Neural correlates of audiovisual speech processing in a second language

1* Alfonso Barrós-Loscertales, 1*Noelia Ventura-Campos, 2,1* Maya Visser, 3Agnès Alsius, 4Christophe Pallier, 1Cesar Avila, and 2,5 Salvador Soto-Faraco

1 Dept. de Psicologia Básica, Clínica y Psicobiología, Universitat Jaume I, Castellón, Spain
2 Dept. de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain
3 Dept of Psychology, Queens University, Ontario, Canada
4 INSERM-CEA Cognitive Neuroimaging Unit, Neurospin center, Gif-sur-Yvette, France
5 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*The first three authors have contributed equally to the current study

Figures: 4
Tables: 4
Pages: 29
Words Abstract: 203

Correspondence to: Corresponding author: Maya Visser, department of Psicologia Básica, Clínica y Psicobiología, Universitat Jaume I, Avda. Sos Baynat s/n, 12071, maya.visser@gmail.com, phone number: +349582246789, fax number: +34964729267

Acknowledgements: This research was 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).

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
Abstract
Neuroimaging studies of audiovisual (AV) speech processing have exclusively addressed the listeners’ native language. Yet, several behavioural studies now show that AV processing plays an important role during non-native speech perception. The current fMRI study examined audiovisual speech processing in the second language of bilingual (Spanish/English) speakers. Participants were presented with auditory, visual, audiovisual congruent and audiovisual incongruent utterances in their native (L1) and non-native (L2) language. A first analysis restricted to multisensory regions whose response to the AV congruent condition was higher than either unimodal condition highlighted the posterior superior temporal sulcus (pSTS) in L1 as well as in L2. Differences in AV processing as a function of language background were not expressed in this region. However, unimodal regions in the occipital lobe had a stronger response to congruent than to incongruent AV speech in the non-native language, as compared to native. This fits well with a framework in which pSTS integrates the multimodal input in close collaboration with unimodal regions. According to these results, language background differences are predominantly expressed in these unimodal regions, whereas multisensory area pSTS is equally involved in AV integration regardless of language proficiency (L1 or L2).

Keywords: Audiovisual speech, bilingualism, fMRI
1 Introduction

Audiovisual (AV) binding is an integral aspect of language processing during natural face to face conversations as well as in modern media such as TV, cinema, or video-conferencing. Not only visual cues can strongly support the perception of auditory speech when they are correlated (especially under noisy environments; e.g., Sumby and Pollack, 1954; Ross et al., 2006), but when seen articulatory gestures do not correspond with acoustic speech they can dramatically alter auditory perception even under optimal hearing conditions (e.g., McGurk and MacDonald, 1976). The neural correlates underlying audiovisual integration of speech have been addressed oftentimes using fMRI, MEG and EEG (Callan et al., 2010; Calvert, 2001; Calvert et al., 2000; Calvert et al., 2001; Colin et al., 2002a; Colin et al., 2004; Colin et al., 2002b; Miller and D'Esposito, 2005; Skipper et al., 2005; van Wassenhove et al., 2005). However, an important, yet barely investigated aspect of audiovisual speech integration relates to its contribution during second language comprehension.

Previous 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). Amongst other things, these results indicate that visual information can be decoded to some extent even in an unfamiliar (or non-native) languages. From a theoretical point of view, therefore, one can predict an involvement of visual speech and thereby, of audiovisual integration, in second language perception when visual cues are available to the listener. What is more, from an informational point of view, the potential gain on overall comprehension that arises from the integration of vision with speech sound tends to be larger the less information is available from sound (Sumby & Pollack, 1954), which is precisely the situation one finds when trying 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
phonological discrimination between non-native sounds which are otherwise undistinguishable on the basis of auditory cues alone (Navarra and Soto-Faraco, 2007; Reisberg et al., 1987; Hirata and Kelly, 2010) resulting an improvement in overall comprehension of L2 (though the contribution of AV integration in L2 perception is not always effective; Hazan et al., 2006). Moreover, a recent study showed that available visual mouth movements improved auditory L2 learning (Hirata and Kelly, 2010). This aligns well with the anecdotal observation that people tend to focus more on visual speech information (e.g., lip movement), when conversing in a second language. That the influence of visual speech is more important in L2 has been shown by a study of Wang et al. (2008). This study showed that (1) adding visual speech information to auditory speech resulted in improved phoneme perception in L2 but not L1 and (2) audiovisual integration as shown by a stronger McGurk effect in L2 compared to L1. This may suggest that perception of non-native speech is more influenced by visual speech, whereas native speech was dominantly auditory based.

Although the function of AV integration in second language processing has been investigated in some behavioural studies (cited above), its neural correlates are still largely unknown. As far as we know, the neuroimaging literature on second language processing has exclusively concentrated on unimodal speech, such as auditory speech comprehension and visual reading, but not on multimodal aspects (for reviews see; Abutalebi, 2008; Indefrey, 2006; van Heuven and Dijkstra, 2010). Unisensory literature has generally shown that overlapping brain regions often underlie unimodal language processes in the native (L1) and non-native (L2) language. However, L2 processing sometimes has been found to result in stronger activations in frontal and temporal regions, suggesting that more neural resources are taken up to accomplish the same task in L2 as compared to L1. This seems logical as information at the phonetic, syntactic and semantic level will be harder to extract and parse in L2 (Abutalebi, 2008).

---

1 The McGurk effect refers to the occasion when an audio [ba] and a visual [ga] result in the
In the current study, we address differences in the multisensory processing brain network for AV speech perception in L1 and L2. Based on previous unimodal literature, we expect that a similar network of brain regions underlies audiovisual speech integration in L1 and L2, but that the consequences of this integration processes may vary depending on language background (native or not). We expect that multisensory regions will 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 discussed above. In line with this, we would expect occipital regions to play an important role during these multimodal 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, suggesting that multisensory processes may involve the orchestration of a network that engages classical association areas as well as regions regarded as unisensory (e.g., Driver and Noesselt, 2008). For example, auditory regions have been found to respond to visual speech stimuli presented in silence (i.e., lipreading; Calvert et al., 1999; Miller and D’Esposito, 2005; Kayser et al., 2005).

Prior literature on AV speech integration (always in the native language) has identified some regions of interest. Most notably, a considerable amount of evidence has associated the posterior part of the superior temporal sulcus (pSTS) with audiovisual integration during language processing (for reviews see, Amedi et al., 2005; Beauchamp, 2005; Campbell, 2008). This cortical region responds to visual as well as 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 other, usually meaningful, stimuli). The enhancement effect has been highlighted as a key aspect defining the responses of the STS as well as other regions of multisensory integration (Beauchamp, 2005; Calvert et al., 2000; Campbell, 2008).
particularly convincing study was able to associate the pSTS with the phenomenology of audiovisual speech integration (Miller and D’Esposito, 2005). In Miller and D’Esposito’s study, AV synchrony was varied over time whilst subjects rated whether they perceived the audio-visual signals as fused or not. The pSTS did not respond during trials when audiovisual information was not perceived as a fused object, but it showed activity even for asynchronous stimuli that were nevertheless perceived as fused.

Multisensory responses in the pSTS has been shown for speech information at the semantic level (Beauchamp, 2005; Calvert et al., 2000; Stevenson et al., 2010), the phonological level (e.g., non-words: Miller and D’Esposito, 2005; letters: van Atteveldt et al., 2010) and time-varying stimulation in the non-speech domain (e.g., sinusoidal moving visual information aligning with sinusoidal auditory noise; Werner and Noppeney, 2010; Bischoff et al., 2007). Furthermore, activation in this region is dominantly bilateral (van Attenveld et al., 2010; Werner and Noppeney, 2010; Bischoff et al., 2007; Beauchamp, 2005), although some studies have reported unilateral left (Miller and D’Esposito; 2005; Calvert et al., 2000) or right (Stevenson et al., 2010) pSTS activation. However, the left and right STS might be functionally different. Miller & D’Esposito have shown that the left STS responds to audiovisual stimuli when perceived as fused, whereas the right STS shows a higher BOLD response when perceiving audiovisual stimuli as unfused. They did not provide an interpretation for this pattern and future research is required to investigate this laterality difference further. In addition, Calvert et al.(2000) suggested that the left and right STS might be involved in speech and non-speech stimuli, respectively. However, more recent studies on both speech and non-speech stimuli, generally seems to show a higher BOLD signal in the left STS for all stimulus types (van Attenveld et al., 2010; Werner and Noppeney, 2010; Bischoff et al., 2007; Beauchamp, 2005). To sum up, the literature above suggests that the bilateral pSTS is a critical region for audiovisual integration in (but not necessarily only) language, and that this pattern seems to be stronger in left pSTS.
Initial reports of multisensory enhancement considered those areas displaying blood oxygenation level dependent (BOLD) responses to bimodal speech stimuli that 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 some advantages (as it is safe against false positives from areas that contain separate populations of unisensory neurons). However, it can be overly conservative (Beauchamp, 2005; Goebel and van Atteveldt, 2009; Laurienti et al., 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 here we chose to use the max criterion, as described by Beauchamp (2005), to examine multimodal responses to L1 and L2 AV speech processing. He proposed that the multisensory response should be higher than the maximum of the unisensory responses. (For further reading on these and other issues see: Beauchamp, 2005; Goebel and van Atteveldt, 2009; Laurienti et al., 2005). For regions that survived the max criteria, we further explored the pattern of multisensory interaction by looking at no-linearity using the approach described by van Attenveldt et al. (2007). This measurement calculates the difference between the total percentage BOLD signal change of the AV condition and the unisensory conditions with the max response.

In addition to the logic based on the aditivity criterion, the congruency criterion has also been successful in revealing the regions associated with audiovisual processing (van Atteveldt et al., 2010; Calvert, 2001). The hypothesis behind this criterion is that if a region’s BOLD response is different for congruent than incongruent AV information, then this means that it is involved in some kind of multisensory integration. 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 (Campbell, 2008).
To sum up, there is a considerable body of imaging research on AV integration in the native language, and on bilingualism using unisensory auditory stimulation, but second language processing and AV speech integration has received no attention. Given prior behavioural studies, the potential for audio-visual 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 areas of study by investigating the neural correlates of audiovisual integration in L1 vs. L2. Our hypothesis is that similar regions will be involved in AV integration for L1 and L2. However, we hypothesize that the BOLD signal might be stronger in L2 and that L2 AV speech processing might rely more on the visual network, as visual information seems to be relatively more important in L2.

2 Methods

2.1 Subjects

Forty two bilingual volunteers (age range 20 to 46 years old) proficient in English and Spanish were included in the study. Half of the sample (n=21:10 females) spoke Spanish as their native language and English as their non-native language, and the other half (n=21: 9 females) spoke English as their native language and Spanish as their non-native language. By pooling two equivalent groups of participants with the reverse language dominance pattern we are able to cancel out possible group effects 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 as to their onset age of exposure to their second language, as assessed by a questionnaire of language use (Costa et al., 2008). Furthermore, groups did not differ about their L2 proficiency (self-ratings) in terms of comprehension, fluency, reading and writing skills (see details in Table
1). All participants were in good health without personal history of psychiatric or neurological diseases, and normal auditory acuity as well as normal or corrected-to-normal (visual lenses from VisuaStim, Magnetic Resonance Tech.) visual acuity. All they gave informed consent prior to participation in the study.

Table 1 about here

2.2 Stimuli and design
Stimuli comprised 5s long sentences presented auditorily only (A), visually only (V) or audiovisually (AV). The audiovisual stimuli could be audiovisually congruent (AVc) or audiovisually incongruent (AVi). The AVi stimuli were cross-dubbed so the sound track of one sentence was combined with a different visual sentence of the same duration. Equivalent sets of stimuli were generated in each test language (English and Spanish) for every single condition, from recordings of an English-Spanish well-balanced bilingual speaker. 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 a no-stimulus baseline condition (B) were presented in a blocked design-fashion, with block order pseudo-randomized (consecutive blocks of the same condition were avoided). Different version of the experiment ensured that each particular sentence could be presented in any display condition across participants. Each block was repeated four times, lasted 40 s and included eight stimuli of five seconds in duration. Run order was also counterbalanced between subjects. The stimuli were presented through visual and auditory MRI compatible systems (Visuastim, Resonance Technologies, Inc). Subjects were instructed to listen to each sentence and to keep
their attention on the screen (even during baseline conditions), since we informed them that they would be asked to perform a recognition test after scanning. In 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 a total of 16 trials; 1 target and 1 foil per condition (i.e., A, V, AVc and AVi) per language (L1 and L2). This was included just to ensure an attentive strategy during stimulus presentation (Calvert et al., 2000; Beauchamp, 2005).

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 T2*-weighted images depicting BOLD contrast were sparsely acquired over 10-min and 40-s with a TR of 8-s (TE=60-s, TA= 2s; flip-angle=90°, voxel-matrix=64x64; voxel-size= 3.94x3.94, 5-mm thick and 0.5-mm gap, 1 interleave). 25 coronal slices were acquired, perpendicular to the Sylvian fissure covering the whole brain. In our sparse sampling design, five 2 s volumes were acquired per block. The first volume was acquired three seconds after the onset of the stimuli. The following four volumes were acquired with six second gaps (hence a TR of eight seconds). 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 = 256x224x176).

2.4 FMRI data analysis

Preprocessing: Prior to time-series statistical analyses, the data from each subject was preprocessed using SPM5 (Welcome Department of Cognitive Neurology, London). 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 next co-registered to the mean image and segmented into gray and white matter partitions. Normalization parameters were extracted from the segmentation of each subject anatomical T1-weighted scan and applied to their corresponding functional scans (rescaled voxel size 3x3x3-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 analysis: Conditions of interest corresponding to A, V and AVc and AVi for both L1 and L2 were modeled using a box-car function. Low frequency drifts were removed using 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 during the realignment step of 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 are collapsed across language dominance groups (i.e., the English and Spanish native speakers; see Table 1). For each contrast, we first ensured that there were no significant differences between the two groups. In addition, we later checked whether the pattern that arises from this contrast also holds when examining the groups separately. Finally, 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 during the audiovisual congruent condition (AVc) compared to the unimodal (A and V) conditions at L1 and L2 group level we performed the conjunction of $[(AVc>A) \cap (AVc>V) \cap (A>B) \cap (V>B)]$ in which B refers to the baseline or
´rest´ condition. This contrast is referred to as the “max criterion” and is commonly applied in multisensory research (van Atteveldt et al., 2007; Beauchamp, 2005). The result from this conjunction gave a statistical value for each voxel as the minimum of the statistical values obtained from the four included contrasts (van Atteveldt et al., 2004). Clusters were considered as significant whenever they showed a significant interaction at p<0.001, uncorrected and a cluster size of 8 adjacent voxels (216 mm$^3$). Van Atteveldt et al. (2007) used a multisensory interaction (MSI) measure to visualize the multisensory enhancement in region-of-interests (ROIs) based on the functional data. This measurement calculates the difference between the total percentage BOLD signal change of the AV condition and the unisensory conditions with the max response. Van Atteveldt et al., used the total percentage 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 functional ROIs as clusters that 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 will use 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 & Noppeney (2011). We first looked for any regions showing a supra- or sub-additive pattern using the contrasts (AVc) $>$ (A + V) and (AVc) $<$ (A + V). Then, we checked whether the max criterion applies to these regions, by using $[(AVc>A) \cap (AVc>V) \cap (A>B) \cap (V>B)]$ and $[(AVc<A) \cap (AVc<V) \cap (A>B) \cap (V>B)]$. We examined these contrasts 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 audiovisual congruency, as compared to audiovisually incongruent stimulation. In this case, we directly compared synchronized (AVc) with mismatched (AVi) audiovisual speech. This contrast is interesting because the two terms contain equivalent amounts of sensory input in each modality, and they only differ in 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 Behavioral results of the recognition test during the scanning session

The participants performed a recognition test after the scanning session (see Methods). This test included too few observations per condition and it does not directly inform about online comprehension, since it was mainly included to ensure the participants would remain 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. (Data from two participants were missing due to experimenter error). The average recognition results are presented in Table 2. A two (language; L1 and L2) by four (modalities; A, V, AVc, AVi) ANOVA revealed that there was a significant effect of language ($F(1) = 6.9, p = .012$) and modality ($F(3) = 10.0, p < .001$), but the interaction was not significant ($F(3) = .336, p = .8$). The overall mean of L1 ($M= 0.67, SD = 0.34$) was higher compared to L2 $M= 0.57, SD = 0.37$. This pattern of performance indicates that participants could recognize words in the
sentences from their native language slightly better than the non-native language, but that all had a good understanding of the sentences in the non-native language. Our aim was to make sure they would comprehend and therefore attend the stimuli in both languages. Small differences at the level of comprehension probably makes it more likely to find differences in the neural network, and therefore do not constitute a problem for us.

*Table 2 about here*

### 3.2 Neural correlates of multisensory integration

As described in the method section, we used the max criterion to reveal multisensory regions involved in audiovisual speech processing for L1 and L2. This criterion requires that the response to the AV condition is higher than the maximum response in the unimodal conditions. The clusters that survived the max criterion are presented in Table 3 and Figure 1, at a corrected level for multiple comparisons (using Monte Carlo simulations in AFNI). For both L1 and L2, we found bilateral activation of the posterior superior temporal sulcus (pSTS). We further examined the mean percent BOLD signal change of the different conditions that make up the max criterion (e.g., audio, visual and audiovisual congruent). These are presented in Figure 1B. Van Atteveldt et al (2007) used a multisensory interaction (MSI) measure to calculate the multisensory enhancement (see Methods). The MSI values of the current study are presented in Figure 1C. A paired-sampled t-test did not show significant differences in MSI between L1 and L2 in both hemispheres \((t(40) = .16; p = .88\) and \((t(40) = 1.0; p = .32\) for the left and right hemisphere respectively). This is in line with research suggesting 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 the neural responses to multisensory integration, we ran a second set of analyses following a
method recently used by Lee and Noppeney (2011). These analyses first test whether any regions showed a sub- or supra-additive pattern. As a second step, the max or minimum criterion is used \([(AVc>A) \cap (AVc>V) \cap (A>B) \cap (V>B)] \) or \([(AVc<A) \cap (AVc<V) \cap (A>B) \cap (V>B)] \) (see Method section). Only one contrast showed a significant result, namely the sub additive contrast for L1. This showed the bilateral inferior frontal gyrus (MNI coordinates; 54, 15, 27 and -54, 21, 27; corrected for multiple comparisons using Monte Carlo simulations in AFNI) a region that is commonly associated with multisensory processing, including the processing of AV speech (Lee & Noppeney, 2011; Calvert 2001; Campbell, 2008). However, since it did not respond to the follow-up contrast, we cannot make a strong claim on the function of this region. Further research is required to investigate the function of this region with respect to AV speech in bilinguals. Note that the pSTS did not come out significant for the interactive effect, that is, we cannot assume a pattern beyond additive.

*Table 3 about here*

*Figure 1 about here*

### 3.3 Congruency effects

We further examined regions that responded to the congruency contrast within each language dominance condition (L1 and L2). The multisensory regions that had been highlighted by the max criteria in the previous analyses did not respond significantly to this contrast. Instead, the contrast \(AVc > AVi\) in L2 resulted in the significant activation of two clusters in visual areas, the right middle occipital lobe (BA 18/19) and the left lingual gyrus (BA 17/18). This is a classically defined unisensory area. Figure 4 shows the location of the regions and the percentage BOLD signal change for the congruent and incongruent conditions in L1 and L2.
The percentage BOLD signal change of both visual clusters showed a significant congruency effect and a significant language by congruency interaction. (For cluster -24 -92 -3, (F(1) = 12.66; p = 0.001) and (F(1) = 9.98; p = 0.003), respectively. For cluster 30 -83 -11, (F(1) = 18.94; p < 0.001) and (F(1) = 8.80; p = 0.005), respectively.

Table 4 about here

Figure 2 about here

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 have collapsed across language dominance groups (i.e., the English and Spanish native speakers; see Table 1 in the Method section). This introduces the desirable feature that none of the effects observed could be due just to between-group differences (all subjects contributed to L1 and L2 BOLD) not to particular aspects of the stimulus language (both English and Spanish stimuli played the role of L1 and L2). However, just to confirm our results, we repeated all analyses for each language group separately (Spanish and English speakers) at an uncorrected level of p < 0.001. The results for each group remained the same as for the collapsed analyses presented in above,
Furthermore, we checked whether the length of Exposure to L2 or the experience speaking L2 modulated some of the L2 effects here. We introduced these factors as a covariate in an ANCOVA. The results of the max criterion as well as the congruency criterion did not change, indicating that these factors do not play a significant role in the current study. This is not surprising, since our participants sample was intended to be as homogeneous as possible in these factors.

4 Discussion

The present study set out to examine the neural correlates of audiovisual (AV) speech integration in second language perception. We first targeted multisensory regions that displayed enhancement effects to AV congruent stimulation in comparison to unisensory stimulation. We found that the pSTS was activated for AV speech processing in the native as well as in non-native language. This is in line with previous audiovisual speech research, with the bilateral posterior superior temporal sulcus (pSTS) being the most frequently identified region (Beauchamp, 2005; Calvert, 2001; Calvert et al., 2000; Calvert et al., 2001; Campbell, 2008; Goebel and van Atteveldt, 2009; Miller and D'Esposito, 2005). Moreover, in our case we found that the BOLD enhancement was equivalent in both language background conditions indicating that the AV integration in the bilateral pSTS underlies the same function in processing L1 and L2. Secondly, we found that the BOLD response in the occipital lobe responded differentially to congruent vs. incongruent stimulation, depending on the language status of the stimulus for the participant. The fact that this unimodal region is influenced by multimodal input suggests their close collaboration with the AV integration system. This finding is in line with previous research revealing a set of sub-networks for dissociable components of AV speech integration (Driver and Noesselt, 2008; Skipper et al., 2009; Hertz
& Amedi, 2010). In the following, we discuss the implications of these findings for the characterization of audiovisual speech processing in the second language.

Multisensory speech integration

Our experiment identified AV integration regions during second language processing. We used a standard criterion requiring that the BOLD signal in an AV region to be higher during AV input compared to the maximum of the two unimodal (visual and auditory) inputs (Calvert, 2001; Calvert et al., 2001; Beauchamp, 2005), and revealed overlapping regions in the bilateral pSTS involved in AV integration for L1 and L2. To our knowledge, this is the first study that links the pSTS with AV processing in a second language. Previous studies of auditory speech perception on 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 et al., 2010). Given this previous result, we examined whether there was a possible difference in the percentage BOLD signal change in the pSTS across language dominance. Yet, the results showed that the percentage BOLD signal change in the pSTS was equivalent for both language modalities. In sum, the present results are clear in that an equivalent or very similar integration system in the bilateral pSTS underlies multisensory gains for speech in L1 and L2 in high proficient bilinguals. Future research will need to verify whether this result holds for low proficient bilinguals as well. Further analyzing the nature of this multisensory response, we did not find evidence of any interaction pattern beyond (or below) additive. This means that, for this particular case, the response of the pSTS was additive, both for L1 and L2.

Congruency effects
The congruency contrast is used to examine regions that respond differently on AV congruent compared to incongruent information. Unimodal regions in the occipital lobe did show differential activity response to this contrast. Instead, the multisensory region identified with the enhancement criterion (pSTS) did not respond selectively to this congruency contrast. Although the congruency criterion has been used to identify multimodal regions, it is important to note that not all multisensory regions might be sensitive to this contrast (Campbell, 2008). In particular, in some previous studies, just like it was the case here, the bilateral pSTS failed to respond to the congruency manipulation (Bushara et al., 2001; Miller and D'Esposito, 2005; Ojanen et al., 2005). In addition, other studies found a higher BOLD response to the incongruent compared to the congruent condition (Benoit et al., 2010; Pekkola et al., 2006) or vice versa (Calvert et al., 2000; van Atteveldt et al., 2004; van Atteveldt et al., 2009; note that Calvert et al. only found left pSTS activation). Therefore, the responsiveness of the bilateral pSTS to congruency is still unclear.

The important finding of the current congruency analysis is that a region traditionally considered as unimodal was responsive to audiovisual congruency (vs. incongruency) of multimodal speech information, as a function of which language (L1 or L2) was being processed. This is a rather interesting finding, and in line with literature suggesting that unimodal regions are influenced by multimodal processes (Driver and Noesselt, 2008; Hertz & Amedi, 2010). Our results provide additional insight in this region by showing that the type of audio-visual pairing partly determines the responsiveness of this region. In particular, when looking at the congruency effects within L2, the active clusters concentrated in the occipital lobe. More precisely, the right middle occipital gyrus (BA 18/19) and the left lingual gyrus (BA 17/18). Occipital regions are involved in visual processing and our data suggest that it has a relatively more important role in AV speech processing in L2. Behaviourally, 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 and Soto-Faraco, 2007; Wang et al., 2008). For example, Wang et al., (2008) compared 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 could extract sufficient information from the auditory signal (and could, therefore, rely more in audition). Interestingly, Wang et al. (2008) measured the perceptual illusion created with AV incongruent syllables, called the McGurk\(^2\) effect and found that they were more pronounced in the Mandarin speakers compared to the native English group, but solely for phonemes that did not exist in mandarin. These results suggest that specially for difficult foreign sounds, the bilinguals tended to make comparatively greater use of visual speech information in L2 (see also Navarra & Soto-Faraco, 2007). Note that these visual regions did not survive the max criterion (looking for enhancement effect) possibly because this criterion required (positive) response to either sensory modality in isolation, and this region did not reveal a significant response to auditory speech alone. Altogether, this pattern suggests that multisensory regions (pSTS) play a modulatory role on the responsiveness of these unisensory area during AV processing.

Overall, the present results strongly suggest that regions sensitive to multisensory speech input such as the bilateral pSTS collaborate closely with unimodal regions that are recruited for the task. This is not a new idea, since it has been proposed in the past that multisensory processing is carried out through the interplay between association (heteromodal) regions and regions traditionally considered to be unisensory (Campbell, 2008; Driver and Noesselt, 2008). What is new in our current findings is that this interplay engages different parts of the network to varying strengths depending on the language background of the speaker / stimuli (native vs. non-native). We would like to emphasize that it is not our

---

\(^2\) 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 and MacDonald, 1976).
claim that this network is specific for speech, since it may well play a role in non-speech stimuli as well (Ghazanfar et al., 2005; Campbell, 2008). What we contend is that it will has a paramount function during speech processing, and that the native vs. non-native nature of the language being processed seems to attune the working of the network in different ways.

Conclusions and future research

We investigated the neural correlates of brain regions involved in audiovisual speech processing when bilinguals use their native vs. their second language. The results showed that the pSTS is involved in audiovisual processing in L1 and L2 to a similar extent in our study (testing high proficient bilinguals at sentence comprehension level). What is clear from our results is that, closely relating to previous behavioral studies showing the effects of AV integration in L2, similar neural responses to AV speech integration were shown for second as well as first language. In addition, clusters in the occipital lobe were dominantly associated with L2 as compared to L1 AV speech processing. The fact that these unimodal regions respond to multimodal stimulation reveals, in our view, a modulatory effect arising from the interactivity between unimodal and multimodal components of the multisensory processing network. In our case, these modulatory effects seem to reflect a stronger reliance on visual processing when perceiving L2. Future research will need to investigate how this multimodal network interacts with brain regions specifically involved in particular aspects of speech processing like phonology or semantics, such as the Supramarginal Gyrus, posterior Superior Temporal Gyrus, Broca’s and the Anterior Temporal Lobe (Bernstein et al., 2008; Myers et al., 2009; Visser & Lambon Ralph, 2011).

In sum, this study helps reveal a new aspect of AV speech processing where a network of areas is engaged in parallel, comprising both unisensory and heteromodal regions. Remarkably, the input language and language dominance of the listener will modulate the interplay between these areas, so that the network is biased toward visual input.

Future
research will need to examine potential functional differences of these regions for L1 and L2 as a function of proficiency (high vs. low proficient bilinguals) and/or at other levels of speech processing (i.e., word level tasks, phonological tasks).
References


Figure 1. Regions involved in multimodal speech processing in the native (L1) and non-native language (L2). A) The bilateral superior temporal sulcus (STS) was identified as a region responding to audiovisual stimulation using the max criterion: AVc conditions resulted in higher activation that the max response of the unimodal conditions. The figure presents a response to AV integration 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 audiovisual congruent (AVc) information. These conditions are used in the max criterion. The max criterion tests whether a region shows a higher response to the AVc condition as opposed to the max of the unimodal conditions. Error bars represent the standard error of the mean. C) The multisensory interaction (MSI) values were calculated in the pSTS clusters that survived the max criterion (see Methods). For the MSI, the bimodal response was calculated relative to the most effective unimodal response (van Atteveldt et al., 2007). Language background differences did not result in significant differences of this value. Error bars represent the standard error of the mean.

Figure 2. Regions responding to AV congruent as opposed to AV incongruent sentences during L2 processing. Clusters survived a corrected family-wise error rate of p>0.05, defined by Monte Carlo simulations using the AFNI program Alphasim.
Table 1. Demographic details of the two bilingual groups.

<table>
<thead>
<tr>
<th></th>
<th>Spanish L1</th>
<th>English L1</th>
<th>Between-group differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>25.29 (6.05)</td>
<td>28.90 (9.49)</td>
<td>t(40) = 1.47, p = .15</td>
</tr>
<tr>
<td>Lateralization (right/ left/</td>
<td>17/1/3</td>
<td>19/1/0</td>
<td>t(40) =.00 , p = 1.00</td>
</tr>
<tr>
<td>bimanual)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 language talk age</td>
<td>10.61(4.96)</td>
<td>16.52(7.93)</td>
<td>t(38) = - 1.41, p = .17</td>
</tr>
<tr>
<td>L2 Self-rated proficiency (1/best to 4/worst)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td>1.62(.50)</td>
<td>1.86 (.86)</td>
<td>t(40) =-1.10 , p = .28</td>
</tr>
<tr>
<td>Reading</td>
<td>1.43(.51)</td>
<td>1.80 (.83)</td>
<td>t(40) = -1.73 , p = .09</td>
</tr>
<tr>
<td>Fluency</td>
<td>1.71(.46)</td>
<td>2.05(.94)</td>
<td>t(40) = -1.46, p = .15</td>
</tr>
<tr>
<td>Writing</td>
<td>1.67(.58)</td>
<td>1.95 (.89)</td>
<td>t(39) = -1.22, p = .23</td>
</tr>
</tbody>
</table>

Table 2. Means and standard deviations of the recogntion test. A
<table>
<thead>
<tr>
<th></th>
<th>Conditions</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Native</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Auditory</td>
<td>0.60</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Visual</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Audiovisual congruent</td>
<td>0.76</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Audiovisual incongruent</td>
<td>0.80</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>All conditions collapsed</td>
<td>0.67</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Auditory</td>
<td>0.54</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Visual</td>
<td>0.44</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Audiovisual congruent</td>
<td>0.63</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>Audiovisual incongruent</td>
<td>0.66</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>All conditions collapsed</td>
<td>0.57</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 2. Means and standard deviations of the recognition test. After the scanner the participants performed a recognition test. They were presented with sentences and they had to press a button if they were of the opinion that they has seen the sentence during the experiment.
Table 3: Location of main activation clusters after applying the max criterion analysis to L1 and L2.

<table>
<thead>
<tr>
<th>Brain region</th>
<th>BA</th>
<th>TAL coordinates</th>
<th>T</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>Y</td>
<td>z</td>
</tr>
<tr>
<td>L1 R. Superior Temporal Gyrus</td>
<td>41</td>
<td>53</td>
<td>-34</td>
<td>10</td>
</tr>
<tr>
<td>L. Superior Temporal Gyrus</td>
<td>13</td>
<td>-45</td>
<td>-46</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>-53</td>
<td>-40</td>
<td>8</td>
</tr>
<tr>
<td>L2 R. Superior Temporal Gyrus</td>
<td>22</td>
<td>56</td>
<td>-37</td>
<td>13</td>
</tr>
<tr>
<td>L. Superior Temporal Gyrus</td>
<td>13</td>
<td>-45</td>
<td>-46</td>
<td>13</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Brain region</th>
<th>BA</th>
<th>TAL coordinates</th>
<th>T</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>AVcL2&gt;AViL2</td>
<td>R. Middle Occipital Gyrus</td>
<td>18/19</td>
<td>30</td>
<td>-82</td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>L. Mammillary Body</td>
<td>0</td>
<td>12</td>
<td>-7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L. Lingual Gyrus</td>
<td>17/18</td>
<td>-21</td>
<td>-87</td>
<td>-1</td>
</tr>
</tbody>
</table>

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

ALL TABLES SHOULD BE FORMATTED CORRECTLY. (IE AT LEAST LIKE TABLE 1, OR BETTER)