Top-down attention regulates the neural expression of audiovisual integration

Luis Morís Fernández [a*]

Maya Visser [b*]

Noelia Ventura Campos [b]

César Ávila Rivera [b]

Salvador Soto-Faraco [a, c]

[a] Multisensory Research Group, Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain.

[b] Departament de Psicología Básica, Clínica y Psicobiología, Universitat Jaume I, Castelló de la Plana, Spain.

[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

* Both authors contributed equally to this work.

Corresponding author: Luis Morís Fernández

Dept. de Tecnologies de la Informació i les Comunicacions
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
Spain

+34 686 17 30 58

luis.moris.fernandez@gmail.com

Abstract

The interplay between attention and multisensory integration has proven to be a difficult question to tackle. There are almost as many studies showing that multisensory integration occurs independently from the focus of attention as studies implying that attention has a profound effect on integration. Addressing the neural expression of multisensory integration for attended vs. unattended stimuli can help disentangle this apparent contradiction. In the present study, we examine if selective attention to sound pitch influences the expression of audiovisual integration in both behavior and neural activity. Participants were asked to attend to one of two auditory speech streams whilst watching a pair of talking lips that could be congruent or incongruent with the attended speech stream. We measured behavioral and neural responses (fMRI) to multisensory stimuli under attended and unattended conditions while physical stimulation was kept constant. Our results indicate that participants recognized words more accurately from an auditory stream that was both attended and audiovisually (AV) congruent, thus reflecting a benefit due to AV integration. On the other hand, no enhancement was found for AV congruency when it was unattended. Furthermore, the fMRI results indicated that activity in the superior temporal sulcus (an area known to be related to multisensory integration) was contingent on attention as well as on audiovisual congruency. This attentional modulation extended beyond heteromodal areas to affect processing in areas classically recognized as unisensory, such as the superior temporal gyrus or the extrastriate cortex, and to non-sensory areas such as the motor cortex. Interestingly, attention to audiovisual incongruence triggered responses in brain areas related to conflict processing (i.e., the anterior cingulate cortex and the anterior insula). Based on these results, we hypothesize that AV speech integration can take place automatic only when both modalities are sufficiently processed, and that if a mismatch is detected between the AV modalities, feedback from conflict areas minimize the influence of this mismatch by reducing the processing of the least informative modality.

Keywords

Multisensory; Audiovisual; Attention; Speech Perception; fMRI; STS

1. Introduction

Almost every event in our everyday life environments engages more than one sensory system at a time. This information, received across the different sensory pathways, is integrated to form unified multisensory objects allowing for a more efficient representation of the external world (G. Calvert, Spence, & Stein, 2004). A prime example of multisensory integration (henceforth referred to as MSI) is speech perception, whereby visual speech cues are extracted from the sight of a speaker's facial gestures and combined with auditory information. Audiovisual (AV) integration of speech has been shown to lead to improvements in understanding, especially under noisy circumstances and in persons with poor hearing (e.g., Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). Moreover, the tendency to integrate AV information is so strong that, when visual and auditory inputs are set in conflict, they can lead to dramatic illusions arising from the fusion between the two modalities, such as the famous McGurk effect (McGurk & MacDonald, 1976). Multiple brain sites responsive to integration have been described in past literature, both in and outside the domain of speech. Regarding the former, these various brain regions have been posited to conform a network that includes classical association brain areas as well as auditory and visual sensory cortices (M. S. Beauchamp, 2005; G. A. Calvert, 2001; Jon Driver & Noesselt, 2008a; Fairhall & Macaluso, 2009).

One of the current debates in MSI is to determine to which degree these sensory integration processes happen independently of the observer's focus of attention and intentions, or if attention is a requisite for integration (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014; Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007; Alsius & Soto-Faraco, 2011; Andersen, Tobias, Tiippana, Laarni, Kojo, & Sams, 2009; Bertelson, Vroomen, de Gelder, & Driver, 2000; Buchan & Munhall, 2011, 2012; Fairhall & Macaluso, 2009; Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006; Senkowski, Talsma, Herrmann, & Woldorff, 2005; Soto-Faraco, Navarra, & Alsius, 2004; Tiippana, Puharinen, Möttönen, & Sams, 2011; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; van Ee, van Boxtel, Parker, & Alais, 2009; Vroomen, Bertelson, & de Gelder, 2001. For reviews see: Navarra, Alsius, Soto-Faraco, & Spence, 2010; Koelewijn, Bronkhorst, & Theeuwes, 2010; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). This question is very relevant because our normal, everyday life environment produces far too many inputs to be fully processed by our senses. Some of these inputs from different modalities will correspond to a common event (i.e., the voice and lips of our conversation partner) and some to completely unrelated sources (i.e., the voice of another person, the sight of a passing car, music…). Thus, the question is: Do the benefits arising from MSI and their neural expression occur when our focus of attention is away from the relevant corresponding inputs? The literature addressing the behavioral correlates of MSI contains widely contrasting approaches and answers to this question.

When using low-level stimuli, such as beep and flash, one of the main stands is that MSI occurs independently of the focus of attention or the attentional manipulation made, it being exogenous or endogenous (Bertelson et al., 2000; Vroomen et al., 2001). Furthermore, some studies not only claim that MSI is immune to attentional effects, but also that the outcome of MSI can summon participants' attention automatically, like in the "Pip and Pop" effect (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; although see Alsius et al., 2010; Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006 for contradictory findings).

The role of attention in MSI has also been an important matter of debate in the specific domain of speech (for reviews see: Koelewijn et al., 2010; Navarra et al., 2010; Talsma et al., 2010). AV speech integration seems to be vulnerable to diverted attention conditions (Alsius et al., 2005, 2007; Tiippana, Andersen, & Sams, 2004; Tiippana et al., 2011; Zion Golumbic, Cogan, Schroeder, & Poeppel, 2013) or to visually crowded scenarios (Alsius & Soto-Faraco, 2011). A recent study by Nahorna and colleagues (2012) revealed that the strength of the McGurk illusion can decrease when the preceding AV context is incongruent. Another study showed that this illusion can be nearly eliminated under hypnotic suggestion (Déry, Campbell, Lifshitz, & Raz, 2014), indicating the malleability of MSI by endogenous factors under some circumstances. However, other studies have highlighted the fact that AV speech integration can be rather unavoidable, and therefore automatic and resilient, even when the relevant stimuli are outside the focus of attention (Driver 1996; Soto-Faraco & Alsius 2004).

This initially simple question has resulted in a mixed pattern of results revealing the complexity underlying the interplay between attention and integration. A paramount contribution to this debate is to understand not only the behavioral consequences of these attentional manipulations, but also their neural expression, especially on the network of brain areas typically involved in MSI. This is precisely the aim of the present study.

Neuroimaging studies measuring attentional effects on AV speech integration

Consistently with the multifaceted nature of the interplay between MSI and attention, it has previously been shown that attentional manipulations of AV integration lead to changes in neural responses to multisensory events at multiple stages and in a variety of brain regions (Fairhall & Macaluso, 2009; Senkowski et al., 2005; Talsma & Woldorff, 2005; Zion Golumbic et al., 2013).

For example, Zion Golumbic et al. (2013) addressed the interaction of attention and visual speech on auditory speech processing using magnetoencephalography (MEG). They presented participants with two auditory messages (both originating from a central location) and two speaking faces (one left and one right), each matching one of the voices. Participants were asked to track one auditory message (voice) and to ignore the other. Zion Golumbic et al. calculated a linear temporal response function that allowed them to estimate the neural response based on the speech signal, and more specifically, to discriminate which of the two signals, attended or ignored, had a larger contribution. This temporal response function revealed a larger contribution of the attended speech signal when compared to the ignored one, indicating that the neural response was more related to the attended speech signal, and had a stronger representation of the attended track in the auditory cortex. What is more: This difference in amplitude was contingent on the visual information, as it did not appear when only auditory information was presented.

In their 2009 study, Fairhall and Macaluso also studied the influence of attention on AV integration using fMRI. In the study, participants were presented with two pairs of speaking lips from different spatial locations (left and right) together with one single auditory speech stream that matched only one pair of lips. Two main findings arose from this study. The first one was related to the superior temporal sulcus (STS), an area classically related to multisensory integration in and outside the speech domain

(Beauchamp, Lee, & Argall, 2004; Calvert, Campbell, & Brammer, 2000; Fairhall & Macaluso, 2009; Miller & D'Esposito, 2005; Nath & Beauchamp, 2012; Toemme Noesselt et al., 2007; Tömme Noesselt, Bergmann, Heinze, Münte, & Spence, 2012; Stevenson, Altieri, Kim, Pisoni, & James, 2010; Stevenson, VanDerKlok, Pisoni, & James, 2011; Stevenson & James, 2009). This study showed a higher BOLD response in the STS when participants focused their visual spatial attention on the lips that were congruent with the auditory stream than when they focused their attention towards the location of the incongruent lips. The second finding in Fairhall and Macaluso's work was that the influence of attention on responses to AV speech was reflected beyond classical heteromodal areas (such as STS). Indeed, attention also had an impact on responses from sensory areas (such as V1, V2) as well as in the fusiform gyrus and the superior colliculus. Previous literature already points out that MSI effects extend beyond heteromodal regions to areas traditionally regarded as unisensory (see Jon Driver & Noesselt, 2008b; Emiliano Macaluso & Driver, 2005; Schroeder & Foxe, 2005 for reviews on this subject), but this study adds to this by showing that attention modulates these expressions of MSI, and that it appears to also affect low-level areas such as V1.

Neuroimaging studies such as these provide important evidence to understand at which stage, or stages, the interaction between attention and MSI occurs, especially if we consider that the brain networks supporting MSI are complex and that the influence of attention can be orchestrated across several components of this network (Talsma et al. 2010). Using non-speech stimuli, Talsma & Woldorff (2005) reported that the gain in electrophysiological response to audio-visual stimuli, compared to unimodal ones, was greater if the bimodal stimulus occurred at an attended region of space than when the audio-visual compound appeared at an unattended region. Interestingly, Talsma & Woldorf found this modulatory effect of attention took place at multiple stages along the ERP signal, starting as early as 90 ms post stimulus and with the latest effect seen at 500 ms. To sum up, past literature suggests that the attentional effects while processing multisensory information take place in classical multisensory regions including, but not restricted to the above mentioned STS, inferior parietal lobe and superior colliculus (as shown in Fairhall & Macaluso, 2009, for example) as well as in unisensory regions (Zion et al., 2013). This possibly reflects that attention has an impact at multiple stages of multisensory processing (Talsma & Woldorff, 2005; Talsma et al. 2010).

Scope of the present study

The hypothesis of the present study is that attention to AV stimuli is necessary for integration to occur in its full strength. If our hypothesis is true, then we expect to see a modulation of the neural activity within the MSI network specifically in the STS when participants attend a congruent AV stimuli compared to when they attended an incongruent AV stimuli. Behaviorally one would expect an increment in the word recognition rate when attention is directed towards AV congruent stimuli as compared to when it is directed to AV incongruent stimuli. We also expect to be able to narrow down the possible mechanistic interpretations by inference from the brain regions in which the attentional modulation of AV integration expressed.

We used speech as it is a prime model for MSI, and used selective attention conditions akin to the cocktail party phenomenon (Cherry, 1953). We asked participants to attend to one of two speech streams (high pitch and low pitch) originating from the same central location, simultaneously with a close-up of a pair of lips that could be congruent with one of the two speech streams or with none of them. In the critical contrasts physical stimulation remained constant; one of the auditory streams always matched the lips in the screen, congruent, while the other one did not match the lips, incongruent. Therefore to create each condition we just manipulated the observer's focus of endogenous attention towards the congruent or incongruent auditory stimulus. To measure the behavioral effect of the interplay between attention and MSI, at the end of the trial participants were asked to recognize which of two words had appeared in the previous messages. In a first experiment we probed participants with words that appeared in the congruent or the incongruent stream and in the attended or unattended stream, this way we could measure the effect of MSI under unattended conditions. In a second experiment we addressed the effect of congruency under unattended conditions by measuring if an unattended auditory stream would have any differential effect if it was accompanied with congruent or incongruent visual information.

The present study introduces several important differences with respect to previous attempts at this question. First, it is important to extend the scope of previous studies beyond visual selection, using selective attention in other modalities such as audition. This is because it is far from trivial that a particular multisensory interaction will generalize across other possible modality pairings. Examples of such lack of generalization can be found in cross-modal attention literature (see Jon Driver & Spence, 1998). Specially relevant are the results by Alsius & Soto-Faraco (2011), using a cross-modal search task where participants had to detect a face-voice match. Alsius & Soto-Faraco reported that when using visual selective attention, search for AV congruency was serial, but if selection was auditory then face-voice match seem to pop-out.This suggests that the effect of MSI can be highly dependent on the modality in which selective attention takes place. Second, above and beyond the interest of generalizing to selection in other sensory modalities, the interest of using audition is to be able to detach the question of attention and MSI from spatial-attention paradigms (Fairhall & Macalusio, 2009; Zion Golumbic 2013). Here we use attention towards a purely auditory feature such as is pitch. This type of auditory selective attention has been previously described, in the context of the 'cocktail party' problem, as object-based attention (Shinn-Cunningham, 2008). The cocktail party paradigm has also been suggested to be ideal to study the effects of selective attention due to its high load conditions and the necesity to fully engage selective attentional processes to ignore the irrelevant input and correctly process relevant information (Hill & Miller, 2010; Lee, Larson, Maddox, & Shinn-Cunningham, 2014). This is of special relevance as the expected effect of attention on MSI seems to be stronger when it's tested in paradigms under high load conditions (see section 4. General Discussion). And finally, we believe that it is important to measure the neural and behavioral expression of any possible modulations within the same paradigm and task. This expands the results by Fairhall & Macaluso (2009) as the linguistic information (i.e. the auditory signal) was irrelevant to their behavioral task (visual detection). In the present paradigm (word recognition) we gauge the interplay between attention and MSI using the well known benefit due to AV integration, proven many times in behavioral literature.

2. General Methods

General methods information for all experiments is presented below. Each of the experiments following will contain their particularities.

2.1. Materials and apparatus

We recorded 72 passages (20 s duration each) from the novel “The Little Prince” by Antoine de Saint-Exupéry (Spanish translation) read by a female native Spanish speaker (Video resolution: 960x720 fps: 50; Audio sample rate: 48kH 16 bits Mono). The clips were edited with Adobe Premiere Pro V. 3.2.0(374).

Video clips were cropped to show only the mouth of the speaker and downsampled in color to a grey-scale. To avoid abrupt onsets and offsets the video and the audio were faded-in and faded-out (ramp duration: 1 second). After edition they were exported: video resolution 800x600, 25 fps compressor Indeo video 5.10, AVI format; audio sample rate 48 kHz 16 bits Mono.

All trials in the experiment consisted of two auditory speech streams presented simultaneously at different pitch plus one visual speech stream, all originating from a central location. First 36 pairs of auditory messages were pseudo-randomly created from the recorded clips. One of the streams in each pair was presented with the pitch shifted three semitones up (high pitch) and the other three semitones down (low pitch). The intensity of all tracks was modified to equalize the loudness of the stimuli; an average of 64.44 dB and 63.47 dB for the high and low pitch tracks, respectively. The design is fully balanced for pitch of the particular clips, so that any influence of pitch on individual clips was cancelled out. The onset of the speech signal was aligned in both streams. The only restriction while creating these pairs was that the content was unrelated to avoid possible priming across streams. Each audio track only appeared in one pair throughout the experiment.

From these 36 unique pairs, we generated three different AV conditions: One in which the video matched the high-pitch audio track, one in which the video matched the low-pitch track and one in which the video did not match any of the audio tracks; giving us a total of 108 AV stimuli. The particular distribution of these stimuli varies from experiment to experiment, and is explained below.

From each of the audio tracks in each of the 36 pairs of messages, two words were selected as targets for the recognition test. The target words were always nouns, never occurred in the first or the last two seconds of the audio track, and they appeared in only one of the tracks of the pair. For each target word a foil word was selected; target and foil words were comparable in frequency of use, number of syllables and imageability using the LEXESP database for Spanish (Sebastián-Gallés, Martí, Cuetos, & Carreiras, 2000). The foils never appeared in any of the audio tracks used in the experimental materials.

Stimuli in behavioral experiments were delivered in a protocol programmed with E-Prime 2.0.8.90. Video was presented centrally on a 19 inch CRT flat screen (Philips 109b) at 800x600 resolution and at 60 Hz refresh rate. Participants sat approximately 50 cm away from the screen. Audio was presented through headphones.

2.2. Procedure

In a given trial (see Figure 1) participants were first instructed by means of an arrow cue to attend the high or the low pitch stream. At the end of each trial participants performed a two alternative forced choice recognition task (2AFCR). They were presented with a pair of words on the screen, one on the left side and one on the right side; in this pair one of the words had been present in one of the audio tracks (target word) the other was present in none of the tracks (foil word); their task was to recognize which of the words was present in any of the audio track. Participants had a three second time limit to provide a response; they responded using the mouse buttons, right or left button for the word present on the right or left side respectively. The target position in the pair (left or right), and the order in which the targets appeared in the track and their order of appearance in the pairs (first or second pair) was counterbalanced. Participants performed two such tests for each trial in order to acquire enough measures per participant and condition. Participants' instructions informed them of the validity of the cue in the different experiments, that both targets will appear on the same track, and that they should perform the task as fast but above all as accurately as possible. Each participant was exposed to each of the 36 pairs only once, to avoid memory effects in the 2AFCR task, and different versions of the experiment were created so all tracks would pass through all conditions in each experiment across participants.

To ensure participants were looking at the video they were asked to visually monitor the central speaking lips during their selective listening task at all moments. In six of the trials (16.6%) the video frame rate slowed down from 25 fps to 3 fps during 1 second. Participants had to respond to this rate change by pressing the control key in the keyboard, and were informed that this slowing would occur in some of the trials but not in all of them. This slowing down occurred always in the second half of the videos, and it was counterbalanced across conditions (during Experiment 1, it only appeared in the attended trials). We set an a priori criterion for the visual task of at least a $d' \geq 2.75$ corresponding to at least 66% hits without false alarms. This was done to exclude subjects who may lead their gaze away from the visual display.

Just before the experiment participants ran two training blocks identical to the experiment but with a different set of stimuli.

2.3. Experiment 1

2.3.1. Participants

Twenty-one native Spanish speakers participated in this behavioral study (7 male, mean age 24 years old). All participants were right-handed, reported normal or corrected-to-normal vision and hearing and they gave written informed consent to participate. They were paid 7€ and the experiment lasted approximately 50 minutes. This study was approved by the Pompeu Fabra University ethics committee.

2.3.2. Materials, apparatus and procedures

We manipulated two main factors of interest: attention and congruency. The attention manipulation was introduced with the validity of the arrow cue so that target words appeared in the cued track 83% of the time (i.e. 30 trials), and in the uncued track in the remaining 17% (i.e. 6 trials). We used a high difference between the amount of attended trials and unattended trials to discourage divided attention as much as possible. For the congruency factor, in half of the trials the visual information matched the track containing the target words (congruent condition); while in the other half they matched the track not containing the target words (incongruent condition).

Therefore, the resulting conditions after crossing the two factors were attended congruent (41.5% of the trials), attended incongruent (41.5% of the trials), unattended congruent (8.5% of the trials), and unattended incongruent (8.5% of the trials) (see Figure 1 for a depiction of the conditions). Note that all trials contained the exact same amount of information, only the focus of attention of the observer and in which track appeared the target words (cued or uncued track) changed from one condition to the other.

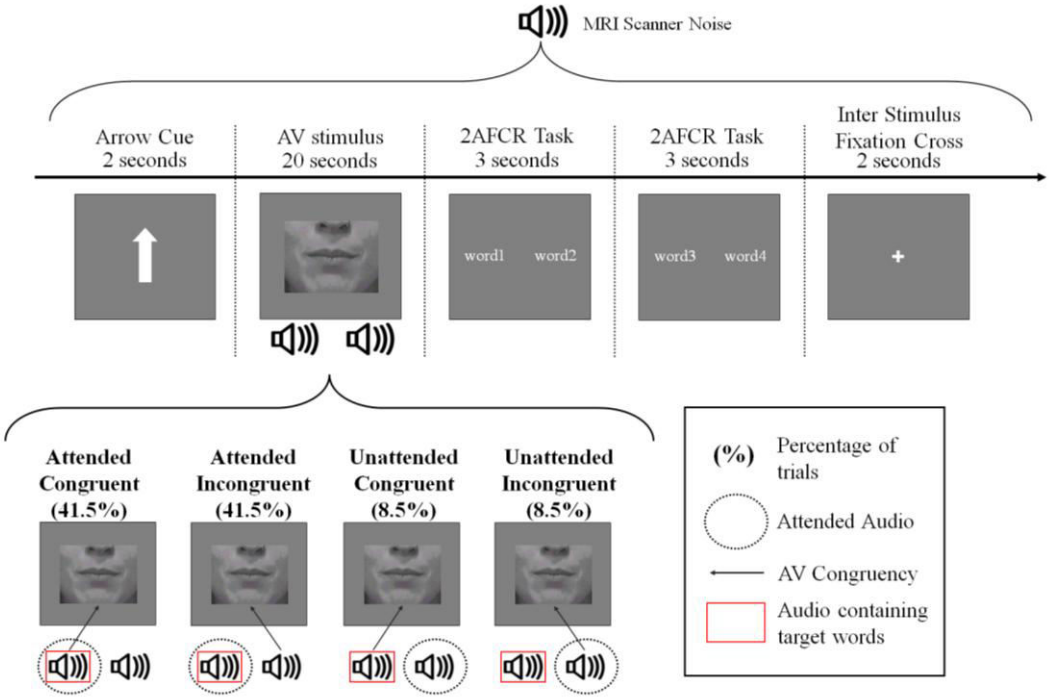

Figure 1. Experiment 1 task scheme. Participants are presented with an arrow cue, which combined with the AV stimulus produced one of the four conditions. All information is kept constant across trials and conditions, only the manipulation of the participants' attention produces each of the conditions. For clarity, audio sources are depicted spatially in the figure, but during the experiment they were always presented centrally and through headphones, so no spatial discrimination was possible. Across the whole experiment an MRI scanner noise was presented in the background.

2.3.3. Data analyses

The data was analyzed using software package R[1]; custom-made scripts were used for the permutation analysis described below. In the auditory task, we calculated the proportion of correct answers in the 2AFC task for each participant and condition, the two measures taken on each trial were treated as independent trials, and responses from trials containing a target in the visual detection task (slowing in the visual stimulus) were not included in the analyses. We performed a 2x2 repeated measures parametric ANOVA (Attention and Congruency as within participant factors). Parametric paired t-tests were used for follow-up analysis in significant interactions. Effect size is reported using partial squared eta (η^2_p) and the generalized eta squared (η^2_G) for the ANOVA, and the Hedges's g for related measures(g_{rm}) for the paired t-tests, all of them as described in Lakens (2013). Means (M), standard deviations (SD) and confidence interval of the mean difference between conditions (95% C.I., Mdiff) are also reported. To compensate for the possible biases due to the difference in variability between conditions with different number of trials (attended and unattended conditions) we also

[1] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

performed the nonparametric permutation version of the ANOVA (Ernst, 2004). All statistical tests were two-sided with α level of 0.05. In the case of the permutation tests we generated five null distributions, three for the F-values (two for the main effects and one for the interaction) and two for the t-values used as follow-up t-tests, using a Monte Carlo simulation with 10,000 iterations. For each of the iterations we randomly shuffled the condition labels within each participant's data (for the t-values only the tested conditions were shuffled); we then calculated the F-value, or t-value, for each of the tests and added it to the correspondent null distribution. The p-value for the permutation tests was calculated as the proportion of values that resulted in a larger statistic than the one observed in our data in the respective null distribution, with a practical lower limit of 10^{-4}, corresponding to the number of iterations. In addition, we calculated the *d'* values of the visual detection task for each participant and a two-sided paired t-test comparing the effect of attending to low or high pitch.

2.4. Experiment 2

2.4.1. Participants

Twenty-six subjects, native Spanish speakers, different from those in Experiment 1, participated in this behavioral study (15 female, mean age 24.5 years old). All participants were right-handed, reported normal or corrected-to-normal vision and hearing and gave written informed consent. Participants were paid 7€ and the experiment lasted approximately 50 minutes. This study was approved by the Pompeu Fabra University ethics committee.

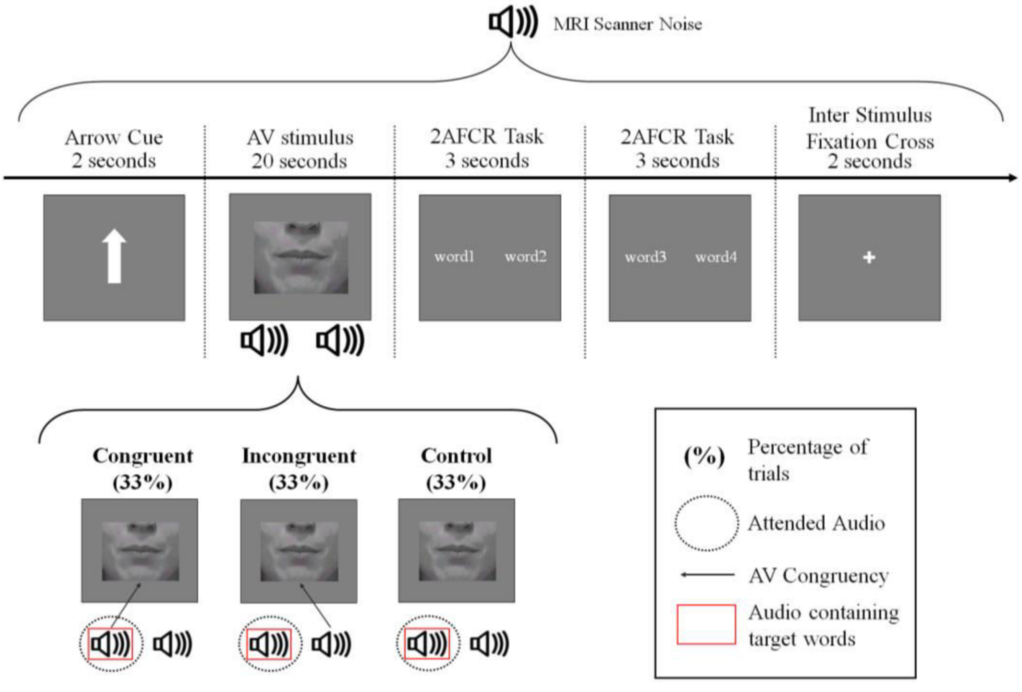

Figure 2. Experiment 2 Task scheme. Participants are presented with an Arrow Cue that combined with the AV stimulus produced one of the three conditions; in this experiment target words are always located in the attended audio track. (100% cue validity). Congruent: lips match the attended audio track; incongruent: lips match the unattended audio track; control: lips don't match the attended or unattended audio track. Across the whole experiment an MRI scanner noise was presented in the background.

2.4.2. Materials, apparatus and procedure

All aspects of the methodology were as in Experiment 1, except for the following. In this case the cue was 100% valid and a new condition where the lips did not match any of the two voices presented (control condition) was included. Therefore, the three resulting conditions for this experiment were the video matched the cued voice (congruent condition), the video matched the uncued voice (incongruent condition) and the video did not match any of the voices (control condition). Trials were equally distributed across the three conditions, resulting in 12 trials (i.e., 33%) for each condition (see Figure 2) and were equally likely to appear during the experiment.

2.5. Experiment 3

2.5.1. Participants

Thirty participants (14 females, mean age 24 years old) were included in the current fMRI study. All participants were right-handed, Spanish native speakers, in good health and without a personal history of psychiatric or neurological diseases, normal auditory acuity as well as normal or corrected-to-normal vision (visual lenses from VisuaStim, Magnetic Resonance Tech.) All gave informed consent prior to participation in the study. Participants were paid 20€ and the experiment lasted approximately 40 minutes (approximately 20 minutes inside the scanner). This study was approved by the Pompeu Fabra University ethics committee.

Data from five of the participants was excluded. Three of them did not perform properly during the behavioral task (less than 75% answers given). The fMRI images of one participant were not available due to a hardware error. One last participant was excluded due to excessive movement inside the scanner (several sudden moves larger than 3 mm). The behavioral data of the visual task for three participants was not acquired because of a failure in the button box; nonetheless, they were included in the imaging analysis. In the end, twenty-five participants were included in the behavioral and imaging analysis.

In this case, we decided not to exclude participants considering the visual task during this experiment in order to achieve a higher statistical power in the fMRI analysis. Experiment 1 and Experiment 2 proved that the rate of exclusions during the behavioral experiments due to the visual task was very low (3 of 47, see sections 3.1 and 3.2) indicating that participants were generally compliant with the instructions. Also visual information must have been harder to ignore during the fMRI experiment because visual information was presented through a pair of goggles attached to the participants head. Finally no difference was found in the d’ values across conditions (congruent d’=3.090; incongruent d’=3.334; control d’=3.099), neither was found a trend among participants indicating that they failed to monitor the visual information more often in the incongruent or control condition than during the congruent condition.

2.5.2. Materials, apparatus and procedure

The procedure was the same as in experiment 2 (see Figure 2) with the following differences: Interstimulus fixation was now 20 seconds long, this allowed the hemodynamic response to descend to baseline after each trial. Trials were presented in a pseudo-random order to avoid trials of different types to be too separated in time as this would produce a loss in the signal during the high pass filtering in the preprocessing step. The clips were presented trough the VisuaStimDigital AV system (Resonance

Technology Inc., Northridge, CA), 800 x 600 pixel resolution at 60 Hz refresh rate (the visual experience equals a 62 inch screen at 150 cm distance). Participants were presented with just one pair of words instead of two, and they responded with the right hand using a button box. They used their left hand to respond in the visual task.

As in previous experiment participants performed an equivalent session outside the scanner using different stimuli to familiarize with the task and apparatus.

2.5.3. fMRI Data acquisition and preprocessing

Images were acquired on a 1.5 T Siemens scanner (Avanto). First, a high-resolution T1-weigthed structural image (GR\IR TR=2200ms TE=3.79ms FA=15º 256 x 256 x 160 1mm isotropic voxel size) was acquired. Immediately after, functional data was acquired in a single run consisting of 660 Gradient Echo EPI functional volumes not specifically co-planar with the Anterior Commissure – Posterior Commissure line, in an interleaved ascending order, using a 64× 64 acquisition matrix a FOV=224, TE=50 ms, TR=2000, voxel size 3.5 x 3.5 x 3.5 mm with a 0.6 mm gap between slices covering 94.3 mm in the Z axis (23 slices) trying to cover the whole brain. Three dummy scans were presented prior to data- acquisition.

Standard spatial preprocessing was performed for all participants images following these steps: Horizontal AC-PC reorientation; realignment and unwarp using the first functional volume as reference, a least squares cost function, a rigid body transformation (6 degrees of freedom) and a 2nd degree B-spline for interpolation, creating in the process the estimated translations and rotations occurred during the acquisition; slice timing correction using the middle slice as reference using SPM8's Fourier phase shift interpolation; coregistration of the structural image to the mean functional image using a normalized mutual information cost function and a rigid body transformation; image was normalized into the Montreal Neurological Institute (MNI) space, voxel size was unchanged during normalization and interpolation was done using a 4th B-spline degree; functional data was smoothed using an 8-mm full width half-maximum Gaussian kernel to increase signal to noise ratio and reduce inter subject localization variability.

2.5.4. fMRI Analysis

The time series for each participant were high-pass filtered at 128 s and pre-whitened by means of an autoregressive model AR(1). At the first level (subject-specific) analysis, box-car regressors modeled the 3 conditions of interest (congruent, incongruent, and control) as 20 second blocks and the response and cue periods as 3 and 2 seconds blocks respectively. All these regressors were convolved with the standard SPM8 hemodynamic response function. The inter stimulus resting periods were not explicitly modeled. Additionally, the six movement regressors provided by SPM during the realign process were also included. The resulting general linear model produced an image estimating the effect size of the response induced by each of the conditions of interest.

The images from the first level were introduced in a second level analysis (inter-subject). In this second level paired t-test models were used for pair-wise comparisons between conditions. First we created a conjunction mask: [(congruent > rest) ∩ (incongruent > rest) ∩ (control > rest]. This allowed to exclude from the analysis the areas that were less activated during the task than during rest and also to reduce the search volume increasing our statistical power and therefore our sensitivity.

Statistical images were assessed for cluster-wise significance using a cluster-defining threshold of P=0.001; the p<0.05 FEW critical cluster size was corrected for multiple comparisons by means of a Monte Carlo simulation performed using AlphaSim (center to center maximum distance was 5mm which provides an 18 connectivity scheme, i.e. edge and face connectivity) included in the AFNI package (Cox, 1996) obtaining a 22 cluster size threshold in the congruent vs. incongruent contrast (FWMH: 12.2mm, 12.2mm, 11.8mm) and a 23 cluster size threshold for the congruent vs. control contrasts (FWMH: 12.4mm, 12.4mm, 11.9 mm); the search volume used was the conjunction mask created in the previous step in both cases. Percent signal change was calculated using the average of all voxels forming each reported cluster, using MarsBaR (Matthew Brett, Jean-Luc Anton, Romain Valabregue, 2002).

An a priori motivated ROI analysis was performed. Two 10mm radius spheres centered over the two peaks reported in Fairhall & Macaluso (2009) for both, left and right, STS in the Attend AV congruence > Attend AV incongruence contrast (MNI coordinates in mm left and right respectively: x=-57 y=-12 z=-9; x=60 y=6 z=-12). Pair-wise t-test comparisons were calculated at the ROI level, summarizing the cluster activity by using the mean of all the voxels in each sphere, using MarsBars.

Inflated brain figures were created using Caret Software over an inflated brain derived from the PALS atlas (Van Essen, 2005; Van Essen, D.C., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H. and Drury, 2001).

See supplementary materials for analysis correlating behavior and BOLD signal.

3. Results

3.1. Experiment 1

The goal of the first experiment was to test to which extent speech information that appears in an unattended auditory stream is amenable to audio-visual integration. If visual speech information is integrated automatically, even with the unattended auditory stream, this should affect audio-visual congruency in both attended and unattended conditions; specifically a facilitation effect should appear in the congruent conditions independently of attention. To test this, we measured behavioral performance in the 2AFC task where target words had appeared in the cued (i.e., attended) or uncued (i.e., unattended) auditory track in an unpredictable way.

Exclusion of participants based on the visual monitoring task

Data from one participant was excluded from the analysis because he performed below the criterion (*d'* ≥ 2.75) during the visual monitoring task. All remaining participants performed individually well above the criterion (mean hit rate: 0.98; mean false alarm rate: 0.02). This indicates that participants were compliant with the instructions and were looking and attentive to the lips on the screen during the experiment.

Results and discussion

As for the results of interest (see Figure 3), in the selective listening task, the main effect of Attention ($F_{1,\ 19}$=43.08, p<0.001; F-test permutation, p<0.001; η^2_p=0.695; η^2_G=0.372) was significant, with word recognition performance in the attended condition streams being higher than to the unattended condition. The main effect of Congruency was not significant overall ($F_{1,\ 19}$=0.599, p=0.449; F-test permutation,

$p=0.452$; $\eta^2_p=0.030$; $\eta^2_G=0.010$), but the critical interaction between attention and AV congruency ($F_{1,\,19}=7.12$, $p=0.015$; F-test permutation, $p=0.015$; $\eta_p^2=0.270$; $\eta^2_G=0.084$) resulted significant. Following up on the interaction, t-tests between congruency conditions at each level of the attention factor revealed a significantly superior performance for the congruent ($M=0.792$, $SD=0.092$) versus incongruent ($M=0.721$, $SD=0.100$) AV streams in attended trials ($t_{19}=2.20$, $p<0.040$; t-test permutation, $p=0.039$; $g_{rm}=0.723$; 95% C.I. of the mean difference M_{diff} = [0.558, 0.138]), but not between the congruent ($M=0.417$, $SD=0.183$) and incongruent ($M=0.558$, $SD=0.277$) AV speech streams when unattended ($t_{19}=-1.78$, $p=0.091$; t-test permutation, $p=0.089$; $g_{rm}=0.594$; 95% C.I. M_{diff} = [-0.025, 0.308]). Performance was not significantly different from chance in any of the two unattended conditions (unattended congruent, $t_{19}=2.03$, p-value=0.056; unattended incongruent, $t_{19}=0.94$, $p=0.358$). For completeness, we compared and found no significant difference between attending high pitch or low pitch ($t_{19}=1.29$, p-value=0.212).

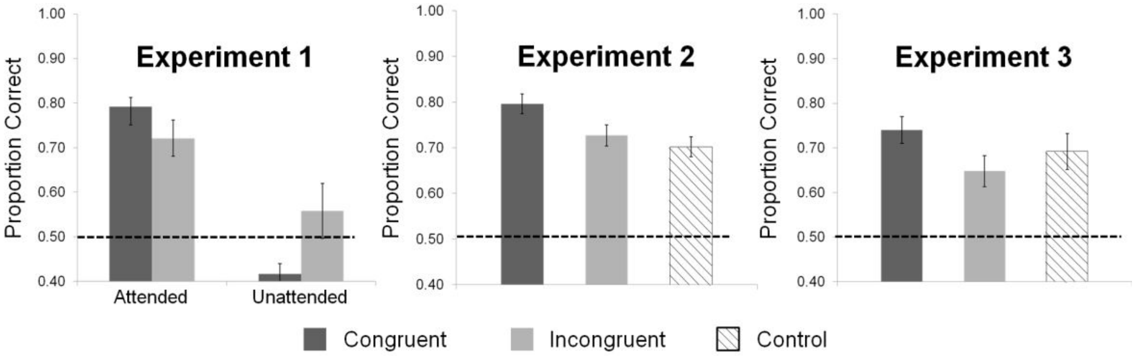

Figure 3. Mean proportion of correct responses across participants (bars), standard error of the mean (error bars) and chance level (dashed line) for experiment 1 (left), experiment 2 (center) and experiment 3 (right).

In Experiment 1, we found a significant improvement of speech perception when participants attended the auditory stream that was congruent with the lips in the screen compared to when they attended the incongruent one. Critically, when the message was unattended, there was no effect of AV congruency.

First, this result indicates that AV integration weakens (or is absent altogether) under unattended conditions suggesting that both attention and AV congruency impact the processing of the stimuli and that an interaction between them exists. And second, albeit the possibility of an effect of AV congruency may exist (undetected due to a reduced power in the unattended condition) the numerical trend in performance for congruent vs. incongruent in the unattended condition is the opposite as the one found (significant) in the attended condition, as indicated by the interaction; therefore the existence of a weak but consistent effect of facilitation due to congruency in the unattended condition is unlikely.

Yet, one possible critique to our conclusion is that attending to incongruent stimuli involved also ignoring AV congruent stimuli, hence the trend toward a reversed pattern of congruency effects in the unattended condition. This is addressed in the next behavioral test (Experiment 2) and the subsequent neuroimaging experiment (Experiment 3) where we decided to include a baseline where both the attended and the ignored streams were audio-visually incongruent.

3.2. Experiment 2

During experiment 2 we put our hypothesis through further test by measuring a possible interference from unattended AV congruent speech; which, if true, would contradict our initial hypothesis. Here we probed participants only on the attended auditory speech stream, and manipulated whether the visual speech stream matched the attended auditory message, the unattended auditory message or neither. According to previous literature, we expected to find higher performance when attending AV congruent speech versus when attending AV incongruent trials and the control trials. Crucially if, against our hypothesis, some integration effect occurs in the unattended stimuli we expect to see a difference between the AV incongruent and control conditions. That would mean that a competing AV matching speech stream outside the focus of attention would interfere more with the primary attended message than a non-matching AV stream. This paradigm constitutes the basis for the subsequent fMRI experiment.

Exclusion of participants based on the visual monitoring task

Two participants were excluded from the analysis because they performed below the criterion (*d'* ≥ 2.75) during the visual monitoring task. All remaining participants performed individually well above the (mean hit rate: 0.97; mean false alarm rate: 0.002). Again, this indicates that participants were compliant with the instructions and were looking and attentive to the lips on the screen during the experiment.

Results& discussion

As for the main selective listening task (see Figure 3), we performed a one-way repeated measures ANOVA that resulted significant ($F_{2,46}$=5,046; p=0.010; η_p^2=0.180; η^2_G=0.121). Two-tailed parametric t-tests revealed a significant difference between the congruent (M=0.796, SD=0.107) and incongruent (M=0.727, SD=0.113) conditions (t_{23}=2.355, p=0.027, g_{rm}=0.614, 95% C.I. M_{diff}=[0.008, 0.129]) and between the congruent and the control (M=0.702, SD=0.107) condition (t_{23}=3.157, p=0.004, g_{rm}=0.861, 95% C.I. M_{diff}=[0.032, 0.155]), as one would expect according to prior literature. Interestingly, we did not detect differences between the control and incongruent condition (t_{23}=-0.764, p=0.452, g_{rm}=0.224, 95% C.I. M_{diff}= [-0.043, 0.093]). This result indicates, as the hypothesis of no AV integration for unattended conditions would predict, that the distractor (unattended) stream interferes equally regardless of whether it matches the central lips or not. Again, no significant difference between attending high pitch or low pitch was found (t_{23}=1.23, p-value=0.230).

As in Experiment 1, AV congruence had a beneficial impact on behavior when it was attended (improvement over the two other conditions) and yet, when unattended, AV congruence did not show any further significant impact (by interfering with the relevant message) than an AV incongruent control stream. To be on the cautious side, one could say that even if there was an effect of AV incongruent speech, it was of a considerably smaller size compared to its attended counterpart. Together with Experiment 1, our behavioral data so far suggests that MSI of AV speech vanishes, or at least is substantially weaker, in the absence of attention. In the next experiment we further tested this hypothesis using fMRI.

3.3. Experiment 3

The two previous behavioral experiments suggested that the behavioral expression of MSI during speech perception is strongly modulated by attention. In this experiment we addressed what are the brain mechanisms that underlie this attentional modulation of MSI. Therefore, we tested a new sample of participants using the similar paradigm to that in Experiment 2 while measuring BOLD responses in an fMRI protocol.

Under our initial hypothesis, in this experiment we expected to see modulation of brain regions that have been previously involved in MSI. For example, several high-level association brain areas have been previously highlighted as possible nodes of a MSI network (for reviews see, M. S. Beauchamp, 2005; G. A. Calvert, 2001; Campbell, De Gelder, & De Haan, 1996) One candidate area was the STS, a region that has been proved in many past studies to be responsive to multisensory stimulation in speech and other, meaningful, stimuli (Beauchamp, Lee, & Argall, 2004; Calvert, Campbell, & Brammer, 2000; Fairhall & Macaluso, 2009; Miller & D'Esposito, 2005; Nath & Beauchamp, 2012; Toemme Noesselt et al., 2007; Tömme Noesselt, Bergmann, Heinze, Münte, & Spence, 2012; Stevenson, Altieri, Kim, Pisoni, & James, 2010; Stevenson, VanDerKlok, Pisoni, & James, 2011; Stevenson & James, 2009). We hypothesized that we should see a higher BOLD response in the STS in the AV congruent condition as compared to the AV incongruent and control conditions. Again, in keeping with the behavioral results, a strong prediction of our hypothesis is that no differences were expected to appear between AV incongruent and control conditions in this area. That is, ignoring AV congruent vs. ignoring AV incongruent stimuli would not make a difference. Above and beyond higher level association areas, many authors claim nowadays that putatively unisensory areas reflect the expression of MSI (Jon Driver & Noesselt, 2008b; Fairhall & Macaluso, 2009; E. Macaluso, 2000; Emiliano Macaluso & Driver, 2005; Miller & D'Esposito, 2005; Pekkola et al., 2005; Schroeder & Foxe, 2005). We were therefore interested in exploring the response of unisensory areas to attentional modulations of MSI as well.

3.3.1. Behavioral results

The pattern of behavioral results (see Figure 3) for the task ran inside the scanner was similar to that of Experiment 2, but the one way ANOVA revealed only a marginally significant effect between the measures in the three different conditions ($F_{2, 48}$=2,441; p=0,098). This is probably due to the reduction in the number of measures. For completeness and comparison with Experiment 2, parametric paired t-tests between conditions are shown. In the same trend as in Experiment 2, performance in the congruent condition (M=0.74, SD=0.147) was higher than in the incongruent (M=0.648, SD=0.169) and control (M=0.692, SD=0.198) conditions, albeit only significantly so in comparison to the incongruent condition (t_{24}=2.255, p=0.034) but not when compared to the control condition (t_{24}=1.266, p=0.218). As expected, there was no significant difference between the incongruent and the control conditions (t_{24}=0.960, p=0.347). No significant difference between attending high pitch or low pitch was found (t_{24}=0.28, p-value=0.780).

To compensate the reduced number of measures, we merged the behavioral data from this experiment and Experiment 2 in a 2x3 ANOVA, with experiment as a between subject factor and congruency as a within subject factor. The congruency factor was significant ($F_{2, 94}$=5.701; p=0.005). The effect of experiment was not significant ($F_{1, 47}$=2.913; p=0.094) and there was no significant interaction between the terms of the ANOVA ($F_{2, 94=}$0.911; p=0.405) indicating that results from both experiments followed the same trend. Parametric paired t-tests between conditions revealed a significant difference between AV congruent condition and the other two, incongruent condition

($t_{48=}$3.218; p=0.002) and control condition (t_{48}=2.912; p=0.005), but not between incongruent and control condition (t_{48}=-0.358; p=0.721).

fMRI results

		Number		Coordinates (mm)		
Hemisphere	**Region**	**of voxels**	**Z - Score**	**X**	**y**	**z**
Congruent > incongruent						
R	Inf. temporal gyrus	859	5.71	45	-61	-5
			5.70	48	-73	-8
			5.48	45	-43	-17
L	Inf. occipital gyrus	519	5.51	-45	-73	-2
			4.98	-42	-49	-17
			4.48	-36	-76	-8
R	Precentral gyrus	77	4.67	45	-1	31
			4.06	60	-7	40
			3.68	48	-7	37
L	Postcentral gyrus	116	4.57	-48	-7	55
			4.10	-51	-1	49
			4.02	-54	-13	37
L	Sup. temporal gyrus	81	4.17	-42	-34	19
			3.59	-39	-19	22
			3.58	-51	-19	19
L	Sup. temporal sulcus	39	3.95	-48	-31	1
R	Sup. temporal sulcus	50	3.74	54	-34	-8
			3.67	48	-22	-8
			3.38	60	-19	-11
Incongruent > congruent						
R	Inf. frontal gyrus	30	4.28	30	41	16
L	Ant. insula	48	4.15	-30	20	10
			3.76	-36	20	1
L	Supplementary motor area	31	3.68	-12	5	61
			3.43	-3	14	49
			3.38	0	8	55
R	Ant. insula	26	3.66	33	26	7
			3.28	36	17	-2
Congruent > control						
L	Fusiform gyrus	223	4.82	-39	-46	-17
			4.26	-42	-76	-5
			3.80	-45	-52	4
R	Inf. temporal gyrus	297	4.56	51	-73	-8
			4.13	42	-58	-20
			4.09	33	-94	-8
L	Inf. frontal gyrus	36	3.91	-51	35	1
			3.66	-48	35	10
Control > congruent						
L	Supplementary motor area	40	3.88	-3	14	46
L	Ant. Insula	31	3.68	-30	20	7
			3.37	-33	11	7

Table 1. Table showing activation clusters in all contrasts showing the hemisphere and approximate anatomical region using AAL atlas (Tzourio-Mazoyer et al., 2002), the Z-score and location of each of the three highest maxima in a cluster that were more than 8mm away. Spatial coordinates correspond to the MNI reference space. Contrasts that do not appear in this table did not show any significant results.

Congruent vs. incongruent

This contrast involves the exact same physical stimulus display, with the only difference being which one of two competing auditory messages the observer is paying attention to. In the congruent condition the observer is paying attention to the message congruent with the centrally presented lips, in the incongruent condition the observer is attending the incongruent one, and in both the observer is watching the lip movements (necessary in order to detect visual targets). Here, the [congruent > incongruent] contrast revealed a higher hemodynamic response in the STS bilaterally, as it was expected based on the results of other studies. The ROI analysis over the peaks reported in Fairhall & Macaluso (2009) corroborated this result (left and right respectively t=3.22, p=0.002; t=2.73, p=0.006). Other areas, known to respond to AV speech, were also responsive to this contrast, including the sensory-motor cortex (precentral and postcentral gyri) bilaterally, and a cluster covering the left superior temporal gyrus (STG), the supramarginal gyrus and the inferior parietal cortex. We also found a robust modulation in extrastriate visual areas covering occipital and inferior temporal areas. The reverse contrast [incongruent > congruent] revealed a higher bilateral response in the anterior insula and a cluster going from the supplementary motor area to the anterior cingulate cortex (ACC) (BA 6 and 32). This pattern of brain activity reveals that attending congruent AV speech engages a network previously linked to AV integration, while a very different network is engaged when the congruent AV speech is unattended (Figure 4, Table 1).

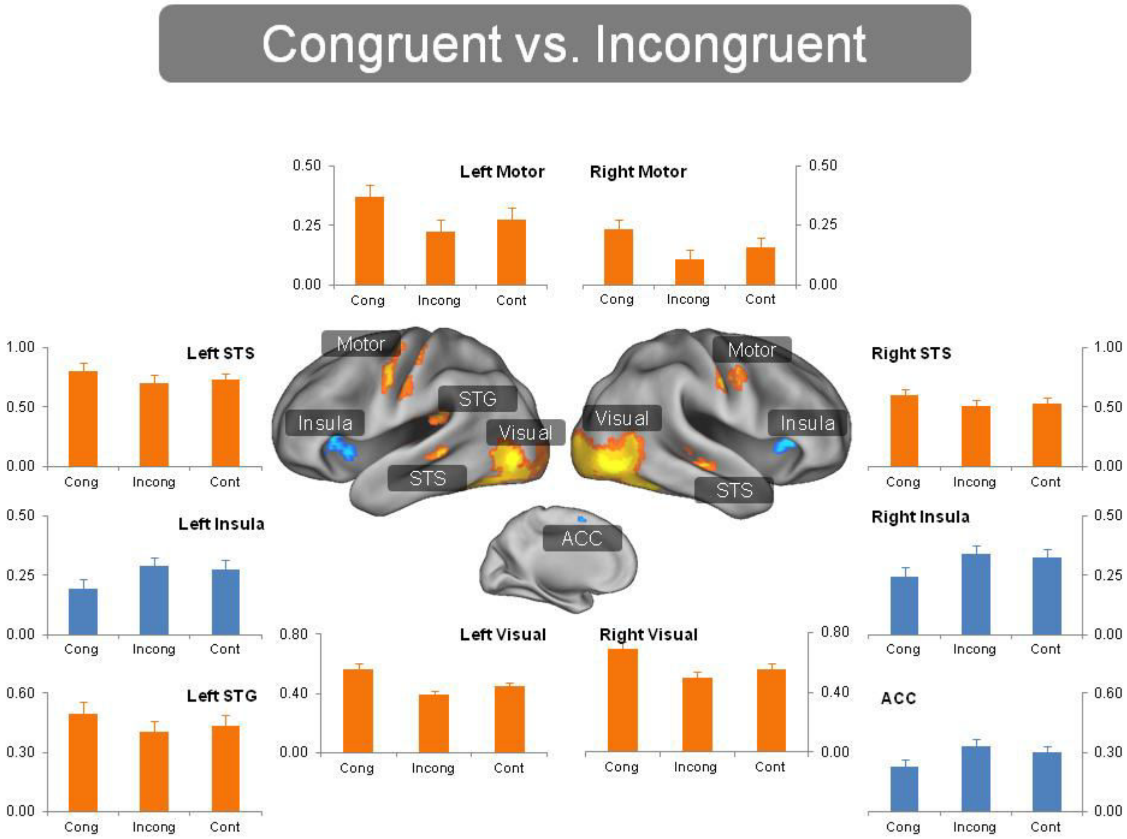

Figure 4. t-maps showing the positive effects on the [congruent > incongruent] contrast in hot color and the [incongruent > congruent] contrast in cold colors, along with corresponding the percent signal change for all three conditions. Maps are presented at p<0.001 threshold and have a minimum cluster extent of k=22. (Cong: Congruent, Incong: Incongruent, Cont: Control, STG: Superior Temporal Gyrus, STS: Superior Temporal Sulcus, ACC: Anterior Cingulate Cortex)

Congruent vs. control

Next we assessed the effect of attending congruent AV speech vs. attending incongruent AV speech but in this case the unattended stream was always AV incongruent in both cases. In this contrast [congruent > control], we found a higher BOLD response in the visual areas covering occipital and inferior temporal areas bilaterally but with a larger extent in the right hemisphere, and in the left inferior frontal gyrus. This network of brain areas is included in that found for the comparison between congruent vs. incongruent (above), where the irrelevant (to be ignored) stream in the incongruent condition was congruent. Indeed, when relaxing the voxel significance threshold (to p<0.005), the map resulting from the [congruent vs. control] contrast overlaps to a large degree with the one found on the [congruent > incongruent] (see Figure 5). In the [control > congruent] contrast we found a higher response in the left anterior insula and the ACC (BA 32). Again, this matches rather well with what was found in the comparable contrast for the [incongruent > congruent] conditions. All in all what these contrast suggests is that neural activity differences between conditions is driven by the congruency, or incongruency, between the attended auditory stream and the visual stream, while the congruency between the unattended stream and the visual stream is irrelevant.

This supports the idea that unattended congruent AV speech may not engage integration processes, or at least not nearly as strongly, as attended AV speech.

Furthermore the ROI analysis confirmed that the STS showed a higher activation during the congruent condition than during the control condition (statistical values for left and right hemisphere respectively: t=2.43, p=0.011; t=1.91 p=0.033).

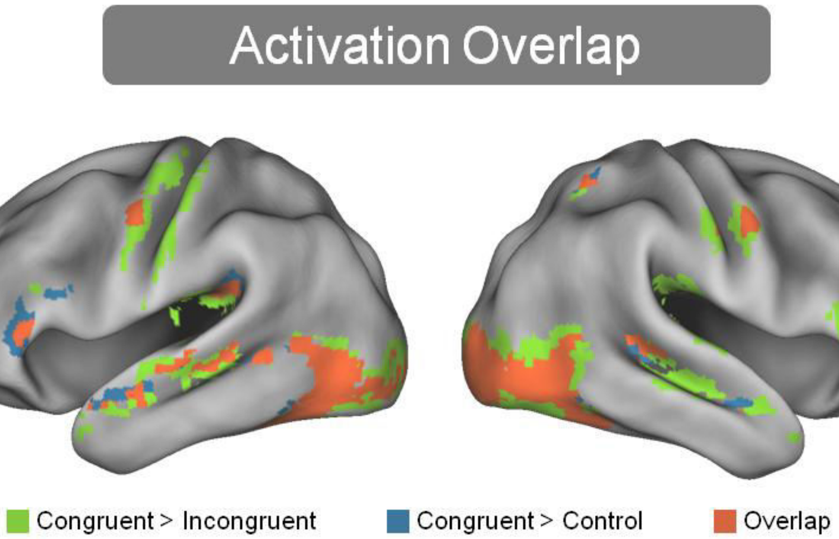

Figure 5. Maps showing the projection of the [congruent > incongruent] contrast (green color), the [congruent > control] contrast (blue color) and the overlap between both (orange color) on a PALS atlas standard brain. Both contrasts are thresholded using an uncorrected voxel p-value<0.005 and a cluster size threshold of 10 voxels.

Incongruent vs. control

No significant differences were found for the contrasts involving AV incongruent compared to the control condition in either direction. For comparison with the analysis above, we also applied a more relaxed significance threshold (same as that used to reveal a pattern in the congruent vs. control contrast; p<.005), but in this case it did not lead to any significant differences in the contrasts [incongruent > control] and [control > incongruent] (Table 1). The ROI analysis in the STS also failed to reveal any significant differences between the AV incongruent and control conditions in the left (t=1.39, p=0.088) or right (t=0.22, p=0.413) hemispheres. This particular finding would lead one to conclude that, attending to one out of two competing auditory messages engages a very similar pattern of BOLD responses, regardless of whether the ignored message is AV congruent or AV incongruent. This result is complementary to the one found in the [congruent vs. control] contrast that when ignored AV congruency does not engage any distinct brain areas compared to AV incongruency.

4. General Discussion

The present findings illustrate how auditory selective attention to speech can strongly modulate AV integration processes in a multisensory context, not only at the behavioral level but also the underlying brain mechanisms revealed through fMRI. We have used a paradigm that allowed us to study the interaction between multisensory integration and auditory selective attention by manipulating exclusively the focus of attention of the participant while keeping constant the sensory stimulation (i.e., in terms of amount of information in the display and the AV congruency relationships). The present stimulation conditions are not unlike everyday life environments, where long and meaningful fragments of speech have to be parsed in the context of other, irrelevant, speech messages. The main conclusion to be extracted from the results of our three experiments point in the same direction: integration only occurs (or at least it occurs to a much larger degree) when both components of the AV stimuli are in the focus of the listener's attention. Below, we discuss the implications of these findings in terms of current knowledge about the interplay between attention and MSI, and derive an explanation that accounts for these, and previous results.

The results of Experiments 1 and 2 (behavior) indicate that attention must be placed in both sensory inputs of an AV stimulus for integration to have a consequence in responses. Indeed, no AV integration seemed to take place when the AV stimulus was present but unattended. At the very least, present but unattended AV stimuli did not gave rise to any measurable behavioral effect within our paradigm. This result argues against a strictly automatic account of AV speech integration, at least in the context of auditory selection. In addition, please note that, different from other studies, attentional selection here was object based as opposed to spatially-based. In this way, stimuli were presented from the same (central) location and selection could not be influenced by spatial factors (i.e. ventriloquism effect, Driver, 1996). Current evidence for the automaticity of MSI comes mostly from the ventriloquist illusion produced by a beep and a co-occurring flash (Bertelson et al., 2000; Vroomen et al., 2001) and the Pip and Pop effect (Van der Burg et al., 2008). According to this view multisensory coincidence outside the current focus of attention can trigger an orienting of attention to that location or time (Van der Burg et al., 2008). If this had been the case also in our study, then we should have seen a drop in performance when people are attending incongruent streams and trying to ignore an AV congruent stream, due to the unwanted shifts of attention to the congruent stimulus. In parallel, this should have improved word recognition performance when the target came from unattended but AV congruent trials (for exactly the same reason). Yet, none of these effects occurred.

The present conclusion stands in stark contrast to the conclusions from previous ventriloquist and pip-and-pop studies, which had indicated rather unavoidable automatic integration of AV stimuli. Unfortunately, such studies differ in many substantial ways from the present "cocktail-party" type of task to allow for a thorough comparison. Indeed, as several authors have pointed out MSI should not be seen as a monolithic process but as a multifaceted phenomenon that may express differently depending on the task or nature of the stimuli presented (Talsma et al., 2010). In fact, it is precisely the complexity of speech in terms of the rapid variation in temporal structure that might explain part of the discrepancy between AV speech and other demonstrations of AV integration automaticity like the mentioned pip and pop and ventriloquist effects. It has been suggested that when the limit in information processing capacity is reached during the cocktail-party situation (e.g., under high load conditions), attentional processes are engaged to actively select and ignore the appropiate input (Hill & Miller, 2010; Lee et al., 2014). Therefore, when AV automaticity has been tested under high perceptual load (arising from complex stimuli such as speech or other events with quick temporal variation, as the cocktail party we used in our experiment) the conclusion is more often than not that AV integration needs attention to occur (Alsius et al., 2005, 2007; Alsius & Soto-Faraco, 2011; Tiippana et al., 2004, 2011; Zion Golumbic et al., 2013). Perhaps one notable exception to this is Driver (1996) which illustrated the existence of AV integration prior to attentional selection in a set of elegant behavioral experiments. In one experiment, Driver showed that observers could select (and track) one out of two, co-located, auditory messages solely on the basis of AV congruence. Strikingly, participant's performance improved when the visual information was located away from the sound sources, as compared to when it was co-located with them, putatively as a consequence of the ventriloquist effect pulling only the matching sounds toward the visual stream. In their crucial second experiment, Driver showed that the AV congruency could also interfere with selective attention if it resulted in the illusory proximity between two physically disparate auditory messages. Unfortunately, for the purpose of full comparison with the present experiment, the control condition for their experiment consisted of occluding the speaker's lips, instead of an incongruent face. We believe this is important to rule out that the matching lips introduced an additional source of interfering information. According to our results we would speculate that no difference would have been found between Driver's experimental condition and a condition in which the lips did not match any of the voices, indicating, as in our results, that AV congruency when unattended is irrelevant.

The present fMRI findings support and extend the behavioral results in the comparable task. Namely, using the same paradigm as in the behavioral protocol tested in Experiment 2, BOLD responses in Experiment 3 revealed that attending to congruent AV speech engaged a set of integration areas in the brain, including heteromodal regions (such as the STS) as well as unisensory areas (such as the extrastriate cortex). Although the elements of this network will be discussed in more detail below, the relevant message here is that these brain areas were engaged when AV congruent speech was present and attended, but not when present but ignored. Even more the pattern of BOLD responses to the control condition, where AV congruency was never present (neither attended nor unattended), resembled that of ignoring AV congruency and differed from that obtained when attending to AV congruency.

The actual network of areas unveiled by attending (vs. ignoring) AV congruent speech included the STS as well as visual areas (extra-striate cortex), auditory areas (STG), and motor areas. The STS has been often highlighted as one of the main loci for MSI, especially but not exclusively, regarding AV integration of speech (M. Beauchamp et al., 2004; G. a Calvert et al., 2000; Fairhall & Macaluso, 2009; Miller & D'Esposito, 2005; Nath & Beauchamp, 2012; Toemme Noesselt et al., 2007; Tömme Noesselt et al., 2012; Stevenson et al., 2010, 2011; Stevenson & James, 2009). In particular, two different parts of the STS have been found to activate depending on the nature of the audiovisual manipulation. When unimodal conditions are compared with multimodal conditions (i.e. the often used contrast between the sum of unimodal conditions, A+V, vs. the multisensory condition, AV) a more posterior part of the STS is usually found to be responsive (M. S. Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; G. a Calvert et al., 2000; Nath & Beauchamp, 2012). Instead, studies specifically contrasting congruency or fusion conditions with incongruent conditions have failed to find activation in the pSTS (Fairhall & Macaluso, 2009; Jones & Callan, 2003; Miller & D'Esposito, 2005) and usually, as in our study, find responses in a more anterior area, the middle STS (mSTS) (Fairhall & Macaluso, 2009; Miller & D'Esposito, 2005; Stevenson et al., 2011).

In our study, the activity in the mSTS resulted higher during the congruent condition than in any of the other two conditions (albeit we had to turn to a more sensitive analysis to find this difference with the control condition) and this pattern is similar to that found in Fairhall & Macaluso, 2009. This pattern of activity supports that the mSTS needs both AV congruency and attention to be engaged.

Our results regarding the STS align well with the findings of Fairhall & Macaluso 2009, which compared attended vs. unattended AV congruency. It is encouraging to see this coincidence even across the two very different paradigms used (spatial visual attention in their study compared to auditory attention in ours). However, going beyond the results of Fairhall & Macaluso (2009), the linguistic information in our paradigm was relevant to the behavioral task. That is, as opposed to their task (which consisted in monitoring visual targets unrelated to the task to make sure participants attended to the cued face), we measured participant's performance on the actual speech message (recognizing target words). The behavioral improvement found in our study for attended congruent speech could possibly be related to the increased activity in the STS area, as it has been previously reported to be involved in the creation of multisensory percepts (Miller & D'Esposito, 2005; Stevenson et al., 2011) that may have entailed a better comprehension of the auditory track during (attended) congruent conditions. This increased performance can also be related with the increased activity in the auditory area (STG) during the congruent condition, an area previously found to be active when presenting AV speech (Calvert et al., 2000; Miller & D'Esposito, 2005; see Pekkola et al., 2005 for auditory activation during lip reading). Nonetheless we were unable to find a significant correlation between the BOLD signal in these areas and the behavioral score of individual participants (see Supplementary materials). One possible reason is that the behavioral measure was noisier during the fMRI experiment because of the reduced number of measures (see section 3.3.1). Therefore the direct relation between the behavior and the activity remains speculative for our data, and is based on the interpretation given for activity in these areas in previous studies.

Another focus of BOLD activity when comparing attended vs. unattended AV congruent speech was found in sensory-motor areas. This cluster was roughly located bilaterally in the mouth motor area according to probabilistic atlases (Fox et al., 2001). These areas have previously been found in relation to the processing of AV congruent speech, especially in the context of studies advocating for motor theories of speech perception (Skipper, Nusbaum, & Small, 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007). Activity in motor or pre-motor areas is often interpreted as the participation of motor circuits in the perception of speech (Wilson, Saygin, Sereno, & Iacoboni, 2004) and has been specifically shown to be stronger if auditory information is accompanied by visual information (Skipper et al., 2005).

Of relevance is the increased BOLD signal of the visual areas when attention was directed to congruent speech compared to when it was directed to incongruent speech the (incongruent and control conditions). This increased activity during the congruent condition probably reflects a deeper processing of the visual information which, as proved by our behavioral data, improved participants performance. A compatible and probably complimentary explanation to this difference in BOLD signal would be the inhibition of visual processing in the incongruent and control conditions due to the mismatch between the visual information and the relevant auditory message. If we place the focus on the putative decrease BOLD activity in visual areas during the incongruent and control conditions, it becomes relevant to consider the areas that displayed stronger responses when attention was directed to incongruent AV speech (compared to congruent AV speech). Indeed, as reported in the results section, we found an increase of activity bilaterally in the anterior insula and in the frontier between the supplementary motor area and the ACC (BA 6 and 32 respectively). In general, this pattern of activity is consistent with the role of the cingulate cortex (Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; Roberts & Hall, 2008; Shenhav, Botvinick, & Cohen, 2013) in conflict detection and resolution. Specifically this network is activated when participants have to override an automatic behavior in favor of a non-automatic one (e.g., Stroop or Flanker task). This type of conflict responses in the ACC have been reported before in multisensory contexts in written words, letters or pictures combined with corresponding or non-corresponding auditory counterparts (Uta Noppeney, Josephs, Hocking, Price, & Friston, 2008; Orr & Weissman, 2009; Weissman, Warner, & Woldorff, 2004; Zimmer, Roberts, Harshbarger, & Woldorff, 2010) and in AV speech (Miller & D'Esposito, 2005; Pekkola et al., 2006; Szycik, Jansma, & Münte, 2009).

Given the above, one has to consider that in our results of visual down-regulation accompanied by the increment in the cingulate gyrus and the insula when attending to incongruent AV speech could reflect detection of the AV speech conflict (ACC) and, as a consequence of this conflict, a decrease in the processing of the least relevant modality, in this case the visual one (see Navarra et al., 2010, for a similar hypothesis based on behavioural findings). Following this interpretation, we propose a more general hypothesis that explains our results. First, that this MSI process (AV speech integration) is hardly accessible (if at all) by our volition (i.e. we cannot decide to integrate or not integrate AV information). Second that for AV integration to occur unisensory information must be processed to a certain degree, something that does not happen when information is unattended. If processing of the relevant unisensory streams unfolds to that certain degree (i.e., when attended), then an attempt to integrate will be made independent of the AV congruency. In case of attended AV incongruency, this attempt to integrate AV information will lead to a cost (maybe a discomfort) due to the conflict between the auditory and visual information. Therefore the only way of reducing the impact of this incongruency, and of modulating the ensuing integration process, is to inhibit the processing of the least relevant modality for the task at hand. By extension this implicates that the only way of modulating AV speech integration is to modulate the input from the low-level areas.

One could express this idea in a Bayesian framework by proposing that there is a strong prior to fuse AV speech inputs as compared to treating them separately. Fusing AV information is clearly an optimal strategy if we think about how rarely we find incongruent AV speech arising from the same location in our environment, compared to how often we experience congruent (hence potentially helpful) AV speech information. Moving beyond the pure statistical prevalence of AV speech congruency, previous studies report the benefit of using AV information when perceiving speech during noisy situations (Ross et al., 2007; Sumby & Pollack, 1954) or when acquiring language, at very early stages of life (Teinonen, Aslin, Alku, & Csibra, 2008). Indeed, infants perceive AV speech illusions such as the McGurk effect very early in life (Burnham & Dodd, 2004; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum, Schmuckler, & Johnson, 1997), reinforcing the idea that AV speech is integrated by default (see Noppeney, Ostwald, & Werner, 2010 for a similar hypothesis formulated under the compatibility bias described by Yu, Dayan, & Cohen, 2009 or Nahorna, Berthommier, & Schwartz, 2012).

Multisensory integration has been recently framed in terms of predictive coding (see for example: Arnal & Giraud, 2012; Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011; Schroeder & Lakatos, 2009; Talsma, 2015; Wassenhove, Grant, & Poeppel, 2004). Following this idea our hypothesis may be explained and expanded also in those terms, by proposing that in AV speech integration we can make a very strong top-down prediction based on the massive experience producing and experiencing speech. Therefore when confronted with incongruent speech information, the first reaction within this framework would be to update the priors, or the model, to adjust to the new situation due to the failure of the predictions. In our case the update of the model, as suggested by our data, is to stop relying in the less informative modality by dampening its processing.

Our hypothesis is distinct from pre-attentive integration, as we propose that prior to AV integration enough processing of the unisensory information must occur, and the processing of the unisensory information, as supported by our data is not enough under unattended conditions. It is possible that in the past this tendency to integrate AV speech automatically has been confused with it being resistant to attentional manipulations when in fact the main issue was that attentional resources were not completely depleted, therefore enough resources were still available and AV stimuli were still sufficiently processed (Lavie, 1995) and thus, integrated.

Nonetheless a possible alternative explanation to our results can be that attention only acts in a modulatory way, and therefore MSI process still had an effect that our paradigm was not sensitive enough to detect in any of the experiments, neither in the behavioral or neural measures. Of course, such modulation already imposes a limit to AV integration, but it would mean that some MSI always occurs. Our data are not supportive though are neither able to completely negate this possibility.

5. Conclusions

Our results are in line with previous demonstrations that MSI is sensitive to the inner goals and voluntary direction of attention, and that this sensitivity can be generalized to attention in modalities away from visual and from spatial attention. Our data suggest that for the MSI network to express neurally or to manifest in behavioral enhancements it was not enough that AV congruent stimulus were present, but they had to be attended, thus suggesting attention to be a necessary factor. The neural expression of *attending* to AV stimuli encompassed both association brain areas as well as unisensory areas, auditory and visual, previously reported in literature. Brain areas associated with conflict detection were also active when attention was deployed to incongruent AV speech stimuli, implying that this incongruency is enough to activate conflict related processes. We propose that the AV integration process is automatic once the independent modalities are processed, and in the case of AV conflict regulation occurs at low-level areas.

6. Acknowledgements

This research was supported by the *Ministerio de Economía y Competitividad* (PSI2013-42626-P), AGAUR *Generalitat de Catalunya* (2014SGR856), and the European Research Council (StG-2010 263145).

7. References

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Frontiers in Psychology*, *5*(July), 727. doi:10.3389/fpsyg.2014.00727

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology : CB*, *15*(9), 839–43. doi:10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *183*(3), 399–404. doi:10.1007/s00221-007-1110-1

Alsius, A., & Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *213*(2-3), 175–83. doi:10.1007/s00221-011-2624-0

Andersen, Tobias, S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Communication*, *51*(2), 184–193. doi:10.1016/j.specom.2008.07.004

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–8. doi:10.1016/j.tics.2012.05.003

Beauchamp, M., Lee, K., & Argall, B. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. Retrieved from http://www.sciencedirect.com/science/article/pii/S0896627304000704

Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics*, *3*(2), 93–113. doi:10.1385/NI:3:2:093

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190–2. doi:10.1038/nn1333

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, *62*(2), 321–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11436735

Buchan, J. N., & Munhall, K. G. (2011). The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. *Perception*, *40*(10), 1164–1182. doi:10.1068/p6939

Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. *Seeing and Perceiving*, *25*(1), 87–106. doi:10.1163/187847611X620937

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–20. doi:10.1002/dev.20032

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 : CB*, *10*(11), 649–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10837246

Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, *11*(12), 1110–1123. doi:10.1093/cercor/11.12.1110

Calvert, G., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. MIT press.

Campbell, R., De Gelder, B., & De Haan, E. (1996). The lateralization of lip-reading: a second look. *Neuropsychologia*, *34*(12), 1235–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8951835

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975. doi:10.1121/1.1907229

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–73.

Déry, C., Campbell, N. K. J., Lifshitz, M., & Raz, A. (2014). Suggestion overrides automatic audiovisual integration. *Consciousness and Cognition*, *24*, 33–7. doi:10.1016/j.concog.2013.12.010

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, *381*(6577), 66–8. doi:10.1038/381066a0

Driver, J., & Noesselt, T. (2008a). Multisensory interplay reveals crossmodal influences on “sensory-specific” brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23. doi:10.1016/j.neuron.2007.12.013

Driver, J., & Noesselt, T. (2008b). Multisensory interplay reveals crossmodal influences on “sensory-specific” brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23. doi:10.1016/j.neuron.2007.12.013

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, *2*(7), 254–62. doi:10.1016/S1364-6613(98)01188-7

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, *19*(4), 676–685. doi:10.1214/088342304000000396

Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *The European Journal of Neuroscience*, *29*(6), 1247–57. doi:10.1111/j.1460-9568.2009.06688.x

Fox, P. T., Huang, a, Parsons, L. M., Xiong, J. H., Zamarippa, F., Rainey, L., & Lancaster, J. L. (2001). Location-probability profiles for the mouth region of human primary motor-sensory cortex: model and validation. *NeuroImage*, *13*(1), 196–209. doi:10.1006/nimg.2000.0659

Fujisaki, W., Koene, A., Arnold, D., Johnston, A., & Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings. Biological Sciences / The Royal Society*, *273*(1588), 865–74. doi:10.1098/rspb.2005.3327

Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex (New York, N.Y. : 1991)*, *20*(3), 583–90. doi:10.1093/cercor/bhp124

Jones, J. a, & Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport*, *14*(8), 1129–33. doi:10.1097/01.wnr.0000074343.81633.2a

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–84. doi:10.1016/j.actpsy.2010.03.010

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants, *105*(32).

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(November), 863. doi:10.3389/fpsyg.2013.00863

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(3), 451–68. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7790827

Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111–20. doi:10.1016/j.heares.2013.06.010