NeuroImage 119 (2015) 272–285

Contents lists available at ScienceDirect

NeuroImage

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

Top-down attention regulates the neural expression of audiovisual integration

Luis Morís Fernández [a,*,1], Maya Visser [b,1], Noelia Ventura-Campos [b], César Ávila [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

ARTICLE INFO

Article history:
Received 5 February 2015
Accepted 18 June 2015
Available online 26 June 2015

Keywords:
Multisensory
Audiovisual
Attention
Speech perception
fMRI
STS

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 while 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 automatically only when both modalities are sufficiently processed, and that if a mismatch is detected between the AV modalities, feedback from conflict areas minimizes the influence of this mismatch by reducing the processing of the least informative modality.

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 (Calvert et al., 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 et al., 2007; Sumby and 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 and 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 (Beauchamp, 2005; Calvert, 2001; Driver and Noesselt, 2008a; Fairhall and 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 et al., 2014; Alsius et al., 2005; Alsius et al., 2007; Alsius and Soto-Faraco, 2011; Andersen et al., 2009; Bertelson et al.,

* Corresponding author at: Dept. de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona, Spain.
E-mail address: luis.moris.fernandez@gmail.com (L. Morís Fernández).

[1] Both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.neuroimage.2015.06.052
1053-8119/

2000; Buchan and Munhall, 2011, 2012; Fairhall and Macaluso, 2009; Fujisaki et al., 2006; Senkowski et al., 2005; Soto-Faraco et al., 2004; Tiippana et al., 2011; Van der Burg et al., 2008; Van Ee et al., 2009; Vroomen et al., 2001; for reviews see: Navarra et al., 2010; Koelewijn et al., 2010; Talsma et al., 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 et al., 2008; although see Alsius and Soto-Faraco, 2011; Fujisaki et al., 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 et al., 2004, 2011; Zion Golumbic et al., 2013) or to visually crowded scenarios (Alsius and Soto-Faraco, 2011). A recent study by Nahorna et al. (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 et al., 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 et al., 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 and Macaluso, 2009; Senkowski et al., 2005; Talsma and 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 et al., 2004a; Calvert et al., 2000; Fairhall and Macaluso, 2009; Miller and D'Esposito, 2005; Nath and Beauchamp, 2012; Noesselt et al., 2012; Beauchamp et al., 2004a, 2012; Stevenson et al., 2010; Stevenson et al., 2011; Stevenson and 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 toward 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 Driver and Noesselt, 2008b; Macaluso and Driver, 2005; Schroeder and 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 and 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 and Macaluso, 2009, for example) as well as in unisensory regions (Zion Golumbic et al., 2013). This possibly reflects that attention has an impact at multiple stages of multisensory processing (Talsma and 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 congruent AV stimuli compared to when they attend incongruent AV stimuli. Behaviorally one would expect an increment in the word recognition rate when attention is directed toward 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.