often shows stronger responses when speech stimuli are simultaneously presented in the two sensory modalities (e.g., speech with co-occurring and correlated mouth movements, amongst others, usually meaningful stimuli). The enhancement effect has been highlighted to be a key aspect that defines the responses of the STS and other regions of multisensory integration (Beauchamp, 2005; Calvert et al., 2000; Campbell, 2008). One particularly convincing study associated the pSTS with the phenomenology of AV speech integration (Miller & D'Esposito, 2005). In Miller and D'Esposito's study, AV synchrony varied over time while subjects rated whether they perceived the AV signals as fused or not. The pSTS did not respond during trials when AV information was not perceived to be a fused object, but it displayed activity even for asynchronous stimuli, which were nevertheless perceived as fused. Furthermore, disruption with single-pulse transcranial magnetic stimulation in the pSTS affects AV integration (Beauchamp, Nath, & Pasalar, 2010).

Multisensory responses in the pSTS have been shown for speech input at the semantic level (Beauchamp, 2005; Calvert et al., 2000; Stevenson, VanDerKlok, Pisoni, & James, 2010), the phonological level (e.g., non -words: Miller & D'Esposito, 2005; letters: van Atteveldt, Blau, Blomert, & Goebel, 2010) and time-varying stimulation with non-speech stimuli (e.g., sinusoidal visual motion aligned with sinusoidally modulated sounds; Bischoff et al., 2007; Werner & Noppeney, 2010). These latter studies have demonstrated that the AV integration system in the pSTS is not language-specific, but is responsive to AV correspondence in some important features in the speech signal. Furthermore, activation in this region is dominantly bilateral (Beauchamp, 2005; Bischoff et al., 2007; van Atteveldt, Blau, Blomert, & Goebel, 2010; Werner & Noppeney, 2010), although some studies have also reported unilateral left (Calvert et al., 2000; Miller and D'Esposito, 2005) or right (Stevenson et al., 2010) pSTS activation. Indeed, the left and right STS might be functionally different; for example, Calvert et al. (2000) suggested that the left and right STS might be involved in speech and non-speech stimuli, respectively. Miller and D'Esposito found that the left STS responded to AV stimuli when perceived as fused, whereas the right STS showed a higher BOLD response when perceiving AV stimuli as not fused. However, they did not provide an interpretation of this pattern and future research is required to further investigate this laterality difference. More recent studies on both speech and non-speech stimuli seem to generally reveal a higher blood oxygenation level-dependent (BOLD) response in the left than the right STS to AV correspondence for all stimulus types (Beauchamp, 2005; Bischoff et al., 2007; van Atteveldt et al., 2010; Werner & Noppeney, 2010). To summarise, extant literature suggests that the bilateral pSTS is a critical (but not necessarily the only) region for AV integration in language, and that this pattern is stronger in the left pSTS.

Initial reports of multisensory enhancement considered those areas displaying BOLD responses to bimodal speech stimuli which were significantly larger than the sum of the BOLD responses to each unimodal (visual or auditory) speech stimulus when presented in isolation (called the super additivity effect: Calvert et al., 2000). This set of criteria, inherited from single cell physiology, has been shown to have its advantages (it is safe against false-positives from areas containing separate populations of visual and auditory unisensory neurons). However, it may be overly conservative (Beauchamp, 2005; Goebel & van Atteveldt, 2009; Laurienti, Perrault, Stanford, Wallace, & Stein, 2005) due to the saturation effects in the BOLD signal and its dependence on the relative proportion of multisensory to unisensory neurons in a given region (e.g., pSTS: Laurienti et al., 2005). Therefore, we decided to use the max criterion, as described by Beauchamp (2005), in the present study to reveal the multimodal responses to L1 and L2 AV speech processing. Beauchamp proposed that the multisensory response should be greater than the maximum of the unisensory responses (for further reading on these and other related issues, see

[2] The McGurk effect refers to the occasion when an audio /ba/ and a visual /ga/ result in the perception /da/, indicating that audio and visual speech information is integrated.

Table 1
Demographic details of the two bilingual groups.

	Spanish L1	English L1	Between-group differences
Number of subjects	21	21	
Age	25.29 (6.05)	28.90 (9.49)	$t(40) = 1.47$, $p = .15$
Lateralization (right/left/bimanual)	17/1/3	19/1/0	$t(40) = 0.00$, $p = 1.00$
L2 Age of acquisition	10.61 (4.96)	16.52 (7.93)	$t(38) = -1.41$, $p = .17$
L2 Self-rated proficiency (1/best to 4/worst)			
Comprehension	1.62 (.50)	1.86 (.86)	$t(40) = -1.10$, $p = .28$
Reading	1.43 (.51)	1.80 (.83)	$t(40) = -1.73$, $p = .09$
Fluency	1.71 (.46)	2.05 (.94)	$t(40) = -1.46$, $p = .15$
Writing	1.67 (.58)	1.95 (.89)	$t(39) = -1.22$, $p = .23$

Beauchamp, 2005; Goebel & van Atteveldt, 2009; Laurienti et al., 2005). For regions which survived the max criteria, we further explored the multisensory interaction pattern by looking at non-linearity using the approach described by van Atteveldt et al. (2010). This measurement calculates the difference between the total percentage BOLD signal change of the AV condition and the unisensory conditions with the max response.

Apart from the logic based on the additivity criterion, the congruency criterion has proven successful to reveal the regions associated with AV processing (Calvert, 2001; van Atteveldt et al., 2010). The hypothesis states that the congruency criterion is that if a region's BOLD response differs for congruent information from that for incongruent AV information. If this were the case, it means that this region is involved in some kind of multisensory integration. Note, however, that from a logical (and empirical) point of view, the reverse is not necessarily true; that is, not all multisensory regions might be sensitive to stimulus congruency in one domain or more (Campbell, 2008).

In short, there is a considerable body of imaging research on AV integration in native language, and also on bilingualism using unisensory auditory stimulation. Nonetheless no attention has yet been paid to second language processing and AV speech integration. Given the results of former behavioural studies, the potential for AV enhancement in second language processing is, at least, as important as in the first language. The current study therefore aims to bridge the gap between these two study areas by investigating the neural correlates of AV integration in L1 vs. L2. Our hypothesis is that similar regions are involved in AV integration for L1 and L2. However, we hypothesize that the BOLD signal might be stronger in L2 than in L1, and that L2 AV speech processing might rely more on the visual network as visual information seems relatively more important in L2.

2. Methods

2.1. Subjects

Forty-two bilingual volunteers (age range 20–46 years old), proficient in English and Spanish, were included in the study. Half the sample ($n = 21$, 10 females) spoke Spanish as their native language and English as their second non-native language, while the other half ($n = 21$, 9 females) spoke English as their native language and Spanish as their second non-native language. By pooling two equivalent groups of participants with the reverse language dominance pattern, we were able to cancel out possible group effects which correlated with language background or stimulus-based effects. Participants were late bilinguals who had lived a considerable amount of time in the second language environment (English or Spanish). The groups did not differ in terms of their onset age of exposure to their second language, as assessed by a questionnaire of language use (Costa, Hernández, & Sebastián-Gallés, 2008). Furthermore, the groups did not differ in terms of their L2 proficiency in comprehension, fluency, reading and writing skills, as assessed by a self-rated questionnaire (see details in Table 1). All the participants were in good health, had no personal history of psychiatric or neurological diseases, and had normal auditory acuity and normal or corrected-to-normal (visual lenses from VisuaStim, Magnetic Resonance Tech.) visual acuity. They all gave informed consent prior to participation in the study.

2.2. Stimuli and design

Stimuli comprised 5-s (5-s) long speech fragments made up of sentences which were used as stimuli. For the auditory condition (A) auditorily only, speech was presented with a blank screen; for the visually only condition, speech was presented visually without sound (V). There were two audiovisual conditions (AV): in one, audio and video were congruent (AVc), but audio and video channels were incongruent in the other (AVi). In the AVi stimuli the auditory, the sound track of one sentence was combined with a different visual sentence of the same duration. Yet another condition, without auditory (silence) or visual speech (blank screen), was included as baseline (B). Equivalent sets of stimuli were generated in each test language (English and Spanish) for each single condition from recordings of an English–Spanish well-balanced bilingual speaker[3] used in a previous experiment (Navarra et al., 2010). The speech fragments were made up of sentences selected from a set of 224 sentences (112 in Spanish and 112 in English) digitally recorded by the bilingual speaker, showing the frontal view of the entire face and shoulders. Sentences were obtained from different (non-popular) tales appearing in literature-specialized web pages. Most sentences were slightly modified to match the number of syllables required and infrequent words were avoided (or replaced, if necessary). As all the materials (English and Spanish) were obtained from the same sources, were confident that sentences were equivalent in terms of the frequency of use of the words and familiarity. The video clips (720×576 pixels presented in 25 frames/s) were edited using the Adobe Premiere software and were compressed with a single avi video codec for their use in the Presentation® software (Neuro Behavioral Systems Inc.). In order to achieve a smooth transition between the clips within a block, a fade-in and fade-out of 720 ms and 560 ms were introduced at the beginning and the end of each video clip in both the audio and video channels.

Each participant was presented with each stimulus language in a different test run. Within each run, the four different modality conditions (A, V, AVc and AVi), plus baseline condition (B), were presented in a blocked design fashion, with a pseudo-randomized

[3] The speaker was a Spanish-born 28-year-old male with a very high proficiency in English. He was schooled in English since the age of 3, and had lived in English-speaking countries (the US and the UK) since he was 18. A group of English natives (10) and Spanish natives (12) evaluated his proficiency on a scale of 0–10 (10 = native sounding) using a sample of the materials included in the experiment. In Spanish, all the judges considered that his Spanish was perfect (mean score of 10). The English evaluators judged our speaker's English as close to perfect (mean score of 8).

block order (avoiding consecutive blocks of the same condition). Each block type was repeated 4 times, lasted 40 s and included eight speech fragments (or baseline) of 5-s durations. The run order was also counterbalanced between subjects. Stimuli were presented by visual and auditory MRI compatible systems (Visuastim, Resonance Technologies, Inc.). In order to prevent potential familiarity effects by recognizing sentences employed in previous trials during the study, each participant was presented with each sentence only once during the experiment. Different versions of the experimental materials ensured that each particular sentence was presented in all the modality conditions across all the participants (e.g., sentence 1 was presented in condition A to Participant 1, in condition AVc to Participant 2, etc.).

Subjects were instructed to listen to each sentence and to focus on the screen (even during auditory and baseline conditions) since we informed them that they would be asked to perform a recognition test after scanning. During the recognition test, some of the experimental stimuli, plus a number of comparable foils, were presented, and participants were asked to judge whether they had seen/heard that utterance before or not. This test included 16 trials; 1 target and 1 foil per condition (i.e., A, V, AVc and AVi) per language (L1 and L2). This was included to ensure an attentive strategy during stimulus presentation (Beauchamp, 2005; Calvert et al., 2000).

2.3. Image acquisition

Gradient-echo echo-planar (EPI) and anatomical MR images were acquired using a 1.5-Tesla scanner (Avanto, Siemens). A total of 100 volumes per run of the $T2^*$-weighted images depicting the BOLD contrast were sparsely acquired over 10 min and 40 s with an 8-s TR (TE = 60-s, TA = 2 s; flip-angle = 90ª, voxel-matrix = 64 × 64; voxel-size = 3.94 × 3.94, 5-mm thick and 0.5-mm gap, 1 interleave). Twenty-five coronal slices, which were perpendicular to the Sylvian fissure covering the whole brain, were acquired. In our sparse sampling design, 5 × 2-s volumes were acquired per block. The first volume was acquired 3 s after the onset of the stimuli. The following four volumes were acquired with 6 s gaps (hence a TR of 8 s).

Anatomical scans were also obtained using a contiguous 1-mm sagittal images across the entire brain with a T1-weighted fast-field echo sequence (TE = 4.2 ms, TR = 11.3 ms, flip angle = 90; FOV = 24 cm; matrix = 256 × 224 × 176).

2.4. FMRI data analysis

Pre-processing: prior to the time-series statistical analyses, the data from each subject were pre-processed by SPM5 (Welcome Department of Cognitive Neurology, London, UK). Slice-timing was not applied. Functional images were realigned with a two-pass procedure in which functional volumes were registered to the first volume in the series in a first step, and to the mean image of all the realigned volumes in a second step. Anatomical scans from each subject were then co-registered to the mean image and were segmented. Normalization parameters were extracted from the segmentation of each subjecťs anatomical T1-weighted scan and were applied to their corresponding functional scans (rescaled voxel size 3 × 3 × 3-mm^3, template provided by the Montreal Neurological Institute). Finally, functional volumes were smoothed with a Gaussian kernel of 6-mm FWHM.

Processing and statistical analyses: the conditions of interest corresponding to A, V and AVc and AVi for both L1 and L2 were modelled using a box-car function. Low frequency drifts were removed with a temporal high-pass filter (default cut-off of 128-s) and temporal autocorrelations corrected between observations. Furthermore, six different additional covariates, corresponding to the parameters of movement correction obtained in the realignment step of the functional scans, were applied to regress out movement effects. The estimated parameters for each participant were entered in a within-participants ANOVA to perform tests at the group level. The current fMRI analyses were collapsed across language dominance groups (i.e., the English and Spanish native speakers; see Table 1). We first ensured that there were no significant differences between the two groups per condition of interest. In addition, we checked whether the pattern that arose when we collapsed across groups also held when examining groups separately. Finally, in order to correct for multiple comparisons, we used a voxel-wise threshold of $p < 0.001$ in combination with a cluster criterion (Forman et al., 1995) determined by Monte Carlo simulations using the AFNI program Alphasim. This resulted in a cluster-size criterion of 13 voxels for a family-wise error rate of $p < 0.05$.

2.5. Multisensory enhancement

To test for multisensory enhancement in the AV congruent condition (AVc) compared to the unimodal (A and V) conditions at the L1 and L2 group levels, we performed the conjunction of [(AVc > A) ∩ (AVc > V) ∩ (A > B) ∩ (V > B)], where B refers to the baseline or ́rest ́ condition. This contrast is referred to as the ''max criterion'' and is commonly applied in multisensory research (Beauchamp, 2005; Van Atteveldt, Formisano, Blomert, & Goebel, 2007). The result obtained from this conjunction gave a statistical value for each voxel as the minimum of the t-statistical values obtained from the four included contrasts (Beauchamp, 2005). Van Atteveldt et al. (2007) used a multisensory interaction (MSI) measure to visualize the multisensory enhancement in region-of-interests (ROIs) based on functional data. This measurement calculates the difference between the total percentage of BOLD signal change of the AV condition and the unisensory conditions with the max response. Van Atteveldt et al., used the total percentage of BOLD signal change (baseline [100%] + signal change, e.g., 101.4%) to calculate the MSI instead of the BOLD signal change (e.g., 1.4%) in order to avoid extreme outliers in the MSI values. In the current study, we defined the functional ROIs as clusters which survived the max criterion in the group data for both L1 and L2, and we calculated this MSI index for each participant in these functional ROIs.

Although we used the max criterion as described by Beauchamp (2005) in combination with the MSI measure (see above) in the current study, there are different methods to investigate the multisensory network, as mentioned in the Introduction. In order to further characterize our results, we examined whether any regions showed a non-additive interaction (super- or sub-additive effects) following the approach described in Lee and Noppeney (2011). We first looked for any regions showing a non-linear response with either a supra- or a sub-additive pattern using the contrasts (AVc) > (A + V) and (AVc) < (A + V). The resulting significant AV interactions were then characterized as multisensory enhancement [(AVc > A) ∩ (AVc > V) ∩ (A > B) ∩ (V > B)] or multisensory suppression. [(AVc < A) ∩ (AVc < V) ∩ (A > B) ∩ (V > B)]. We examined these AV interactions for L1 and L2 separately and examined possible language differences.

2.6. Congruency effects

We constructed a second set of analyses to address the neural consequences of AV congruent as compared to AV incongruent stimulation. In this case, we directly compared the corresponding (AVc) with the mismatched (AVi) audio-visual speech. This contrast is interesting because the two terms contain equivalent amounts of sensory input in each modality, and they differ only in terms of the degree of cross-modal congruency. This contrast allowed us to perform a whole brain analysis including Language (L1, L2) and Condition (AVc, AVi) in a repeated measures ANOVA.

3. Results

3.1. Results of the recognition test during the scanning session

The participants performed some recognition trials after the scanning session (see Section 2). This test included only one observation per condition, which did not directly inform about online comprehension as it was included mainly to ensure that the participants remained in an attentive state during the scanning session without having to perform an online task leading to interference. Nevertheless, we present the results for completeness (the data from two participants, one from each language group, were lost due to an error made by the experimenter). The average recognition results are presented in Table 2. A two (language dominance; L1 and L2) by four (modalities; A, V, AVc, AVi) ANOVA revealed the significant effect of language dominance ($F(1) = 6.9$, $p = 0.01$) and modality ($F(3) = 10.0$, $p < .001$), but interaction was not significant ($F(3) = 0.34$, $p = .8$). The overall mean of L1 ($M = 0.67$, SD = 0.34) was higher as compared to L2M = 0.57, SD = 0.37.

3.2. Neural correlates of multisensory integration

As the Method section describes, we used the max criterion to reveal the multisensory regions involved in AV speech processing for L1 and L2. This criterion requires the response to the AV congruent condition to be higher than the highest response in any unimodal condition. The clusters that survived the max criterion are presented in Table 3 and Fig. 1 at a corrected level for multiple comparisons (using the Monte Carlo simulations in AFNI). For both L1 and L2, we observed the bilateral activation of the posterior superior temporal sulcus (pSTS). We further examined the mean percentage BOLD signal change of the different conditions making up the max criterion (e.g., audio, visual and AV congruent). These are presented in Fig. 1B. In addition, we also included Van Atteveldt et al.'s (2007) multisensory interaction (MSI) measure to calculate the multisensory enhancement (see the Method section for a description), which are presented in Fig. 1C. A paired-sample *t*-test showed no significant differences in MSI between L1 and L2 in either hemisphere ($t(40) = 0.16$; $p = 0.88$ and ($t(40) = 1.0$; $p = 0.32$ for the left and right hemisphere, respectively). This is in line with research which has suggested that the neural language system is similarly engaged in L1 and L2, at least in proficient (as opposed to low-proficient) bilinguals (Abutalebi, 2008).

As the max criterion does not necessarily inform about potential non-linearities in neural responses to multisensory integration, we ran a second set of analyses in accordance with a method recently used by Lee and Noppeney (2011). In this method, the results need to survive two steps. The first step tests whether any regions showed a sub- or supra-additive pattern, whereas the second step uses the max or minimum criterion $[(AVc > A) \cap (AVc > V) \cap (A > B) \cap (V > B)]$ or $[(AVc < A) \cap (AVc < V) \cap (A > B) \cap (V > B)]$ (see Method section). In the first step, L1 showed a subadditive effect in the bilateral inferior frontal gyrus (IFG, MNI coordinates; 54, 15, 27 and −54, 21, 27; corrected for multiple comparisons using the Monte Carlo simulations in AFNI). This region is commonly associated with multisensory processing, including AV speech processing (Lee & Noppeney, 2011; Calvert 2001; Campbell, 2008). However in the second step, this region did not survive either the max or the minimum criterion. Therefore, the possible differential role of IFG in L1 and L2 must remain speculative for the time being. It is perhaps interesting to note that the pSTS was not significant for the interactive effect; that is, we cannot assume a pattern beyond additive in the present study.

Table 2
Means (standard deviations) of the recognition test. After the scanner session, the participants performed a recognition test. They were presented with sentences and instructed to press a button if they believed they had seen the sentence during the experiment in the scanner.

Conditions	Native	Non native
Native		
Auditory	0.60 (.34)	0.54 (.35)
Visual	0.50 (.34)	0.44 (.36)
Audiovisual congruent	0.76 (.32)	0.63 (.37)
Audiovisual incongruent	0.80 (.27)	0.66 (.38)
All conditions collapsed	0.67 (.34)	0.57 (.37)

3.3. Congruency effects

We further examined the regions showing a congruency effect. When examining within each language dominance separately (L1 and L2), the multisensory regions that had been highlighted by the max criteria in the previous analyses showed no significant congruency effect. Instead AV congruency in L2 resulted in the significant activation of two clusters in the visual areas: the right middle occipital lobe (BA 18/19) and the left lingual gyrus (BA 17/18) for the congruent condition. These are classically defined unisensory areas. Otherwise, no regions showed any significant congruency effect when testing the opposite contrast (AVi < AVc) either for L1 or L2. We followed-up on the occipital regions showing AV congruency effects in L2 in order to confirm differential effects in accordance to language dominance (L1 vs. L2). Fig. 2 shows the location of the regions and the percentage BOLD signal change of the occipital clusters for the congruent and incongruent conditions in L1 and L2. The percentage BOLD signal change of both visual clusters showed not only a significant congruency effect, but also significant language dominance by congruency interaction. for cluster −21 −87 −1; ($F(1) = 12.66$; $p = 0.001$) and ($F(1) = 9.98$; $p = 0.003$), respectively; for cluster 30 −82 −6; ($F(1) = 18.94$; $p < 0.001$) and ($F(1) = 8.80$; $p = 0.005$), respectively. To further examine the BOLD pattern in these regions, we compared L1 to L2 in the congruent condition and in the incongruent condition separately (i.e., AVcL1 *vs.* AVcL2 and AViL1 *vs.* AViL2). These results reveal that there were no significant differences between languages in the incongruent condition. However, both the left and right occipital lobes presented a stronger response to AV congruency in L2 as compared to L1 (i.e., AVcL2 > AVcL1; $T(40) = 2.84$; $p = 0.007$ and $T(40) = 3.14$; $p = 0.003$ for the left and right occipital lobe, respectively). This suggests that visual regions are more strongly engaged in processing AV congruent speech in L2 as compared to L1, and is in line with the idea that visual speech and AV integration are more important during L2 AV speech perception. Attention resources focus less on visual speech in L1, resulting in a lower BOLD response. Note, however, that successfully filtering out the visual component of a speech event (say, for the incongruent condition) is relatively unlikely because the strong illusions arising when incompatible AV stimuli are presented (i.e., McGurk illusion) cannot be avoided voluntarily (McGurk & MacDonald, 1976; Soto-Faraco, Navarra, & Alsius, 2004) (see Table 4).

3.4. Possible effects of group, length of L2 exposure and experience

In the fMRI analyses above (i.e., max criterion and congruency effects), we collapsed across language groups (i.e., the English and Spanish native speakers; see Table 1 in Section 2). This introduces the desirable feature as none of the effects observed can be due to only between-group differences (all subjects contributed to L1 and L2 BOLD) or to only particular aspects of the stimulus language (both English and Spanish stimuli played the role of L1 and L2). However in order to confirm our results, we repeated all the analyses for each language group separately (Spanish and

Table 3
Location of main activation clusters after applying the max criterion analysis to L1 and L2.

Brain region	BA	TAL coordinates			T value	Cluster size
		X	*Y*	*Z*		
L1						
R. Superior Temporal Sulcus	41	53	−34	10	4.90	56
L. Superior Temporal Sulcus	13	−45	−46	13	3.78	38
L. Superior Temporal Sulcus	22	−53	−40	8	3.52	
L2						
R. Superior Temporal Sulcus	22	56	−37	13	4.12	20
L. Superior Temporal Sulcus	13	−45	−46	13	3.65	13

Clusters survived a corrected family-wise error rate of $p < 0.05$, defined by Monte Carlo simulations using the AFNI program Alphasim.

English speakers) at an uncorrected level of $p < 0.001$. The results of each group did not differ significantly, as with the collapsed analyses presented above. Therefore, these results generalize our findings to two different languages and populations.

Furthermore, although our study did not intend to study the effects of second language proficiency due to the possible influence of this factor on language processing, we checked whether the length of L2 exposure or L2 proficiency modulated some of the effects noted herein. These linguistic parameters were measured by a questionnaire on language use (Costa et al., 2008; see Section 2 and Table 2). To calculate the length of L2 exposure, we subtracted participants' age of acquisition from their current age. The self-rated proficiency test included comprehension, reading, fluency and writing. We introduced these factors as covariates in an ANCOVA. The results of both the max criterion and the congruency criterion did not change if compared to the above-described results, indicating that these factors do not play a significant role in the current study. Moreover, the length of exposure to L2 and L2 proficiency were included in two separate regression analyses. The extracted parameter estimates from the left and right clusters of the significant multisensory integration effects, and those from the occipital clusters showing congruency effects, were regressed with these two variables separately. Neither length of exposure to L2 nor L2 proficiency gave a significant correlation with brain activation in terms of the multisensory integration or congruency effects for L2 in either the bilateral pSTS or the posterior occipital regions, respectively.

4. Discussion

The present study aims to examine the neural correlates of AV speech processing in second language perception. We first targeted the multisensory regions that displayed enhancement effects to AV congruent stimulation in comparison to unisensory stimulation. We found that the posterior superior temporal sulcus (pSTS) was activated in AV speech processing in both native and non-native language. This area (pSTS) is well in line with previous AV speech research (Beauchamp, 2005; Calvert, 2001; Calvert et al., 2000; Calvert et al., 2001; Campbell, 2008; Goebel & van Atteveldt, 2009; Miller & D'Esposito, 2005). What is more, in our case we found that the BOLD enhancement in the pSTS was equivalent in both language dominance conditions. This indicates that AV integration into the bilateral pSTS underlies a similar functional role in processing L1 and L2. Secondly, we discovered that the BOLD responses in the occipital lobe responded differentially to congruent vs. incongruent stimulation in accordance with the language status of the stimulus for the participant. The fact that this unimodal region is influenced by multimodal input suggests the close collaboration of putatively unisensory regions with the AV integration network. This finding is in line with previous research which revealed a set of subnetworks for dissociable components of AV speech integration (Driver & Noesselt, 2008; Hertz & Amedi, 2010; Skipper, Goldin-Meadow, Nusbaum, & Small, 2009). We now go onto discuss the implications of these findings for the characterization of AV speech processing in the second language.

4.1. Multisensory speech integration in first and second languages

Our experiment identified AV integration regions during second language processing. We used the max criterion, which requires the BOLD signal in an AV region to be higher during AV input as compared to the maximum of the two unimodal (visual and auditory) inputs (Beauchamp, 2005) to reveal overlapping regions in the bilateral pSTS involved in AV integration for L1 and L2. As far as we are aware, this is the first study that links the pSTS with AV processing in a second language. Previous studies on auditory speech perception in bilinguals have shown that many speech processes engage overlapping regions for L1 and L2, albeit sometimes with different BOLD intensities (for reviews, see Abutalebi, 2008; Indefrey, 2006; van Heuven & Dijkstra, 2010;). Based on this previous result, we examined whether there was possibly a difference in the percentage BOLD signal change in the pSTS across language dominance. Nonetheless, the results reveal that the percentage BOLD signal change in the pSTS was equivalent for both language dominance levels. To summarize, the present results clear evidence that an equivalent or a very similar integration system in the bilateral pSTS underlies multisensory processing for speech in L1 and L2 in high-proficient bilinguals. Future research is required in order to verify whether this result also holds for low-proficient bilinguals. Despite further analyzing the nature of this multisensory response, we found no evidence for an interaction pattern beyond (or below) additive, which means that the response of the pSTS was additive both for L1 and L2, at least for this particular case.

4.2. Congruency effects

The congruency contrast is used to examine regions that respond differently in AV congruent compared to incongruent information. The multisensory region identified with the enhancement criterion (pSTS) did not respond selectively to the congruent condition. Although the congruency criterion has been used to identify multimodal regions, it is important to note that not all multisensory regions are sensitive to the congruent–incongruent effects (Campbell, 2008). In some previous studies, as in the case presented herein, the bilateral pSTS particularly failed to respond to congruency manipulation (Bushara, Grafman, & Hallett, 2001; Miller & D'Esposito, 2005; Ojanen et al., 2005). In addition, other studies found a higher BOLD response to the incongruent condition *vs.* the congruent condition (Benoit et al.,2010; Pekkola et al., 2006), or vice versa (Calvert et al., 2000; van Atteveldt, Formisano, Goebel, & Blomert, 2004; van Atteveldt, Roebroeck, & Goebel, 2009; note that Calvert et al. found only left pSTS activation). Therefore, the responsiveness of the bilateral pSTS to congruency remains unclear.

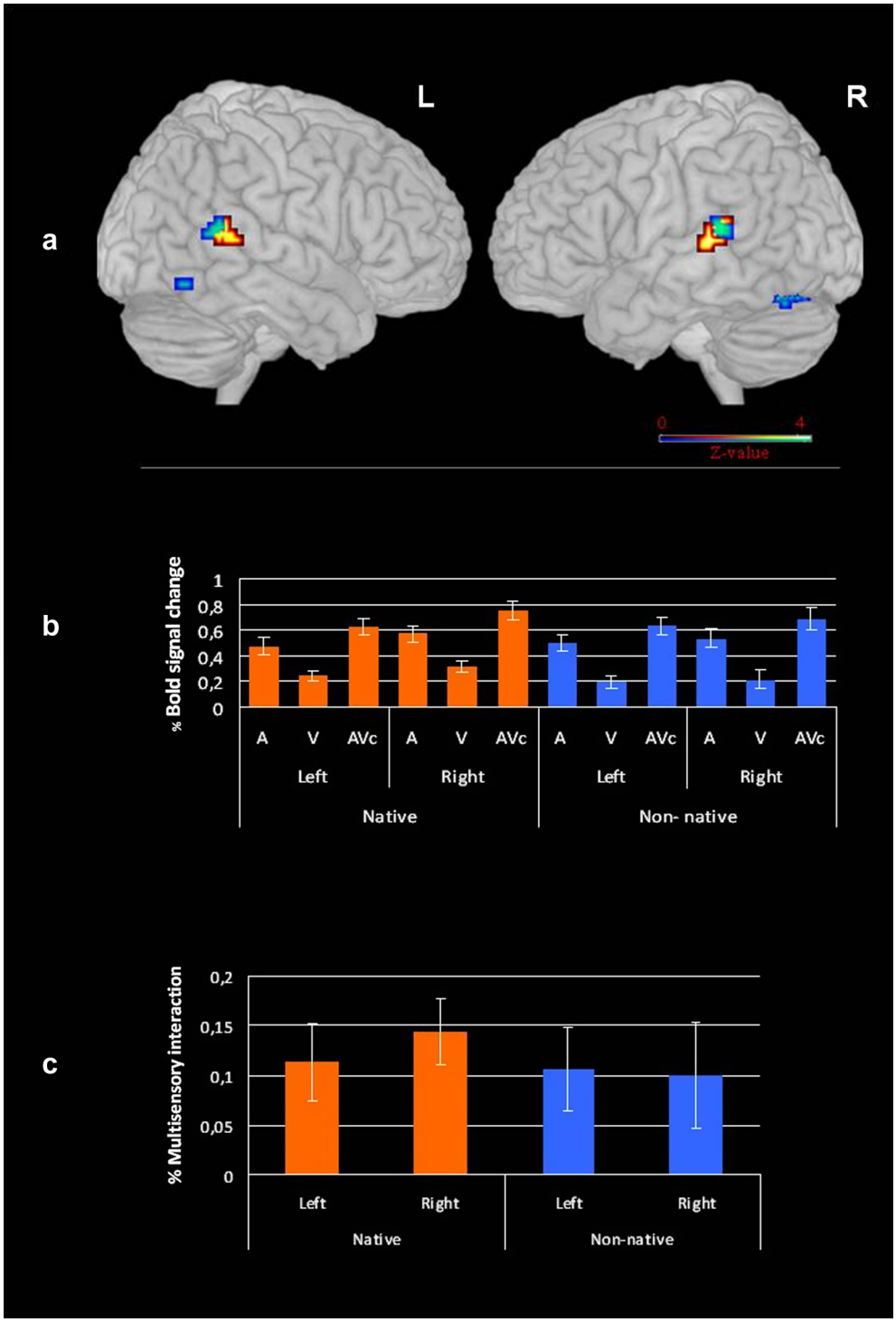

Fig. 1. The regions involved in audiovisual speech processing in native (L1) and non-native language (L2). (A) The bilateral superior temporal sulcus (STS) was identified as a region that responds to audiovisual stimulation using the max criterion: AVc conditions resulted in higher activation that the max response of any unimodal condition. The figure presents the response to AV speech in L1 (red-yellow) and L2 (blue-green). (B) Percentage BOLD signal change in the pSTS during speech processing of auditory (A), visual (V) and AV congruent (AVc) information, the three conditions used to compute the max criterion. (C) The multisensory interaction (MSI) values were calculated in the pSTS clusters that survived the max criterion (see Section 2). To assess this MSI index, the bimodal response was calculated in relation to the most effective unimodal response (van Atteveldt et al., 2007). As in the initial max criterion analysis, language background differences did not result in significant differences in this value. Error bars represent the standard error of the mean.

The important finding of the present congruency analysis is that the percentage BOLD signal change in a region traditionally considered unimodal is responsive to AV congruency (vs. incongruency) of multimodal speech information according to which language (L1 or L2) is being processed. This is a most interesting finding, and is in line with the literature as it suggests that unimodal regions are engaged by multimodal processes (Driver & Noesselt, 2008; Hertz & Amedi, 2010). Our results provide further insight by demonstrating that the AV pairing type partly determines the responsiveness of this region. In particular when looking at the congruency effects within L2, the active clusters concentrate in the occipital lobe; more precisely, the right middle occipital gyrus

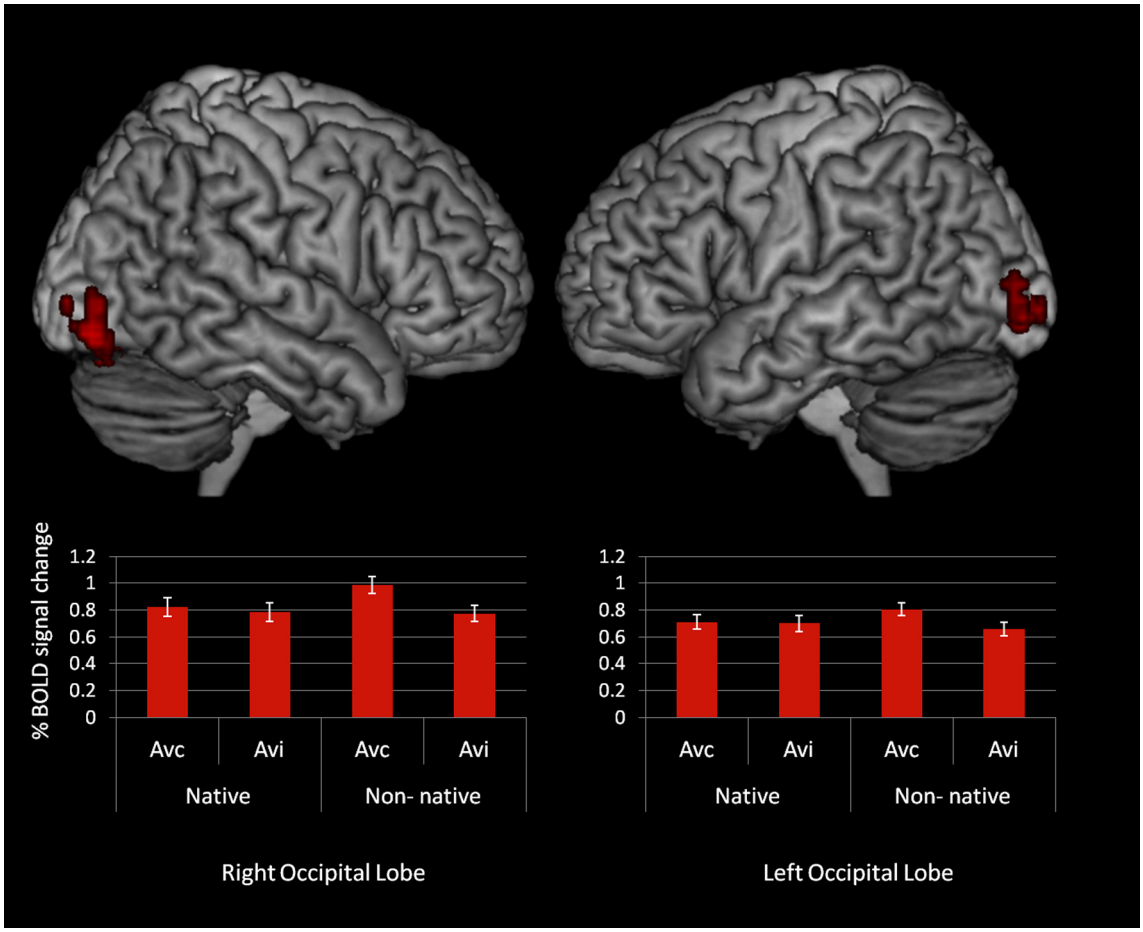

Fig. 2. Above the clusters were significant for the contrast AVc > AVi in L2. Clusters survived a corrected family-wise error rate of $p < 0.05$, defined by the Monte Carlo simulations using the AFNI program Alphasim. Note that no clusters were significant when this same contrast was applied in L1. Below, the percentage BOLD signal change in these clusters for the congruent and incongruent conditions in L2 and L1 per hemisphere are presented.

Table 4
Regions responding to the AV congruent as opposed to the AV incongruent sentences during L2 processing.

Contrast	Brain region	BA	TAL coordinates			*T* value	Cluster size
			X	*Y*	*Z*		
AVcL2 > AViL2	R. Middle Occipital Gyrus	18/19	30	−82	−6	4.59	75
	L. Mammillary Body		0	12	−7	3.95	15
	L. Lingual Gyrus	17/18	−21	−87	−1	3.79	43

These clusters survived a corrected family-wise error rate of $p < 0.05$, defined by Monte Carlo simulations using the AFNI program Alphasim.

(BA 18/19) and the left lingual gyrus (BA 17/18). These occipital regions are involved in visual processing and our data suggest that they play a relatively more important role in AV speech processing in L2. This is quite remarkable because the spatio-temporal alignment/misalignment (in the congruent/incongruent conditions, respectively) is physically the same for both language types. That is, the only difference lies in whether the participant has previous native experience with that particular language from infancy, or has else learned the language later.

In behaviour terms, it has been shown that congruent visual information presented simultaneously with auditory information can improve second language speech perception, and in some tasks, it can do more so for non-native speech (Navarra & Soto-Faraco, 2007; Wang et al., 2008). For example, Wang et al. (2008) compared the speech perception of English phonemes in Mandarin–English bilinguals and native English speakers. Congruent visual speech information improved English speech perception in bilinguals, but not in native English speakers, indicating that these native speakers can extract sufficient information from the auditory signal (and can, therefore, rely more on audition). Interestingly, Wang et al. (2008) measured the McGurk[4] effect, a perceptual illusion created with AV incongruent syllables, using English material on English natives and Mandarin-speaking learners of English. They found that the illusions were more pronounced in the Mandarin speakers when compared to the native English group, but solely for those phonemes that did not exist in Mandarin. These results suggest that bilinguals tend to make comparatively better use of the visual speech information in L2, especially for difficult foreign sounds (see also Navarra & Soto-Faraco, 2007). Note that the visual regions arising in the L1 vs. The L2 comparison of AV congruence in our fMRI experiment did not survive the max criterion (when

[4] This is a perceptual illusion in which the incongruent visual information results in a misperception of the auditory speech information. For example, the speech sound /ba/ presented simultaneously with visual speech information of /ga/, will result in a perceived /da/ (McGurk & MacDonald, 1976).

seeking the enhancement effect). It is possible that this region does not have a significant response to auditory speech alone, thus it fails to meet the max criterion, which requires a positive BOLD response to either sensory modality in isolation. Altogether, this pattern suggests that multisensory regions (pSTS) play a modulatory role in the responsiveness of these unisensory areas during AV processing.

In all, the present results strongly suggest that sensitive regions to multisensory speech input, such as the bilateral pSTS, collaborate closely with the unimodal regions recruited for the task. This is not a new idea because, in the past, it has been proposed that multisensory processing is carried out through the interplay between association (heteromodal) regions and the regions traditionally considered unisensory (e.g., Campbell, 2008; Driver & Noesselt, 2008). However, our results offer a new finding: this interplay engages different parts of the network, and at varying strengths depending on the language background of the speaker/stimuli (native vs. non-native). We would like to emphasize that we do not claim that this network is specific for speech since it may well also play a role in non-speech stimuli (Campbell, 2008; Ghazanfar, Maier, Hoffman, & Logothetis, 2005). What we contend is that it performs a paramount function during speech processing, and that the native vs. non-native nature of the language being processed seems to attune this functional network in different ways.

4.3. Conclusions and future research

We investigated the neural correlates of brain regions involved in AV speech processing when bilinguals use their native vs. their second language. The results show that the pSTS is involved in AV processing in L1 and L2 and to a similar extent (testing high-proficient bilinguals at the sentence comprehension level). What is clear from our results is that, in close relation to previous behavioural studies showing the effects of AV integration in L2, similar neural responses to AV speech integration are shown for second and first languages. In addition, the clusters in the occipital lobe are dominantly associated with L2 as compared to L1 AV speech processing. We believe that the fact that these unimodal regions respond to multimodal stimulation reveals a modulatory effect arising from the interactivity between the unimodal and multimodal components of the multisensory processing network. In our case, these modulatory effects seem to reflect stronger reliance on visual processing when perceiving L2. Future research is required to investigate how brain regions, such as the Supramarginal Gyrus, posterior Superior Temporal Gyrus, Broca's and the Anterior Temporal Lobe, interact in this multimodal network, which involves particular aspects of speech processing like phonology or semantics (Bernstein, Lu, & Jiang, 2008; Myers, Blumstein, Walsh, & Eliassen, 2009; Visser & Lambon Ralph, 2011).

In short, this study helps reveal a new aspect of AV speech processing where a network of areas is engaged in parallel, and comprises both unisensory and heteromodal regions. Remarkably, the listener's input language and the language dominance modulates the interplay between these areas, so the network is biased towards visual input. Future research is needed to examine the potential functional differences of these regions for L1 and L2 in accordance with proficiency (high. *vs.* Low-proficient bilinguals) and/or at other levels of speech processing (e.g., word level tasks, phonological tasks).

Acknowledgments

This research has been supported by the *Spanish Ministry of Science and Innovation* (PSI2010-15426, PSI2010-20168, and Consolider INGENIO CSD2007-00012), *Comissionat per a Universitats i Recerca del DIUE-Generalitat de Catalunya* (SRG2009-092), and the European Research Council (StG-2010 263145).

References

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica, 128*(3), 466–478.
Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*(3–4), 559–571.
Beauchamp, M. S. (2005). Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics, 3*(2), 93–113.
Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). FMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30*, 2414–2417.
Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping, 31*(4), 526–538.
Bernstein, L. E., Lu, Z. L., & Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research, 1242*, 172–184.
Bischoff, M., Walter, B., Blecker, C. R., Morgen, K., Vaitla, D., & Sammera, G. (2007). Utilizing the ventriloquism-effect to investigate audio–visual binding. *Neuropsychologia, 45*(3), 578–586.
Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory and visual stimulus onset asynchrony detection. *The Journal of Neuroscience, 21*(1), 300–304.
Callan, D., Callan, A., Gamez, M., Sato, M. A., & Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage, 51*(2), 844–858.
Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*(12), 1110–1123.
Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593–596.
Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 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*(2), 427–438.
Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B-Biological Sciences, 363*(1493), 1001–1010.
Colin, C., Radeau, M., Soquet, A., Dachy, B., & Deltenre, P. (2002). Electrophysiology of spatial scene analysis: The mismatch negativity (MMN) is sensitive to the ventriloquism illusion. *Clinical Neurophysiology, 113*(4), 507–518.
Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115*(9), 1989–2000.
Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–Macdonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113*(4), 495–506.
Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ant task. *Cognition, 106*(1), 59–86.
Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on sensory-specific' brain regions, neural responses, and judgments. *Neuron, 57*(1), 11–23.
Forman, S., Cohen, J., Fitzgerald, M., Eddy, W., Mintun, M., & Noll, D. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med., 33*, 636–647.
Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25*, 5004–5012.
Goebel, R., & van Atteveldt, N. (2009). Multisensory functional magnetic resonance imaging: A future perspective. *Experimental Brain Research, 198*(2–3), 153–164.
Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America, 119*(3), 1740–1751.
Hertz, U., & Amedi, A. (2010). Disentangling unisensory and multisensory components in audiovisual integration using a novel multifrequency fMRI spectral analysis. *Neuroimage, 52*(2), 617–632.
Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language and Hearing Research, 53*(2), 298–310.
Indefrey, P. (2006). A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language Learning, 56*, 279–304.
Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology, 15*(21), 1943–1947.
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. *Experimental Brain Research, 166*(3–4), 289–297.
Lee, H. L., & Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *Journal of Neuroscience, 31*(31), 11338–11350.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature Neuroscience, 264*(5588), 746–748.
Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25*(25), 5884–5893.
Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science, 20*(7), 895–903.
Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research-Psychologische Forschung, 71*(1), 4–12.
Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Res., 1323*, 84–93.
Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25*(2), 333–338.
Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., et al. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: An fMRI study at 3 T. *NeuroImage, 29*(3), 797–807.
Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd, & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading.* Hillsdale: LEA.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147–1153.
Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology, 19*(8), 661–667.
Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25*(1), 76–89.
Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92*, B13–B23.
Soto-Faraco, S., Navarra, J., Vouloumanos, A., Sebastián-Gallés, N., Weikum & Werker, J. F. (2007). Discriminating languages by speechreading. *Perception & Psychophysics, 69*(2), 218–231.
Stevenson, R. A., VanDerKlok, R. M., Pisoni, D. B., & James, T. W. (2010). Discrete neural substrates underlie complementary audiovisual speech integration processes. *NeuroImage, 55*(3), 1339–1345.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212–215.
van Atteveldt, N. M., Blau, V. C., Blomert, L., & Goebel, R. (2010). fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neuroscience, 11*(11).
Van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the integration of letters and speech sounds. *Cerebral Cortex, 17*(4), 962–974.
van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*(2), 271–282.
van Atteveldt, N., Roebroeck, A., & Goebel, R. (2009). Interaction of speech and script in human auditory cortex: Insights from neuro-imaging and effective connectivity. *Hearing Research, 258*(1–2), 152–164.
van Heuven, W. J. B., & Dijkstra, T. (2010). Language comprehension in the bilingual brain: fMRI and ERP support for psycholinguistic models. *Brain Research Reviews, 64*(1), 104–122.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186.
Visser, M., & Lambon Ralph, M. A. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience, 23*(10). 3121–1331.
Wang, Y., Xiang, J., Kotecha, R., Vannest, J., Liu, Y., Rose, D., et al. (2008). Spatial and frequency differences of neuromagnetic activities between the perception of open- and closed-class words. *Brain Topography, 21*(2), 75–85.
Wang, Y., Xiang, J., Vannest, J., Holroyd, T., Narmoneva, D., Horn, P., et al. (2011). Neuromagnetic measures of word processing in bilinguals and monolinguals. *Clinical Neurophysiology, 122*(9), 1706–1717.
Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316*(5828), 1159.
Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *Journal of Neuroscience, 30*(7), 2662–2675.