# Audiovisual Speech Processing

*The Role of Attention and Conflict*

Luis Morís Fernández

---

Year of completion: 2016

PhD Supervisor:

Dr. Salvador Soto Faraco

Department of Information and Communication Technologies

upf. Universitat Pompeu Fabra Barcelona

A mi unidad familiar.

*"Aunque la halles pobre, Ítaca no te ha engañado.*
*Así, sabio como te has vuelto, con tanta experiencia,*
*entenderás ya qué significan las Ítacas."*

Kavafis, Konstantino. *Itaca*

## Agradecimientos

En estas doscientas páginas quedan reflejados cuatros años y medio de trabajo, no solo mío sino de mucha más gente.

Agradezco a Salva la oportunidad de realizar aquí mi tesis doctoral, de aprender de él cómo se hace ciencia y cómo sobrellevar la frustración que viene con ella. Agradezco su paciencia y psicoanálisis a lo largo de estos años, amén de su apoyo y comprensión plasmados en la recurrente expresión "Pero, ¿a ti cuánto te queda?".

Gracias a toda la gente del grupo, pasados y presentes, porque todos han puesto alguna idea en esta tesis. En especial a Manu, por haber disfrutado conmigo de la experiencia de ir al *Burger King* después del último participante de *reso* en Castellón. A Manu y a Mireia porque siguen sin salir corriendo al oírme decir "Tengo una pregunta facilita", o bien el condicionamiento no funciona tan bien como dicen, o bien tienen muy buena voluntad.

Gracias al grupo de neuroimagen de Castellón, con los que di los primeros pasos en el análisis de imagen funcional.

Gracias a Emiliano por haberme abierto las puertas de su grupo en Roma, por lidiar con el efecto McGurk, por sus consejos y su *feedback* en el segundo trabajo de esta tesis.

Gracias a Mario por sus pausas para *debatir sobre el estado de la nación*. Y gracias también a Miguel Lechón, que se le echa de menos por el despacho, incluso después de un año.

A toda la gente con la que he compartido despacho en las diferentes etapas, desde la más tranquila hasta la más *cachacachacachinski*.

Gracias a Florencia y Cristina por ayudarnos y explicarnos algo aún más indescifrable que los artículos científicos, el papeleo. A Silvia y Xavi, los amos de la mazmorra, por tenerlo todo a punto (no toquéis el material sin avisar a laboratoris.cbc@upf.edu).

Gracias a toda la gente del departamento porque muchos antes o después me han echado una mano.

Gracias a todos los participantes que hicieron cosas por 10€ que yo no haría por 100€.

Gracias a David Moratal que empezó conmigo esto de la investigación y me ha echado un cable a lo largo de estos años aunque sea a cambio de dulces asturianos.

A mis amigos de toda la vida, que han aguantado esta tesis, todo lo que me ha llevado a esta tesis y a todo lo que me lleve esta tesis. A Miriam por revisar la portada que el interlineado antes de ella era un crimen. En especial a Miguel por su compañía durante la aventura barcelonesa. Gracias a Nacho por estar ahí todos estos años y por supuesto, al resto de los *niños* y las *niñas*.

Y como siempre se dice que la ciencia se hace a hombros de gigantes, gracias a los gigantes que me han subido en sus hombros para que yo pudiera terminar esta tesis y ser lo que soy, que son mi unidad familiar: mis padres, mi hermano y Natalia. En especial a Natalia porque celebrar nuestro aniversario dentro de un escáner de resonancia viendo estímulos de McGurk y leerse la tesis entera para poner las comas en su sitio correspondiente es de mérito. Sin vosotros no sé si las cosas serían imposibles pero seguro que serían bastante más complicadas y tendrían bastante menos sentido.

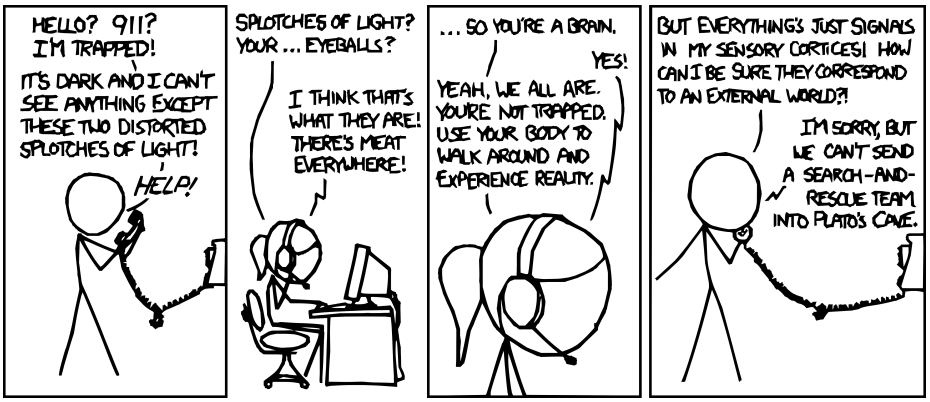
## Abstract

Events in our environment do rarely excite only one sensory pathway, but usually involve several modalities offering complimentary information. These different types of information are usually integrated into a single percept through the process of multisensory integration. The present dissertation addresses how and under what circumstances this multisensory integration process occurs in the context of audiovisual speech. The findings of this dissertation challenge previous views of audiovisual integration in speech as a low level automatic process by providing evidence, first, of the influence of the attentional focus of the participant on the multisensory integration process, particularly the need of both modalities to be attended for them to be integrated; and, second, evidence of the engagement of high level processes (i.e. conflict detection and resolution) when incongruent audiovisual speech is presented, particularly in the case of the McGurk effect.

## Resumen

Los eventos que suceden a nuestro alrededor, no suelen estimular una única modalidad sensorial, sino que, al contrario, suelen involucrar varias modalidades sensoriales las cuales ofrecen información complementaria. La información proveniente de estas diferentes modalidades es integrada en un único percepto a través del proceso denominado integración multisensorial. Esta tesis estudia cómo y bajo qué circunstancias ocurre este proceso en el contexto audiovisual del habla. Los resultados de esta tesis cuestionan los enfoques previos que describían la integración audiovisual como un proceso automático y de bajo nivel. Primero, demuestra que el estado atencional es determinante en el proceso de integración multisensorial. Más concretamente, presenta pruebas de la necesidad de atender a ambas modalidades, visual y auditiva, para que ocurra el proceso de integración. Y en segundo lugar, presenta pruebas de la participación de procesos de alto nivel (i.e. detección y resolución de conflictos) cuando existe una incongruencia entre la modalidad auditiva y visual, especialmente en el caso del efecto McGurk.

# Prologue




*Trapped,* comic strip by Randall Munroe[1], published in https://xkcd.com/876/

What is reality, how do we experience reality or even is there a reality at all are questions that we have asked ourselves many times but remain unsolved. What is reality and if it exists[2] escape the field of Neuroscience; nonetheless, how we experience it is one of the core questions of this field. Already Plato's Republic (380 BC) outlines an explanation of this experience in his *Allegory of the Cave*, where he describes a group of men who live chained in front of a blank wall. In this wall, the people can observe the shadows projected by things passing between them and a fire located behind them. Unable to see things themselves, the men perceive them only through their projected shadows. This same idea is beautifully (and comically) illustrated by Randall Munroe in the strip above–with a more neuroscientific and updated point

[1] This work is licensed under a Creative Commons Attribution-NonCommercial 2.5 License: http://creativecommons.org/licenses/by-nc/2.5/

[2] For the rest of this dissertation, the author will work on the assumption of the existence of a reality to be experienced, and will avoid any solipsistic approach, allowing the reader the possibility and the pleasure to exist.

of view–by substituting the shadows in the *Allegory of the Cave* by signals in our sensory cortices.

As a matter of fact, through the years, research and observation have proved that the same physical reality can be perceived in dramatically different ways depending on the expectations of the perceiver, his attentional focus or, in general, his inner state. Good examples of this are dramatic failures of perception, such as the inattentional blindness exemplified in the paper *Gorillas in the Midst* by Simons & Chabris (2000), or perceptual illusions such as the McGurk effect discussed in this thesis (McGurk & MacDonald, 1976). This demonstrates that human perception is hardly a direct function of objective reality but a combination of it with the observer's inner state. The *leitmotiv* underlying this dissertation is an effort to shed some light on how we experience reality by explaining how the inputs we receive are perceived and interpreted. Among all the possible experiences of reality, in this thesis we will focus on AV speech.

From the many daily activities we engage on, social interactions are probably among the most frequent ones, and among these social interactions, communicating through speech face to face with others is probably the most common of all. Although auditory signals surely convey most of the information during these direct interactions, speakers have parallel access to other sensory modalities (vision, for example) that can convey complimentary information. For example, the listener can use the lips of the speaker, or even their body or facial gestures, to

achieve a better understanding of the message. This makes AV speech one of the most characteristic and common multisensory experiences.

In the present thesis, I propose a general framework for the AV speech integration process with a special focus on its automatic/non-automatic nature. This framework focuses, first, on the influence of the observer's (i.e., the listener's) inner state on this process, and second, on how it takes place when we are presented with conflicting AV information instead of congruent AV information.

In the Introduction section, I will present relevant literature and its relation with the objective of this thesis to provide the reader with the context and motivation of the research questions addressed here. In a second, experimental section, I will describe three studies addressing the process of AV speech integration using both behavioral and neural measures (EEG and fMRI). In a third section, I will discuss the findings presented in this dissertation and their possible impact in the field of speech multisensory integration. In the fourth section, I will summarize the conclusions of this dissertation followed by the last section proposing possible future lines of research.

# Index

Page

## Abbreviations

AV: Audiovisual

ACC: Anterior Cingulate Cortex

IFG/lIFG: Inferior Frontal Gyrus / Left Inferior Frontal Gyrus

lPC: Left Precentral Cortex

EEG: Electroencephalography

fMRI: Functional Magnetic Resonance Imaging

STS: Superior Temporal Sulcus

BOLD: Blood Level Oxygen Dependent

# 1 Introduction

Almost any given event in our everyday life experience stimulates several sensory modalities. Think for example of a musician playing a piano; at least three different sources of information (vision–by watching his fingers playing–, audition–by listening to the note played–and touch–by the pressure on his fingers) are concurrently part of the experience. Indeed, according to Stein & Meredith, 1993, it would be challenging to find one single situation in which only one sense is involved. What is more, we do not experience these different sensory events as separate and independent, but integrated in a single unique percept (see Searle, 2000, for a philosphical perspective). In our brain, the sources of information coming from the different modalities are not only perceptually bound (i.e. packed together), but in fact they interact with each other, so the emerging perceptual experience is often a mixture of all modalities, different from each of the single modality components (Calvert, Spence, & Stein, 2004; Spence & Driver, 2004; Stein & Meredith, 1993).

One paramount example of multisensory integration is the perception of speech, in which visual information provided by a speaker can be crucial for the recognition and identification of the spoken message on the receiver's end, for example, when the auditory signal is degraded. This benefit when visual information accompanies sounds during speech perception has been known

for many years (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). This effect has been attributed, at least in part, to the close correlation between the acoustic temporal envelope and the articulatory gestures (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Grant & Seitz, 2000). The influence of visual information in speech is not restricted to lip movements; movements of the head, jaws or eyebrows of the speaker, and their hand and body gestures, have also been proven to be influential in the perception of speech (Biau & Soto-Faraco, 2013; McNeill, 1992, 2005). This visually-induced improvement in speech perception extends beyond situations in which the acoustic signal is physically degraded by external noise and has a benefit on hearing-impaired listeners (Grant, Walden, & Seitz, 1998) or even when listening to a second language (Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010).

Forty years ago, McGurk & MacDonald (1976) demonstrated that the role of visual speech input is not simply restricted to a complementary source of information that helps to improve perception when the auditory signal is weak, but it can also change auditory perception dramatically. This claim came about by means of one of the most famous audiovisual illusions: *the McGurk Effect* (see Massaro & Stork, 1998, for an explanation on the serendipitous origin of this effect). In the McGurk effect, an incongruency between visually and acoustically presented syllables produces the illusion of hearing a third, different syllable not corresponding to the auditory or visual ones. The classical combination used in the seminal study by McGurk and

McDonald is created by pairing the auditory syllable /ba/[3] with the video of a mouth uttering the syllable [ga]. This particular combination produces the illusion of listening to the syllable /da/, despite the syllable /da/ never being presented.

These two well-known phenomena (visual enhancement under noise and the McGurk effect) illustrate the intimate interplay between the auditory and visual information during the perception of speech. When the auditory and visual inputs are aligned (that is, they correlate), a benefit results, whereas when they are put in conflict, the perceptual system tends to resolve in a compromise between the two.

This thesis addresses two main issues regarding multisensory integration in the context of audiovisual speech. The first is if multisensory integration is an automatic process or if it can be modulated by the inner state of the perceiver, especially by their attentional focus. This is one of the main historical questions about audiovisual speech integration and multisensory integration as a whole. The second aspect, motivated by the results found in the first study carried out, is whether general mechanisms of conflict detection and resolution play a role in the perception of the McGurk illusion. The questions of automaticity and conflict processing are indeed interrelated, as conflict

[3] Throughout this manuscript, the visual part of a syllable will be written in brackets (i.e. [ba]) while the auditory part will be written between slashes (i.e. /ba/).

processing implies the involvement of executive processes and high level areas that would challenge the vision of AV integration as a low level automatic process, as it will be discussed in more detail during the scope section.

The remaining of this chapter will introduce the current state of research on both topics, followed by a description of the scope of the thesis and the general hypotheses.

## 1.1 The role of attention in multisensory integration

Upon a review of the multisensory literature, one can find instances of multisensory integration encompassing almost every pair of modalities and very disparate types of interactions. Famous examples of behavioral manifestations of multisensory integration are the *ventriloquist effect* (Bermant & Welch, 1976)—in which the spatial source of a sound is perceived to be closer to that of a co-occurring visual event than it really is—, *the rubber hand illusion* (Botvinick & Cohen, 1998)—in which a rubber hand located in an anatomically plausible position, and stimulated the same way as the observer's (hidden) hand, is perceived as real—, *the double flash illusion* (Shams, Kamitani, & Shimojo, 2002)—in which a single flash presented simultaneously with two fast beeps produces the illusion of perceiving two flashes instead of one—, or the previously mentioned *McGurk effect*. Therefore the occurrence of interactions between the different sensory modalities is a well supported and accepted fact (Stein, 2012). However, under which circumstances and how these interactions occur is still a topic of debate.

Indeed, one of the questions that has remained in the spotlight for more than two decades is the following: does multisensory integration occur independently of the observer's focus of attention? The relevance of this question becomes evident if we think of the problem of selective attention in everyday life multisensory environments. The classical problem of selective attention is that, because of our limited processing capacity, some selection must occur to be able to process the relevant subset of all currently available information. Now, in multisensory environments, the information comes in a wide variety of modalities and, in consequence, one could ask: out of all possible pairings between sensory events in different modalities that are available at any given moment (truly correlated or spurious coincidences), which ones are selected for integration?

This apparently simple and relevant question has been extremely controversial, as demonstrated by the amount of literature devoted to this subject and the disparity of results found (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; Driver, 1996; Fairhall & Macaluso, 2009; Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006; Senkowski, Talsma, Herrmann, & Woldorff, 2005; Soto-Faraco, Navarra, & Alsius, 2004; Soto-Faraco, Sinnett, Alsius, & Kingstone, 2005; C Spence & Driver, 1996; Talsma, Doty, & Woldorff, 2007; Talsma & Woldorff, 2005; Tiippana, Andersen,

& Sams, 2004; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; van Ee, van Boxtel, Parker, & Alais, 2009; Vroomen, Bertelson, & de Gelder, 2001; see Koelewijn, Bronkhorst, & Theeuwes, 2010; Navarra, Alsius, Soto-Faraco, & Spence, 2010; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010 for reviews on the topic). The next sections will provide an overview of the state of the art on this subject.

### *1.1.1 Behavioral studies*

The claims of automaticity in multisensory integration or lack thereof go far back, but only more recently have researchers tackled the question explicitly. Bertelson et al. (2000) and Vroomen et al. (2001) are two good examples among the first studies to study if multisensory occurs in an automatic fashion. In their experiments, the authors used the ventriloquist effect preceded by a manipulation of the focus of spatial attention. In Bertelson et al. (2000) participants were instructed to deliberately focus their attention towards or away from the multisensory event, while in Vroomen et al. (2001) the attentional manipulation was done by using a display containing a singleton that would automatically capture participants' attention towards or away from the multisensory event. Both studies failed to find any influence of the focus of spatial attention on the ventriloquist effect, and therefore concluded that multisensory integration occurred independently of said focus of attention, and regardless of the attentional manipulation being endogenous or exogenous.

Van der Burg et al. (2008) went one step further and demonstrated that a multisensory event could summon participants' attention in the so called *Pip and Pop effect*. In this case, participants searched for a target (a horizontal or vertical line) among several distractors (45º tilted lines) randomly colored red or green. Every ~900 ms a few distractors or the target randomly changed from green to red (or vice versa), with two particularities: first, no distractor changed color at the same time as the target; second, every time the target changed color it could be accompanied by a tone. Their results showed that, when the tone—completely uninformative of the spatial location of the target—was present, search times for the target were much faster than when the task was performed in absence of sounds. In fact, when the sound was present, search times became nearly independent of the number of distractors, whereas search times increased steeply with the number of distractors when the sound was not present. They concluded that the irrelevant sound and transient color change of the target were integrated into a multisensory event, thus creating an exogenous cue that captured participants' attention.

Having said that, a previous study by Fujisaki et al. (2006) had offered evidence contrary to that of Van der Burg et al. (2008). In Fujisaki et al. (2006), participants were asked to detect which of several items presented in a display varied one dimension (luminance in one experiment and rotation speed in other) synchronously with an auditory signal. In this case, the search time in their study increased with the number of distractors

present in the display independently of the presence or absence of the auditory signal. This suggested that the crossmodal coincidence could not be detected and integrated preattentively, and therefore, it could not aid performance by capturing participants' attention.

These opposing results show the dispute surrounding the role of attention in multisensory integration, with studies pointing in both directions. A controversial scenario is also found when we turn to audiovisual speech.

The McGurk effect has very often been used as a paramount example of AV integration in speech. Since its discovery, it has been described as an automatic low level effect. Already McGurk & MacDonald (1976) showed in their initial paper that, even when participants were informed of the mechanism behind the illusion, they could not prevent it from happening. Even more, they indicated that participants were unable to distinguish the McGurk stimulus from the normal one (although this inability to separate normal stimulus from McGurk ones has been questioned by other authors: Soto-Faraco & Alsius, 2007, 2009; van Wassenhove, Grant, & Poeppel, 2007). Further studies also showed that babies may experience the McGurk effect (Rosenblum, Schmuckler, & Johnson, 1997); that, despite a gender mismatch between the talking face and its voice, the McGurk effect still occurred in all its strength (Green, Kuhl, Meltzoff, & Stevens, 1991); or that it could evoke the mismatch negativity when measuring evoked related

potentials(Colin et al., 2002). Nonetheless, despite of the fact that all these manifestations may point to a strongly automatic nature of the McGurk effect, none of these studies addressed its independence of the focus of attention in a direct manner.

One of the initial studies dealing with the role of attention in AV speech integration was the one by Jon Driver in 1996. This study comprised three different experiments dealing with the ventriloquist effect in audiovisual speech, for brevity, only the first two will be described. In all experiments participants were required to shadow one out of two concurrently presented auditory streams of words—both produced by the same speaker. A movie displaying the face of the speaker producing one the two auditory streams was concurrently presented (see Figure 1). In a first experiment, both auditory streams were presented from the same spatial location (same loudspeaker) and the participants were asked to shadow the stream of words that matched the speaker's lips in the visual display. Performance—measured as the amount of words retrieved from the target stream—improved when the visual display was located spatially away from the loudspeaker emitting both auditory streams, as compared to when it was located just above it. This improvement was credited to the ventriloquist effect: by virtue of the spatially displaced visual display, the sound of the target word stream was also perceived as spatially displaced, and it was easier to segregate it from the distractor stream. The implication was that, for the ventriloquist effect to aid in spatial attentional selection, it must occur prior to it.

In a second and crucial experiment, each word stream came from different spatial sources, one on each side of the visual display. This time the target stream was selected spatially and the speaker's lips in the visual display always matched the distractor auditory stream. The results showed a degradation in performance when the lips of the speaker in the visual display were visible, if compared to when they were occluded. As in the previous experiment this change in performance was attributed to the ventriloquist effect which , in this case, brought the two word streams perceptually closer and made them more difficult to segregate spatially, akin to what occurs when the sounds are presented from physically close locations. This experiment argued in favor of attention-independent integration for two reasons: first, in this case participants could perform the task without the visual information, it was therefore task-irrelevant and ideally unattended (a result similar to that found by Soto-Faraco et al., 2004 using a speeded classification task); second, audiovisual integration in this case was detrimental for their task. Nonetheless, subsequent studies challenged this stance on the role of attention in AV speech integration.

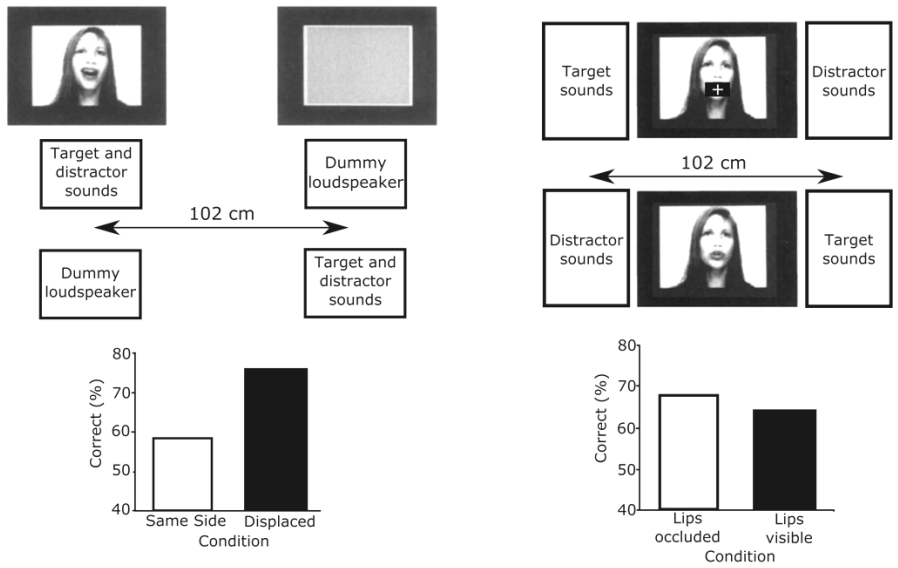



Figure 1. The left panel shows the increase in correct responses when the face is displaced with respect to the origin of the sounds by virtue of the ventriloquist effect. The right panel shows the opposite effect: a decrease in correct responses when auditory signals are brought together. Adapted from Driver (1996).

Tiippana et al. (2004) demonstrated that the McGurk effect was susceptible to attentional manipulations. Particularly, when the visual attention of the participant was diverted away from the speaker's face and onto a falling leaf superimposed on the video clip, the amount of fusion responses (i.e. */ka/* opposed to the unimodal response */pa/*) diminished, if compared to when the attention was not focused on the falling leaf but on the lips of the speaker (see Figure 2). This led them to conclude that attention was required for audiovisual speech integration to occur, a conclusion similar to that reached by the same group in another study using spatial attention (Tiippana, Puharinen, Möttönen, & Sams, 2011).

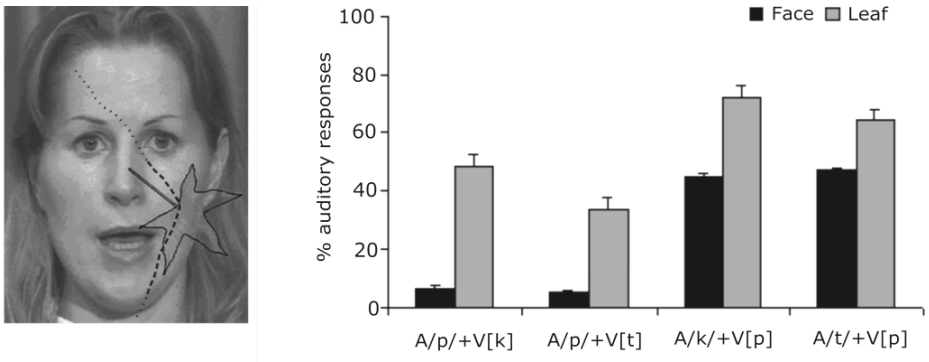



Figure 2 Figure adapted from Tiippana et al., (2004) paper showing the decrease in the perception of the illusion when attention was directed towards the superimposed falling leaf.

Noteworthy are the studies by Alsius et al. in 2005 and 2007 in which perceptual load was manipulated by comparing the proportion of McGurk responses in single vs. double task conditions. In their study, participants were presented with a McGurk-susceptible word (e.g. auditory *bate*, visual *gate*, possible fusion *date*) and a concurrent auditory or visual task—in the first study—, or a concurrent tactile task—in the second study. As in Tiippana et al. (2004), a reduction in fusion responses was found when participants engaged in a double task , if compared to when they did not. Interestingly, no drop in unisensory performance (i.e. when only the sound track or the video track of the speaker was shown) occurred under these dual task conditions. This would indicate that attention can affect the integration process itself, over and above any trivial, detrimental influence on unimodal processing.

In a recent study, Nahorna and colleagues (2012) proved that previous context can also influence the outcome of audiovisual

integration. They found that, when a McGurk stimulus was preceded by a stream of audiovisually incongruent syllables, the amount of fusion responses decreased, if compared to when it was preceded by a congruent audiovisual stream. This study can also be interpreted from an attentional point of view. If we assume that the incongruent context reduces the reliability of the least informative modality–in this case, vision–, the effect can probably be accounted for by top-down regulation of the attention deployed on each modality.

Alsius & Soto-Faraco (2011) employed a similar paradigm to that used by Fujisaki et al. (2006) and Van der Burg et al. (2008), but in this case using audiovisual speech. In a first experiment, an array of speaking faces and a single voice were presented. In a second experiment, several spatially distributed voices and a single face were presented. In both experiments, participants' task was to detect if there was a face-voice match (first set of experiments) or to localize where the matching face or voice was located (second set of experiments). For the localization task, the search time for face-voice matches increased with the number of distractors. This happened regardless of whether the task was to locate the speaking face that matched a single voice out of several speaking faces, or to locate the voice that matched a single centrally presented speaking face out of several voices. A similar pattern was found when participants were asked to detect if there was a face-voice match when many faces were present (reaction times also increased with the number of distractors), but, surprisingly, the time required to accomplish the detection task

when several voices were presented was the same regardless of the amount of distractors present. This result implied that the effectiveness of the attentional capture caused by a multisensory event was dependent both on the task (detect or locate) and on the modality in which selection occurred (visual or auditory).

All the studies and results mentioned above, highlight the complex interaction between the inner attention state (spatial or modality focus and availability of resources) of the participant and the process of multisensory integration. One clear conclusion that is drawn from the results presented so far is that they are not convergent, which pinpoints that, probably, one of the main general questions addressed by this thesis–if multisensory integration is dependent on the attentional focus of the participant–is not bound to have a single, absolute answer. As it was put forward by Talsma et al. (2010), multisensory interactions can happen at a variety of stages in the information processing hierarchy, and hence, top-down processes (including attention) may have an impact at multiple levels and with different consequences.

Therefore, the objective is to test the relevant factors in this interplay, and discern their roles. Regarding the specific case of speech–the focus of this thesis–while some studies point out the apparent independence between attention and multisensory integration, later studies tip the balance in favor of the influence of attention on the multisensory integration process.

One strategy especially well suited to further advance our knowledge about this question is to go beyond behavior experiments and study the neural expression of these attentional manipulations, particularly in areas related to multisensory integration in previous studies.

### *1.1.2 Neuroimaging studies*

*Multisensory integration sites.*

Several brain areas have been found to respond to multisensory events, ranging from deep sub-cortical structures to neocortical areas (Stein, 2012). Although there is a rich and very important body of research on multisensory integration using animal models (Alais, Newell, & Mamassian, 2010; Kayser & Logothetis, 2007), its impact on the issue of speech perception is only indirect. Here, the discussion will concentrate mostly on human studies.

An area that studies have typically linked to multisensory information processing in humans, especially in the domain of speech integration, is the superior temporal sulcus. One of the first studies relating this area to multisensory integration in speech was the one by Calvert, Campbell, & Brammer (2000) using functional resonance imaging. In this study the superior temporal sulcus showed a bilateral supra-additive effect, that is,

BOLD signal during multisensory stimulation was higher than the sum of responses to auditory and visual stimulation alone[4].

Furthermore, other studies found that the STS not only showed a higher response when the multisensory condition was compared to the unisensory ones, but also showed different levels of activity when the relation between the auditory and visual inputs was manipulated: for example, the congruency between the auditory and visual inputs (Calvert et al., 2000; Fairhall & Macaluso, 2009; Pekkola et al., 2006), or the synchrony between them (Miller & D'Esposito, 2005; Stevenson, Altieri, Kim, Pisoni, & James, 2010; Stevenson, VanDerKlok, Pisoni, & James, 2011).

A notable contribution to the understanding of the role of the STS and of the neural processes underlying audiovisual speech processing was made by Miller & D'Esposito (2005). They manipulated the synchrony between the auditory and visual speech components while participants were asked to judge if they perceived the whole stimulus as being fused or not. Their results revealed a higher activity of the left superior temporal sulcus area when participants perceived the stimulus as fused compared to when it was perceived as unfused. Their study also pointed out to other areas being responsive to this difference in fusion perception, namely, the Heschl's gyrus, the middle intraparietal sulcus and the inferior frontal gyrus. Interestingly, this last area was the only one that showed a higher activity during the unfused

[4] Throughout this whole section and the rest of the thesis, when we speak about a higher activity or response in fMRI we will be referring to a higher BOLD signal.

percepts. The authors speculated that this could reflect a shift from an automatic spatiotemporal matching in posterior areas (intraparietal sulcus) to a more controlled processing in frontal areas (inferior frontal gyrus). This result fits well under a new approach, which will be described later on in this thesis (section 1.3.2).

The perception of the McGurk illusion has also been related to activity in the superior temporal sulcus (Szycik, Stadler, Tempelmann, & Münte, 2012). The higher the frequency with which this illusion was reported, the higher the activity in the STS observed (Nath & Beauchamp, 2012). The relationship between activity in the STS and the perception of the McGurk illusion has also been shown using brain stimulation: the disruption of the activity in the STS using transcraneal magnetic stimulation causes a decrease in the amount of perceived McGurk illusions occurs (Nath & Beauchamp, 2012).

Some authors have also highlighted the role of other areas, apart from classical multisensory convergence areas in the temporal lobe such as the STS, like the motor areas. The role of motor areas in decoding speech signals is classically predicted by the motor theories of speech perception (Liberman & Mattingly, 1985), and its modern varieties. Along this line, several studies have also related motor areas to the perception of audiovisual speech (Hasson, Skipper, Nusbaum, & Small, 2007; Skipper, Nusbaum, & Small, 2005).

Finally, recent studies have shown that brain areas classically considered as unimodal also respond to crossmodal manipulations (see Driver & Noesselt, 2008; Macaluso & Driver, 2005; Schroeder & Foxe, 2005 for reviews on the subject). The participation of unisensory areas in multisensory processes is nowadays often accepted, though the question of what their exact role is remains open. In fact, some authors have even suggested that the neocortex as a whole is essentially multisensory (Ghazanfar & Schroeder, 2006).

Summing up, a sparse network of brain areas has been described to be involved in the process of multisensory integration, the STS being one of its paramount exponents, but also including frontal motor and association areas as well as posterior sensory brain regions. After this succinct review of the relevant areas to multisensory integration, the next section addresses the effect of attention on the neural activity in this multisensory integration network.

*Attention and multisensory integration sites*

If multisensory integration is shaped by attentional processes then a logical step would be to test this by recording activity in areas previously recognized as multisensory integration sites under different attentional manipulations.

Talsma & Woldorff (2005) studied this issue with non-speech AV stimuli. They used a paradigm in which participants were induced to direct spatial attention towards or away from the location of an upcoming multisensory event. Their findings

revealed that the electrophysiological signature of audiovisual integration (responses to multisensory stimuli if compared to unisensory) was greater when attention was directed toward these stimuli. Even more, this attentional influence started as early as 90 ms post-stimulus and lasted up to 500 ms, hence encompassing multiple stages of the multisensory integration process.

In another study using fMRI and speech stimuli, Fairhall & Macaluso (2009) showed that both subcortical and cortical areas—heteromodal and unimodal—activated differentially depending on the focus of attention. In their clever design, participants were presented with two close-ups of speaking lips left and right of the center, and a voice that matched one pair of lips or the other was presented centrally (see Figure 3 left panel). They were asked to keep the gaze fixated in the middle of the screen, and therefore, participants could change from attending a congruent audiovisual stimulus to attending an incongruent one by covertly displacing their attention from the face that was congruent with the presented voice to the incongruent one, or vice versa (while all physical stimulation remained constant). This paradigm revealed that activity in the superior temporal sulcus, as well as in the superior colliculus, was higher when attention was directed towards the congruent face (see Figure 3 right panel). This result supported the idea that activity in multisensory areas was modulated by the deployment of attention. Moreover, they also found that this attention modulation happened in visual unisensory areas as low as V1.

Nonetheless, as mentioned by the authors, behaviorally speaking, participants did not engage into any language task; a visual detection task was introduced only to ensure participants adhered to the instructions. This lack of language task implies that there was no measure of the effect of this attentional manipulation–and by extension of AV congruency– on the comprehension of the spoken message.

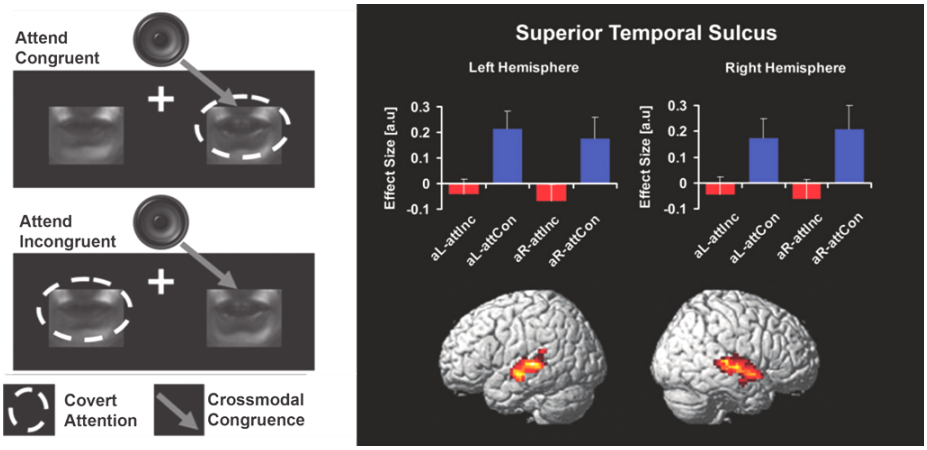



Figure 3. Figure adapted from Fairhall & Macaluso, (2009) showing an increase in activity in the STS when attention was directed towards the Congruent AV stimulus.

Zion Golumbic et al. (2013) addressed the question of the relation between attention and multisensory integration by using magnetoencephalography. Two auditory messages and the corresponding speaking faces were presented to the participants. The two auditory messages were presented centrally, and the two faces were presented spatially in the screen–left and right–while one message was produced by a man and the other by a woman. Participants were then asked to track one of the voices with or without the visual information. Authors then calculated which of

the two auditory messages contributed more to shape the neural signal recorded during the experiment. Their results showed that the attended message was more influential in shaping the neural signal only when the visual information was also present, but no effect was detected when the auditory information was presented alone.

To summarize, although multisensory integration seems to be malleable by attention, as seen in both behavioral and neuroimaging studies, how this influence occurs is still unclear. The type of stimuli used to evaluate said influence seems to be determinant. In *beep and flash* experiments using simple stimuli, attentional influence is harder to find behaviorally (Bertelson, Vroomen, de Gelder, & Driver, 2000; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; Vroomen, Bertelson, & de Gelder, 2001; although check for differences in EEG studies by Senkowski, Talsma, Herrmann, & Woldorff, 2005; Talsma & Woldorff, 2005). In speech, however, a clear influence by attention has been found both behaviorally and in neuroimaging studies (Alsius et al., 2005; Fairhall & Macaluso, 2009; Tiippana et al., 2004). Moreover, the particular modality in which attention is manipulated also seems to be influential (visual vs. auditory Alsius & Soto-Faraco, 2011). Therefore, it would be of interest to extend the results already found to different types of attentional selection. Another important aspect that remains pending is to measure behavioral and neural effects concurrently. These converging measures are essential to understand the relationship

between the neural and behavioral expression of any attention modulation of multisensory integration.

## 1.2 Multisensory integration and conflict

As it has been discussed above one of the most used markers of audiovisual speech integration is the McGurk illusion. Yet, one hardly recognized aspect of the McGurk effect is that this phenomenon is based on a conflict between auditory and speech inputs. This thesis addresses this aspect of the McGurk illusion, and hence, it provides a brief background on conflict.

In our daily routine we cope with many situations in which all the relevant information coming from our environment is congruent; therefore we engage in ordinary, almost automatic behaviors. One example would be the congruency between lip movements and the sounds produced by a speaker. However, when we are confronted with non-routine or challenging actions or when the information in our environment enters into conflict with our goals, the need of additional cognitive control to resolve these situations becomes evident. This might happen when there is a breach in the usual correspondence between lips and sounds, like in the McGurk illusion.

In section 1.2.1, I will review classical conflict tasks such as the Stroop and Eriksen Flanker tasks. In section 1.2.2, I will describe neuroimaging studies dealing with conflict and the commonalities between these classical conflict tasks and multisensory scenarios involving mismatching sources of information.

### 1.2.1 Behavioral studies about conflict

Three paradigms have widely been used to study how conflict is detected and resolved, and how it affects behavior and brain activity: the Stroop task, the Eriksen-Flanker task and the Simon effect.

In 1935 John Ridley Stroop proposed a very simple task: participants were presented with a written word denoting the name of a color, but the letters forming this word could be inked either in the same color denoted by the word or in a different one—**BLUE** (in blue ink) or **BLUE** (in red ink) (see Figure 4). When participants were asked to say aloud the written word the task could be done quickly and flawlessly. However, when they were asked to name the color of the ink a clear interference effect—measured as an increment in the reaction time—was found when the written word and the ink in which it was written differed—**BLUE**—compared to when both were the same—**BLUE**. This interference emerges because of the automatic tendency to name the written word that must be inhibited to perform the ink naming task correctly.

Eriksen & Eriksen (1974) reported a similar interference effect on a letter naming task when the target letter was flanked by task-irrelevant letters that acted as distractors (flankers). When the target letter was congruent with its flankers (i.e. both shared the same response), lower response times were found if compared to the incongruent situation (i.e. flanker and target did not share the same response, see Figure 4).

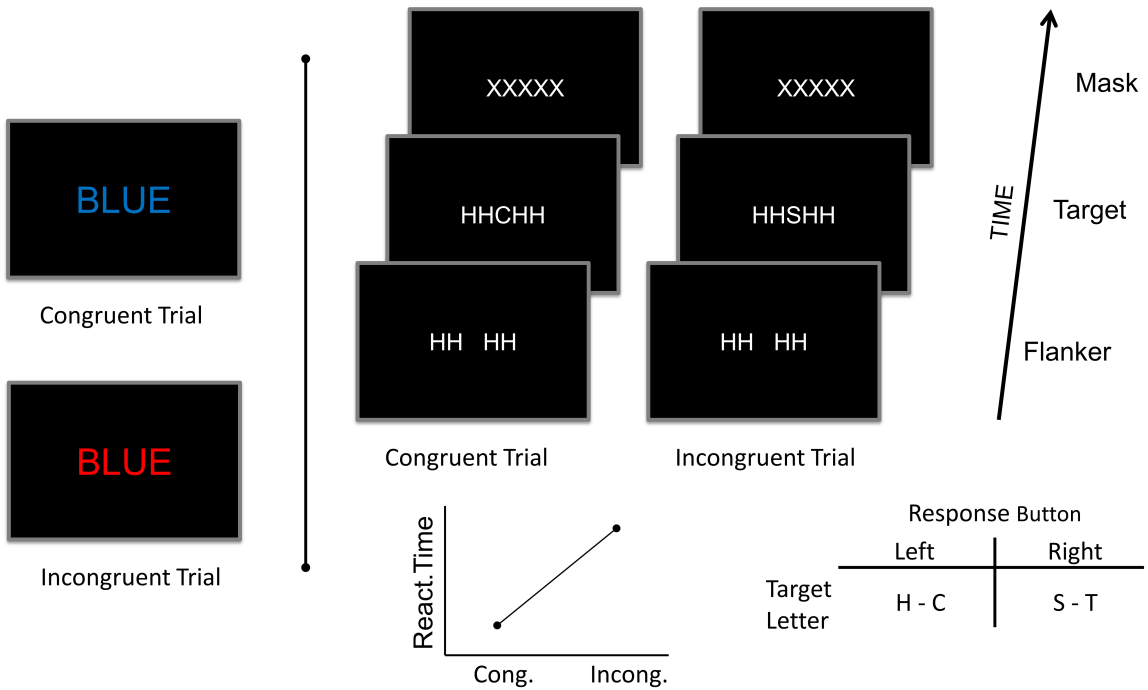



Figure 4. **Left panel:** Example of the two possible conditions in the Stroop task. Congruent, when both the word and the ink refer to the same color, and Incongruent, when the word and the ink refer to different colors. **Right Panel:** Example of an Eriksen Flanker task. While in the congruent trial both the distractors (H) and the target (C) share the same response button (the left one in this particular depiction), in the incongruent trial distractors (H) and target (S) have different response buttons. This conflict between the distractors and the target produces an increase in the reaction times when comparing the incongruent condition with the congruent condition.

Lastly, Simon & Rudell in 1967 reported for the first time the Stimulus Response effect, also known as the *Simon Effect*. In this paradigm, participants heard the words l*eft* or *right* and were instructed to respond using their left or right hand accordingly. This seemingly simple task became in fact daunting, as the words were presented randomly to the left or right ear. The spatial location was not relevant to answer to the identity of the word. However, their results showed that participants had a tendency to respond with the hand corresponding to the side of presentation of the word stimulus. For example, responding to the word *right* presented on the left ear was much more complicated (i.e. the reaction time was longer) than responding to the word *right* presented on the right side (see Figure 5 for an example with colors).

All these conflict tasks had a default action that can be reading (in Stroop), a prevalent response due to the distractor (in the case of the Eriksen Flanker) or to space congruency (in the Simon task). Conflict mechanisms are consequently engaged to override this default behavior in favor of the non-default, but correct one.

### *1.2.2 Neuroimaging of conflict tasks*

Two brain areas are repeatedly found to be involved in the processing of conflict in the tasks described above : the anterior cingulate cortex and the dorsolateral prefrontal cortex (Botvinick, 2007; Botvinick, Cohen, & Carter, 2004; Nee, Wager, & Jonides, 2007; Roberts & Hall, 2008; Shenhav, Botvinick, & Cohen, 2013).

Recent studies (Shenhav et al., 2013) have looked into the involvement of the anterior cingulate cortex (ACC) in the processing of conflict. The ACC has been assigned several roles: first, it detects the presence of a conflict; second, based on the relevant information it selects the optimal action from the set of available ones (e.g. override reading in favor naming the ink in the case of Stroop); and third, it recruits other areas (e.g. dorsolateral prefrontal cortex) that will be in charge of implementing the selected action to reduce the impact of conflict or solve it if possible.

These areas are activated in classical conflict tasks, but also on other situations where a conflict is present, such as multisensory tasks involving some type of incongruency between modalities. For instance, Noppeney and colleagues (2008) found a higher response of the anterior cingulate cortex and the left inferior frontal gyrus to visual primes followed by auditory targets when the visual primes (pictures or words) were incongruent with the auditory target (spoken word or natural sound of the object) than when they were congruent. A similar result was found by Orr & Weissman (2009), Weissman, et al. (2004) or Zimmer et al. (2010) using combinations of spoken letters and written letters.

*EEG studies*

Electrophysiological correlates of conflict processing in classical tasks have also been found using time-frequency analyses. During conflict perception, one common and replicated finding is a non-phase-locked increment of the power in the theta band (4-8

Hz) on the midfrontocentral electrodes (Cavanagh & Frank, 2014; Cohen, 2014; Ergen et al., 2014; Hanslmayr et al., 2008) as seen in Figure 5.

For example, Hanslmayr et al. (2008) ran a study in which participants performed a classical Stroop task while the EEG signal was recorded. They found that the oscillatory power in theta band increased as a function of the degree of interference (congruent, neutral, incongruent or negative priming). This increase began 400 ms after the stimulus onset and reached its peak after 800 ms. Its source was spatially localized in the ACC. A connectivity analysis in the same study revealed an increase in phase coupling between the ACC and the lateral prefrontal cortex in conflicting conditions. They interpreted these data by assigning the ACC a conflict monitoring and detection role, and the conflict resolution role to the lateral prefrontal cortex, based on the increased connectivity between these areas in conflicting conditions.

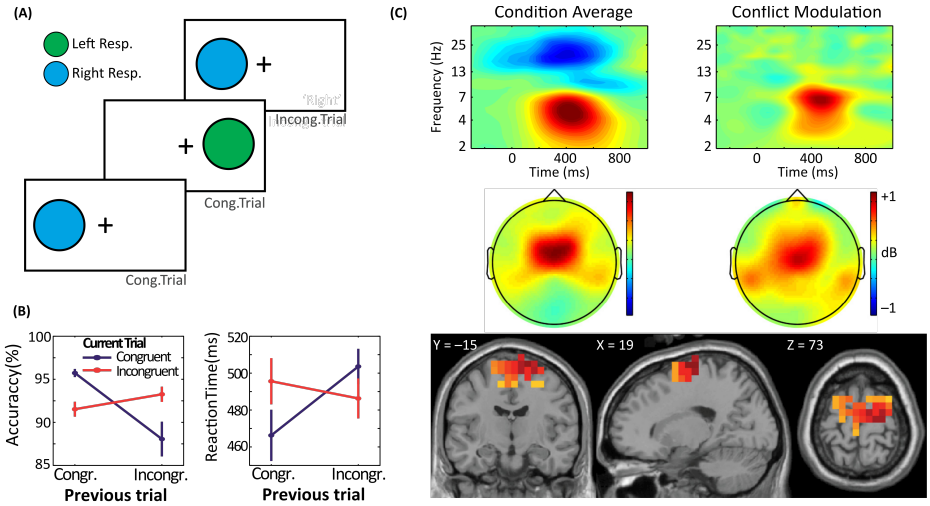



Figure 5. Example of a conflict task and representative findings. **A.** Simon task, in which subjects respond as quickly as possible to the color of the stimuli (task-relevant feature) while ignoring their location (task-irrelevant feature). During 'incongruent' (i.e., high conflict) trials the stimulus is on the opposite side of the required response. This increases reaction times and error rates. **B.** Typical behavioral findings observed on the Simon task. The effect of conflict on the current trial depends on the conflict in the previous trial. **C.** Typical EEG results from the Simon task, showing modulations in non-phase-locked theta-band power over midfrontal scalp electrodes (conflict modulation refers to the difference between incongruent and congruent trials). Brain-space estimation algorithms suggest a source of this conflict modulation in or around the supplementary motor area. Figure and caption adapted from Cohen (2014).

### *Neuroimaging studies using incongruent AV speech*

Activity in conflict related areas (i.e. ACC and IFG) has also been found, but rarely been discussed, in paradigms involving incongruent audiovisual stimulation using speech stimuli. For example, Miller & D'Esposito, (2005) (see page 16 for a description of the experiment) found that the anterior cingulate cortex and the anterior insula had a higher response to stimuli that were audiovisually asynchronous, and the left inferior frontal gyrus showed a higher response when the stimulus was perceived as not fused. Similar results were found also when speech sounds were

dubbed on mismatching visual speech (Morís Fernández, Visser, Ventura-Campos, Ávila, & Soto-Faraco, 2015; Ojanen et al., 2005; Pekkola et al., 2006; Szycik, Jansma, & Münte, 2009).

There are strong reasons to think that the role of conflict in audiovisual speech is very relevant, particularly in the McGurk effect. In spite of the popularity of the McGurk effect as a multisensory model (e.g. Alsius et al., 2005; Bernstein, Auer, Wagner, & Ponton, 2008; Hasson et al., 2007; Munhall, ten Hove, Brammer, & Paré, 2009; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Soto-Faraco et al., 2004; Tiippana et al., 2004; van Wassenhove et al., 2007), a notable difference with regards to regular audiovisual integration is that the McGurk stimulus is, by definition, incongruent. Nonetheless, there has been surprisingly little explicit effort in framing the McGurk effect within the conflict literature.

Interestingly, activity in areas previously related to conflict has also been described in several fMRI studies dealing with the perception of the McGurk illusion, albeit it has rarely been discussed, (Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam, 2010; Bernstein, Lu, & Jiang, 2008; Malfait et al., 2014; Matchin, Groulx, & Hickok, 2014).

In a different tradition, other studies have addressed the perception of the McGurk illusion by means of electroencephalography (Colin et al., 2002; Colin, Radeau, Soquet, & Deltenre, 2004; Sams et al., 1991; van Wassenhove, Grant, & Poeppel, 2005). Of special interest are two recent works, the first

by Keil, Müller, Ihssen, & Weisz (2012) using magnetoencephalography, and the second by Roa Romero, Senkowski, & Keil (2015) using EEG, in which they presented participants with McGurk stimuli.

Keil et al. (2012) found that trials in which the illusion occurred were preceded by an increase in beta power in a distributed network including the superior temporal gyrus, the inferior frontal gyrus and the precuneus. Critically they also found that a frontoparietal network (that included the ACC) was involved during the integration and fusion of the AV information, mainly in the form of a complex coupling-decoupling in the beta band.

In the second study Roa Romero and colleagues compared illusory McGurk perceptions versus congruent trials and found, first, a decrease in amplitude in the N1 evoked related potential; second, an early modulation (0-500 ms) in the beta band in the form of a decrease in power during the illusory McGurk trials; and third, a late modulation in the beta band (500 – 800 ms) also in the form of a decrease power when comparing illusory versus congruent trials. Based on these three effects that unfolded over time they proposed a three-stage process, the first one being the impact of visual context, indexed by N1; the second one, indexed by the early modulation in the beta band, the detection of the AV conflict due to the violation of the visual prediction, followed by the allocation of resources to solve said conflict; and the third one the resolution of the conflict by means of integrating the AV

information and forming a new percept indexed by the late modulation of the beta band.

Remarkably, no attempt to use the theta band power increase as a marker of conflict in the context of incongruent AV speech has been made (to the author's best knowledge).

Summing up, the McGurk stimulus is atypical since it is an incongruent AV stimulus, but very often it has been used to study properties of AV speech integration. Studies dealing with AV incongruency have found that classical areas related to conflict detection and resolution were activated in the case of an AV mismatch. Therefore, it would be of interest to study whether conflict mechanisms are engaged during the perception of the McGurk illusion or not.

## 1.3 Scope and hypotheses

It is well agreed that interactions between the different senses occur during the perceptual process, all falling under the wide umbrella of multisensory integration. Nonetheless, how this integration occurs and under which circumstances is still a subject under research. Among the many possible interactions between senses and possible types of stimulation, this thesis will focus on the domain of audiovisual speech.

This thesis addresses three different but closely interwoven hypotheses integrated in a tentative framework, with the goal of improving our understanding of the interplay between attention

and multisensory integration (see Figure 6 for a diagram of this framework):

1.Integration in AV speech will occur only if both modalities of the stimulus, the auditory and visual inputs, are attended.
2.If both modalities of the AV stimulus are attended an attempt to fuse them together will be made regardless of the congruency between the auditory and visual inputs.
3.If an AV conflict is detected (i.e. auditory and visual inputs are incongruent) conflict resolution processes are engaged to reduce the impact of this conflict in the final percept.

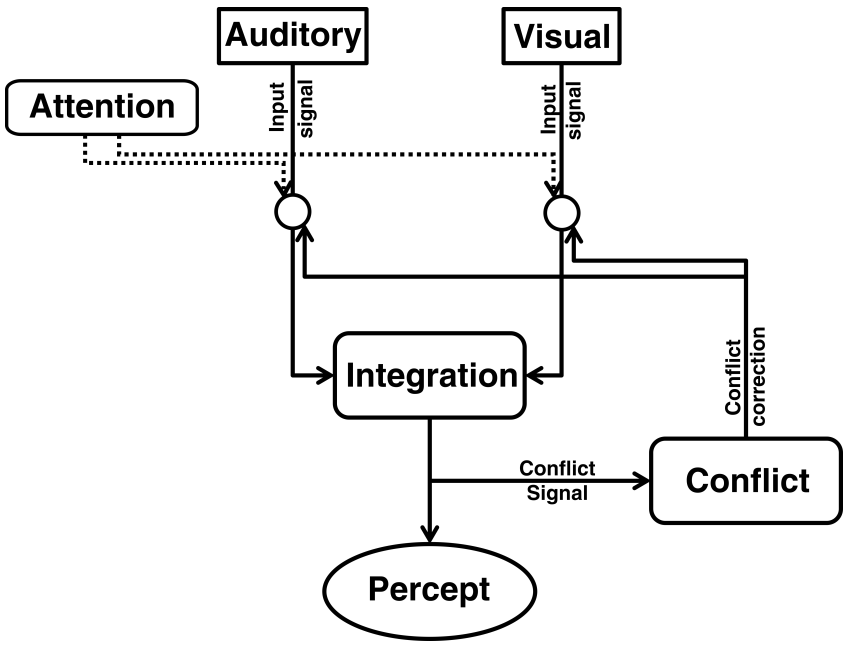



Figure 6. Diagram of the proposed framework. In this framework an automatic attempt to integrate the two inputs, auditory and visual, occurs if, and only if, both modalities are attended. If the integration attempt is successful the integrated percept is created; otherwise, the conflict is detected and a correction of the initial input is made followed by another attempt of integration.

The present framework has two main implications. First, it encompasses results that were concluded as preattentive in the past by describing the AV integration phase as automatic but only when both inputs are attended. Second, it implicates high level executive processes (i.e. conflict detection and resolution) in the perception of incongruent AV speech, challenging the view of AV integration as a low level automatic process. This is especially relevant in the case of the McGurk.

As mentioned before, the McGurk effect has been described since its discovery as result of automatic low level processes; however, this framework explains the McGurk effect as the outcome of a conflict detection and resolution process, high level and non-automatic. On top of this, it also brings to debate the use of this effect as an AV integration paradigm, as different neural routes–although probably partially overlapping–may be engaged depending on the congruency between the auditory and visual modalities (i.e. conflict detection and resolution are activated during an incongruent AV stimulation but they are not during congruent AV stimulation).

In the following sections, I will develop and operationalize the hypotheses presented in this framework given the state of the art described during this Introduction section.

### *1.3.1 The role of attention in the multisensory integration process*

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.

The first hypothesis of this thesis is that attention is necessary for multisensory integration to occur in the context of AV speech. This hypothesis was tested in the first study (section 2.1). In this paradigm, participants' attention was directed towards a congruent or an incongruent stimulus while keeping the amount of information in the display constant. If attention is needed for multisensory integration to take place, then activity in areas previously related to multisensory integration (especially the STS) should be modulated depending on the focus of attention of the participants. Specifically it was expected to observe high STS activity if the congruent stimulus was attended, and low STS activity if the incongruent stimulus was attended.

The results obtained in this first study supported the hypothesis of attention being needed for AV speech integration to occur, as no behavioral or neural correlate of an AV integration effect was found when the AV congruent stimulus was unattended.

A relevant finding in the results of this first study was that activity within a brain network previously related to conflict was found (ACC and insula) when participants attended incongruent AV stimuli, if compared to when they attended to congruent AV inputs. This data was interpreted, post-hoc, with the following explanation: when an observer is presented with AV speech information (i.e. a voice and a speaking face), an attempt to fuse this information, independently of its congruency, is made. When this information is incongruent conflict processes are activated to minimize the impact of the AV mismatch. This finding motivated the second part of this thesis where the possible relation of the conflict network and the perception of incongruent AV speech, in particular the McGurk illusion, were studied.

### *1.3.2 The role of conflict in the process of AV speech integration*

Based on the results of the first study of this thesis (section 2.1), I hypothesize that, whenever we perceive and attend AV speech information, we make an attempt to integrate the AV inputs regardless of their congruency and, if a conflict (AV mismatch) is detected, conflict resolution processes are put into play to reduce the impact of this mismatch. Particularly, in the case of McGurk, I hypothesize that, if the incongruency between the auditory and visual inputs is detected, activity in these areas will be observed, again as compared to AV congruent stimuli. Moreover, if these conflict areas are involved in the perception of the McGurk illusion, differential activity between illusory and non-illusory

outcomes should also be observed. To test these hypotheses, two different studies are presented, one using fMRI and the other EEG (sections 2.2 and 2.3).

In the first study, a higher activity in areas previously related to conflict such as the ACC and the IFG was expected when comparing McGurk stimuli with regular AV congruent stimuli. If, as hypothesized, these areas were not only related to the detection of the AV conflict, but also involved in its resolution and the creation of the illusory percept, then differential activity, between trials in which the McGurk illusion was perceived and those in which the illusion was not perceived, was expected.

In the second study a similar logic was followed. In this case, however, the index of conflict was an EEG marker: an increase in the theta band power in the midfrontocentral electrodes. In the case that the McGurk effect was perceived as a conflict, then one would expect this power increase in the theta band, when comparing the McGurk effect with AV congruent stimuli.

In both cases, the data supported the hypothesis and suggested that conflict areas are active and involved in the perception of the incongruent AV speech, particularly in the perception of the McGurk illusion. In fact these areas showed the expected differential activation depending on the illusory or non-illusory outcome of the McGurk effect. This differential activity indicates that their role may not only include the detection but also the resolution of this AV conflict in this case by reaching a compromise between the auditory and visual information.

Each of the empirical studies reported in the following section of this thesis are presented in the form of articles. They are either already published, submitted, or about to be submitted.

# 2 Experimental Studies

## 2.1 Top-down attention regulates the neural expression of audiovisual integration.

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S.

Top-down attention regulates the neural expression of audiovisual integration.

*NeuroImage 2015*, *119*, 272–285.

doi:10.1016/j.neuroimage.2015.06.052

Morís Fernández L, Visser M, Ventura-Campos N, Ávila C, Soto-Faraco S.
Top-down attention regulates the neural expression of audiovisual integration.
Neuroimage. 2015 Oct 1;119:272-85.
doi: 10.1016/j.neuroimage.2015.06.052.

## 2.2 Audiovisual integration as conflict resolution: The McGurk Illusion engages the conflict processing network

Morís Fernández, L., Macaluso E. & Soto-Faraco, S.

Audiovisual integration as conflict resolution: The McGurk Illusion engages the conflict processing network

Submitted to *Cortex*

# Audiovisual integration as conflict resolution: The McGurk Illusion engages the conflict processing network

Luis Morís Fernández [a]

Emiliano Macaluso [b]

Salvador Soto-Faraco [a, c]

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

[b] Neuroimaging Laboratory, Santa Lucia Foundation, Rome, Italy

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

Corresponding author: Luis Morís Fernández

Dept. de Tecnologies de la Informació i les Comunicacions
Universitat Pompeu Fabra, Office 55.128.
Roc Boronat, 138
08018 Barcelona
Spain

+34 686 17 30 58

luis.moris.fernandez@gmail.com

## Abstract

There are two main behavioral expressions of multisensory integration in speech; The perceptual enhancement produced by corresponding speaker's lip movements, and the illusory change in the sound of a syllable when it is dubbed with incongruent lip movements, in the McGurk effect. These two demonstrations have been used very often as probes to study multisensory integration. We suggest that integration occurring during congruent AV speech perception and the McGurk effect are different in fundamental ways, which can be revealed by their expression in brain activity. More specifically we contend that the McGurk effect arises from the resolution of an AV conflict which engages, at least partially, a general-purpose brain network for conflict detection and resolution (anterior cingulate cortex and left inferior frontal gyrus). This network has been known from studies using different conflict paradigms in other domains (i.e. Stroop, Simon task). To test this hypothesis, we used fMRI to measure BOLD responses upon presentation of AV speech syllables. We manipulated the nature of the stimuli —McGurk or non-McGurk— and integration — integrated or not integrated— in a 2x2 factorial design. Our results indicate that the hypothesized conflict related network showed an increase in BOLD signal when AV conflicting stimuli were presented and, furthermore, this conflict related network showed differential activity depending on the perceptual outcome of the McGurk illusion. We conclude that the AV incongruence in McGurk stimuli triggers the activation of a conflict processing network, and that this network is critical in the resolution of the AV conflict and the posterior outcome of the McGurk illusion.

## Keywords

McGurk; fMRI; Conflict; Speech perception; Multisensory Integration; Audiovisual

# 1. Introduction

One of the most compelling as well as famous perceptual illusions in psychology is the cross-modal effect discovered by Harry McGurk and John MacDonald in the 1970's (McGurk and MacDonald 1976; see Massaro and Stork 1998 for a historical account of how this discovery was made). In the so-called McGurk illusion, an audiovisually (AV) conflicting speech event produces the auditory perception of an illusory syllable which often is neither the one presented acoustically nor the one presented visually. For instance, a sound track containing the syllable /ba/ dubbed onto a video track of a mouth pronouncing the syllable /ga/ may produce the distinct perception of the syllable /da/. Indeed, this illusion demonstrates that visual information can dramatically influence the auditory identity of the perceived syllable, even under good listening conditions (e.g., (e.g. Campbell 2008). Above and beyond its curious phenomenology, this illusion has had a tremendous impact both at a theoretical (in speech perception and multisensory integration research) as well as at an empirical level, as a tool used in investigation. The McGurk effect has been used very often in the context of multisensory integration (MSI) studies as a proxy of AV integration in speech or MSI in general (e.g., Alsius et al. 2005; van Wassenhove, Grant, and Poeppel 2007; Skipper et al. 2007; Tiippana, Andersen, and Sams 2004; Bernstein et al. 2008; Andersen et al. 2009). This effect has proved to be very useful because its perception may vary from trial to trial in the same participant; sometimes the AV mismatch is resolved either as an auditory percept (e.g. /ba/), usually considered a not integrated percept, or as a *fused* (intermediate) percept (e.g. /da/), usually considered an integrated

percept considered a perceptual compromise between the auditory (e.g. /ba/) and visual (e.g. /ga/) conflicting inputs. This variability offers a probe to asses if in a particular trial the AV integration process was successful or not, or in average how often did integration occur. Interestingly, not so many authors have raised the possibility that this might be a distorted model of perception, which needs to be carefully used for generalization to audiovisual integration in normal contexts. Even more, in most behavioral, as well as EEG and fMRI studies, the perception of the McGurk illusion has been often equated to successful AV integration, putatively equivalent to that of a congruent AV stimulus. This assumption has been supported by the idea that the McGurk effect, like other cross-modal illusions, was a consequence of the brain's automatic engagement of multisensory integration mechanisms in a strongly mandatory fashion (Colin et al., 2002; Dekle, Fowler, & Funnell, 1992; Kislyuk, Möttönen, & Sams, 2008; Bernstein, Auer., 2004; Massaro, 1987; McGurk & MacDonald, 1976; Soto-Faraco, Navarra, & Alsius, 2004). Our claim in the present paper is that, despite in both cases (AV congruency and McGurk) there is AV integration their underlying neural processes are very different, due to the conflicting nature of the McGurk stimuli.

Many studies have addressed the neural expression of conflict processing using a variety of different protocols such as the Simon, Stroop or Go/No Go tasks. One of the main recurrent findings to arise from these studies is the involvement of the anterior cingulate cortex and the dorsolateral prefrontal cortex (see Shenhav, Botvinick, and Cohen 2013 for a proposed model on the role of the anterior cingulate cortex or Nee, Wager, and Jonides 2007 for a meta analysis on the neuroimaging studies regarding conflict). Even in the particular case of

audiovisual conflict, the anterior cingulate cortex emerges as one of the main implicated brain areas (Noppeney, Josephs, Hocking, Price, & Friston, 2008; Orr & Weissman, 2009; Weissman, Warner, & Woldorff, 2004; Zimmer, Roberts, Harshbarger, & Woldorff, 2010). For instance, in a study by Noppeney et al. (2008) visual primes followed by incongruent auditory targets were found to activate the anterior cingulate cortex and the left inferior frontal gyrus more strongly than congruent pairings, irrespective of the actual nature of either prime (written words or pictures) or target stimuli (spoken words or sounds). Upon a review of previous fMRI literature addressing AV speech where there is cross-modal conflict (e.g. congruent vs. incongruent syllables; synchronous vs. asynchronous, etc.), we have found that many of the studies have in fact reported the engagement of a network of brain areas very similar to that found during other kinds of conflicting situations, including the anterior cingulate cortex (ACC) and the left inferior frontal gyrus (lIFG) (Miller & D'Esposito, 2005; Morís Fernández, Visser, Ventura-Campos, Ávila, & Soto-Faraco, 2015; Ojanen et al., 2005; Pekkola et al., 2006; Szycik, Jansma, & Münte, 2009) Remarkably, if we focus on neuroimaging literature using the McGurk effect, activity in these conflict-processing areas has been often reported, albeit rarely interpreted (Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam, 2010; Bernstein, Lu, & Jiang, 2008; Hasson, Skipper, Nusbaum, & Small, 2007; Malfait et al., 2014; Matchin, Groulx, & Hickok, 2014). Recent EEG evidence has also pointed to the role of a frontoparietal network, involving the cingulate cortex, to be relevant in the perception of the McGurk illusion (Keil, Müller, Ihssen, & Weisz, 2012).

Given the obvious presence of conflict in McGurk stimuli, and the corresponding activation of conflict-sensitive brain areas seen in neuroimaging studies, it is quite remarkable that the possible role of conflict during the McGurk effect has been mostly overlooked to date. One exception is a recent EEG paper by Roa Romero, Senkowski, and Keil (2015), who studied the neural signature of the perception of the McGurk illusion. Based on their data Roa et al. proposed a three stage process indexed by, first, a reduction of the N1 evoked by the auditory component of the syllable due to the impact of visual context in AV in MSI speech processing; second, an early beta power suppression that indexes the detection of the AV conflict and the allocation of resources; and, third, a late beta power suppression that reflects the resolution of the AV conflict and the formation of the McGurk illusion.

Here, we hypothesize that the McGurk illusion is formed due to the resolution of the conflict between the auditory and visual information. If this hypothesis turns out to be true, we further hypothesize that the brain network involved in integrating non-conflicting (i.e. congruent) AV stimuli is not equal, although probably shares common areas, as the one involved in processing conflicting AV. Moreover if this network is not only involved in the detection of the conflict but in its resolution we expect to see a differential activation depending on the perceptual outcome (illusory or not) of McGurk stimuli which are physically identical.

To this end we manipulated two factors: stimulus nature (McGurk and non-McGurk) and occurrence of AV integration (integrated vs. non-integrated). We measured blood oxygen level dependent (BOLD) responses with functional magnetic resonance (fMRI) while

participants were asked to identify auditory syllables presented in four different experimental conditions that comprised: non-McGurk integrated stimuli (e.g. A /ba/ + V /ba/); non-McGurk non-integrated stimuli (e.g. A /ba/ + reversed video /ba/); McGurk integrated (e.g. A /ba/ + V /ga/ when participants perceive /da/ or /ga/); and McGurk non-integrated (e.g. A /ba/ + V /ga/ when participants perceive /ba/). The main interest of this study was to characterize which brain areas displayed differential effects due to AV integration depending on the nature of the stimuli (i.e. statistically speaking an interaction between the two factors). According to previous literature, if conflict processing mechanisms are triggered during AV speech conflict, then classical conflict areas, such as the ACC and IFG, must be active in the face of McGurk stimuli similar to those engaged under other conflict situations. More specifically, in this study we aimed at finding an influence of these conflict processing mechanisms in the perception of the McGurk illusion. If conflict mechanisms are involved in the perception of the McGurk illusion then we expect to find an interaction between the nature of the stimulus and AV integration. Therefore, we anticipate that this conflict processing mechanisms will be engaged by AV conflict during a McGurk illusory trial (integration with conflicting input), but not in the AV congruent ones (AV integration, without conflict). In addition, our design allowed us also to check for differential activation for AV integration with respect to no integration, both in regular AV congruency and McGurk integrated (illusory) stimuli. According to previous literature, one area that is frequently sensitive to this kind of multisensory integration effect is the pSTS (e.g., Nath and Beauchamp 2012).

# 2. Methodology

## 2.1. Participants

Twenty participants (11 females, mean age= 25.5, std = 3.8) were recruited for the study. All were right-handed, reported normal hearing and normal/corrected to normal vision. All participants gave written informed consent to participate in the study. The study was approved by the independent ethics committee at Fondazione Santa Lucia (Scientific Institute for Research Hospitalization and Health Care).

## 2.2. Stimuli and conditions

Stimulus material consisted in a series of videos of a woman pronouncing the syllables /ba/, /da/ and /ga/. The videos showed the lower part of the face of the speaker (720x576, 1.5s, with white noise at 65/50-dB S/N ratio) These stimuli have been used previously in other studies (Freeman et al., 2013; Soto-Faraco & Alsius, 2009). Four different experimental conditions were created (auditory syllables are denoted within slashes //, and visual syllables within brackets [], subscript R denotes reversed:

- Non-McGurk integrated stimuli (nMI): /ba/+[ba] or /da/+[da]
- Non-McGurk non-integrated stimuli (nMnI): (/ba/+$[ba]_{R}$ or /da/+$[da]_{R}$)
- McGurk stimuli non-integrated (MnI): /ba/+[ga] when perceived as /ba/.
- McGurk stimuli integrated (MI): /ba/+[ga] when perceived not /ba/

We created the incongruent condition based on those in the congruent condition with the exception that the video was reversed; this condition was created under the assumption it would be perceived as incongruent but would never be integrated (i.e. the perception will be always the auditory part of the stimuli).

## 2.3. Experimental procedure

Stimuli were delivered through a set of headphones (MrConfon Cambridge Research Systems) and presented on a monitor viewed through a mirror attached to the MRI head-coil (133 cm from screen to participant) and subtended 6.5° horizontally and 5.2° vertically. The participants' task consisted in watching videos of a speaker pronouncing the same syllable three times and, at the end of the audio-visual stimulation, to report what they had heard. The syllable was repeated 3 times with the aim of maximizing the BOLD response in each trial. The response was given in a four alternative forced-choice task (4AFC) by choosing one of four options (ba, ga, da or other). The participants answered using a four button response box, pressing one button with the middle or index finger, of the right or left hand. The order of the responses was counterbalanced across participants; they were instructed to pay attention to both the video and the audio but to respond strictly to what they had heard and not to what they had seen. The participants were also informed that the three syllables appearing during the video clip presented in each trial were identical.

### 2.3.1. Auditory trials

Prior to the beginning of the experiment, and with the same scanning sequence used during the main experiment, participants performed an

auditory-only task. During this task they were presented 18 auditory trials (9 /ba/ sounds and 9 /da/ sounds randomized, as in the AV stimuli syllables were repeated 3 times per trial) that they had to discriminate using the same 4AFC. This allowed us to regulate the intensity of the audio so each participant heard the stimuli at a comfortable level, and to assess that the audio was clear by itself.

### 2.3.2. Main task

The main task was divided in three runs each of them consisting in 80 trials, in random order, 20 nMI (10 /ba/+[ba] and 10 /da/+[da]), 20 nMnI (10 /ba/+$[ba]_{\mathbf{R}}$ and 10 /da/+$[da]_{\mathbf{R}}$), and 40 McGurk trials that afterwards were classified according to the participant's response in MI (not /ba/ percept) or MnI (/ba/ percept). Before each trial participants saw a 500 ms fixation point that alerted of the beginning of a new trial, followed by the 4.4 seconds long video clips, and a response period with a 2 second deadline. Inter stimulus interval was jittered uniformly between 2 and 4 sec. Each run lasted ~12 min and between the second and the third run a T1 structural scanner was acquired. Ten training trials were presented before the main task to ensure participants were familiar and understood the task correctly.

### 2.3.3. Participant and session selection

As the critical classification within the McGurk conditions as integrated (MI) and not integrated (MnI) was dependent on the participant's response and the perception of the McGurk illusion can be very variable across participant (Basu Mallick, F Magnotti, & S Beauchamp, 2015), we selected participants and sessions based on behavioral data recorded during the fMRI experiment. Five of the twenty participants

were discarded from further analysis as they did not perceive the McGurk illusion or did so very weakly (criterion: less than 12.5% of MI in all 3 sessions). For 12 of the remaining 15 participants we selected two sessions in which the participant perceived the McGurk illusion in at least 12.5% of McGurk trials (i.e. a minimum of 10 MI-trials per session). Three participants reached this criterion only in one single session. Theses participants/sessions were included in the main analysis to maximize statistical power (i.e., n = 15). However, to exclude any possible effect of the different number of sessions between participants, we also performed a control analysis now considering only the data from the 12 participants with two sessions each, and replicated all the main results reported below.

## *2.4. Image acquisition and analysis*

368 volumes per run were acquired in a Phillips Achieva 3T scanner, using an EPI sequence (FOV = 192 x 192 mm, Matrix Size = 64 x 64, Voxel Size = 3 x 3 x 2.5 + 1.25 mm gap, TR = 2.1 seconds, TE = 30 ms, 32 slices in ascending order), covering the whole brain. Image analysis was done using SPM8, ART toolbox[1], SOCKS toolbox (Bhaganagarapu, Jackson, & Abbott, 2013) and MarsBar (Matthew Brett, Jean-Luc Anton, Romain Valabregue, 2002).

### 2.4.1. Preprocessing

The first four image volumes of each run were discarded to allow for stabilization of longitudinal magnetization. Standard spatial

[1] http://www.nitrc.org/projects/artifact_detect/

preprocessing was performed for all participants following the subsequent steps: Horizontal AC-PC reorientation; realignment 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 (mean EPI to EPI template, voxel size was set to 3 mm, isotropic, normalization and interpolation was done using a 4th B-spline degree); functional data was smoothed using a 10-mm full width half-maximum Gaussian kernel to increase signal to noise ratio and reduce inter subject variability.

To further control for any residual effect of head-motion, we used the ART toolbox. With this we created an extra composite movement regressor, in addition to the six provided by SPM, that resumes the movement of the three rotations and three translations. Moreover, every volume meeting any the following conditions was marked as an outlier: a composite movement larger than 0.5 mm with respect to the previous volume or global signal 9 standard deviations away from the global mean of the run (5% on average per participant). The six standard SPM movement parameters, the ART composite movement regressor, plus any outlier volume were included as regressors of no interest in the first-level analyses, so that these effects would not influence the results of our analyses (see below).

As an additional measure to control for noise in the data we also applied an independent component analysis (ICA) method. For this we used the SOCK toolbox, that classifies each component as an artifact or not (Bhaganagarapu et al., 2013) then we removed the independent components classified as artifacts using fsl_regfilt provided by FSL MELODIC (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012).

### 2.4.2. First level analysis

At the first-level (subject-specific) analysis the time series for each participant was high-pass filtered at 128s and prewhitened by means of an autoregressive model AR(1). Box-car regressors modeling the occurrence of the four different conditions [ nMI, nMnI, MI, MnI], specified as events of duration 4.4 seconds corresponding to the presentation of the video, were convolved with the standard SPM8 hemodynamic response function. In addition, the effects of head movement produced by SPM, the effect of composite movement and the outlier regressors produced by ART were included. The resulting general linear model produced an image per session estimating the effect size of the response induced by each of the four conditions of interest per run. Only the selected sessions/runs per participant (see 2.3.3) were included in this first level analysis. For the 12 participants that contributed with two sessions, linear contrasts were used to average the parameter estimates across the two sessions, separately for the four conditions of interest.

### 2.4.3. Second level analyses

At the second (inter-subject) level, the contrast-images were entered into a random effects within-subject ANOVA modeling four conditions, plus the random effect (subject). We calculated the main effects (ME) for the factors stimulus nature (McGurk / Non-McGurk) and AV integration (integrated / non-integrated), and the interaction between the two that will be our critical test. Statistical parametric maps were assessed for cluster-wise significance using a cluster-defining threshold of $p<0.001$; cluster size was defined using random field theory (9.91x9.86x9.78 mm FWHM) obtaining a 37 voxel cluster size threshold for a Family-wise Error of $p<0.05$. In order to describe how the different conditions contributed to the interaction effect, we summarized the activity of the significant clusters using MarsBars (Matthew Brett, Jean-Luc Anton, Romain Valabregue, 2002). For each cluster we computed the mean of all the voxels in the activated cluster and reported pairwise t-tests between the 4 conditions. Brain figures were created using MANGO software[2].

# 3. Results

## *3.1. Behavioral data*

During the prescanning behavioral test using only auditory stimulation, the syllables were correctly identified on average in 84.8% of the trials.

[2] http://ric.uthscsa.edu/mango/mango.html developed by Jack L. Lancaster, Ph.D. and Michael J. Martinez.

This indicates that auditory information was sufficiently clear to properly identify the syllables used in the experiment.

| Non McGurk | | McGurk | | | |
|---|---|---|---|---|---|
| nMI | nMnI | MnI | MI | | |
| | | BA | DA | GA | OT |
| 93.5% | 70.0% | 49.6% | 36.3% | 8.9% | 2.5% |

**Table 1** Behavioral data corresponding to the experimental task as performed during the scanning session. Percentage correct identification of auditory syllables is presented in the Non-McGurk conditions (e.g. heard /ba/ responded /ba/). For the McGurk conditions the percentage of responses for each option is presented, a trial was considered not integrated (MnI) when the participant responded /ba/, in all other conditions /da/, /ga/ or others the trial was considered as integrated (MI).

The behavioral data in AV trials obtained during the scanning protocol is shown in Table 1 for the participants and sessions included in the analysis (15 participants, with a total of 27 sessions); identification in the nMI condition was almost perfect whereas performance in the nMnI condition was lower, indicating a difference between congruent and incongruent AV speech performance. Importantly, in the McGurk conditions approximately half of the responses were classified as integrated (45.2%, mostly reflecting fused percepts /da/ and some /ga/ responses[3]) while the other half of them was classified as non-integrated, reflecting auditory /ba/ percepts (see: Table 1). Please note that both visual dominated percepts /ga/ and fused percepts /da/ can be considered a visual influence on auditory perception, and therefore reflecting the McGurk illusion (according to some authors, any instance

[3] Only one participant classified consistently the McGurk syllable in the "other" category. When debriefed he reported listening to the syllable /fa/ when he pressed "other". He did not classify any syllable as other in any of the remaining conditions. This participant was included in the analysis.

in which the percept is different from the auditory syllable can be considered an instance of the McGurk illusion; see, Tiippana 2014) .

## 3.2. Imaging results

| Hemisphere | Region | Corrected Cluster *P*-Value | Number of Voxels | Z - Score | Coordinates (mm) x | y | z |
|---|---|---|---|---|---|---|---|
| ***Non-McGurk > McGurk*** | | | | | | | |
| R | Inferior Occipital | <0.001 | 2674 | 6.72 | 51 | -55 | -8 |
| L | Inferior Occipital | <0.001 | 1623 | 6.08 | -45 | -73 | -8 |
| L | Angular Gyrus | <0.001 | 251 | 4.64 | -48 | -61 | 46 |
| L | Hippocampus | <0.001 | 216 | 4.63 | -33 | -19 | -5 |
| R | Hippocampus | 0.004 | 64 | 4.15 | 21 | -31 | -2 |
| ***McGurk > Non-McGurk*** | | | | | | | |
| R | Anterior Cingulate Cortex | <0.001 | 355 | 4.88 | 6 | 20 | 37 |
| L | Anterior Insula | <0.001 | 91 | 4.42 | -27 | 23 | 7 |
| R | Anterior Insula | <0.001 | 184 | 4.35 | 36 | 23 | 4 |
| ***Interaction*** | | | | | | | |
| *MI – MnI > nMI - nMnI* | | | | | | | |
| L | Precentral Gyrus | <0.001 | 154 | 5.15 | -33 | -4 | 55 |
| L | Precentral Gyrus | <0.001 | 217 | 4.85 | -54 | 5 | 22 |
| L | IFG | | | 4.54 | -60 | 11 | 19 |
| L | SMA (ACC) | <0.001 | 115 | 4.98 | -3 | 8 | 55 |
| L | Anterior Cingulate Cortex | | | 4.42 | -6 | 17 | 43 |
| ***Interaction*** | | | | | | | |
| *nMI - nMnI > MI - MnI* | | | | | | | |
| R | Supramarginal Gyrus | <0.001 | 208 | 4.19 | 57 | -49 | 28 |
| R | Angular Gyrus | | | 3.85 | 57 | -52 | 37 |
| L | Angular Gyrus | <0.001 | 178 | 4.18 | -48 | -67 | 43 |

**Table 2.** Location Significance and extent for ME and interaction contrasts. P-values are family-wise error (FWE)-corrected at the cluster level. Not reported contrasts didn't have any significant result

### 3.2.1. Overall effect of non-McGurk vs. McGurk

The main effect of McGurk vs. non-McGurk (MI + MnI > nMI + nMnI) showed activation of the anterior insulae bilaterally and the ACC (see: Figure 1 and Table 2). The opposite contrast (non-McGurk vs. McGurk conditions: nMI + nMnI > MI + MnI; regardless of whether the percept resulted in an illusion or not) revealed an extensive bilateral network involving: the lateral occipital complex, posterior temporal lobe, angular gyrus and the hippocampus.

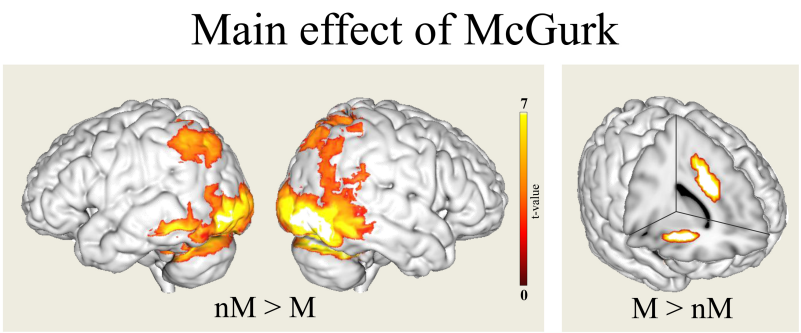



**Figure 1** t-maps showing the main effect of McGurk vs. Non-McGurk. [Non McGurk > McGurk] appears in the left panel, and [McGurk > Non McGurk] appears on the right panel. All t-maps are thresholded at p<0.001 peak level and are p<0.05 FWE cluster corrected (k=37).

| | Pairwise comparison (p-value) | | | | | |
|---|---|---|---|---|---|---|
| | nMI vs. | | | nMnI vs. | | MI vs. |
| ROI | nMnI | MI | MnI | MI | MnI | MnI |
| ACC | $<10^{-4}$ | $<10^{-4}$ | $<10^{-4}$ | 0.073 | 0.301 | 0.043 |
| lFG | $<10^{-4}$ | $<10^{-4}$ | 0.003 | 0.452 | 0.020 | 0.019 |
| lPC | $<10^{-4}$ | $<10^{-4}$ | 0.261 | 0.321 | $<10^{-4}$ | 0.005 |
| rAG | 0.001 | $<10^{-4}$ | 0.008 | 0.007 | 0.219 | 0.003 |
| lAG | $<10^{-4}$ | $<10^{-4}$ | 0.003 | 0.017 | 0.341 | 0.030 |

**Table 3** p-values for each of the pairwise comparisons made in each ROI with 42 degrees of freedom.

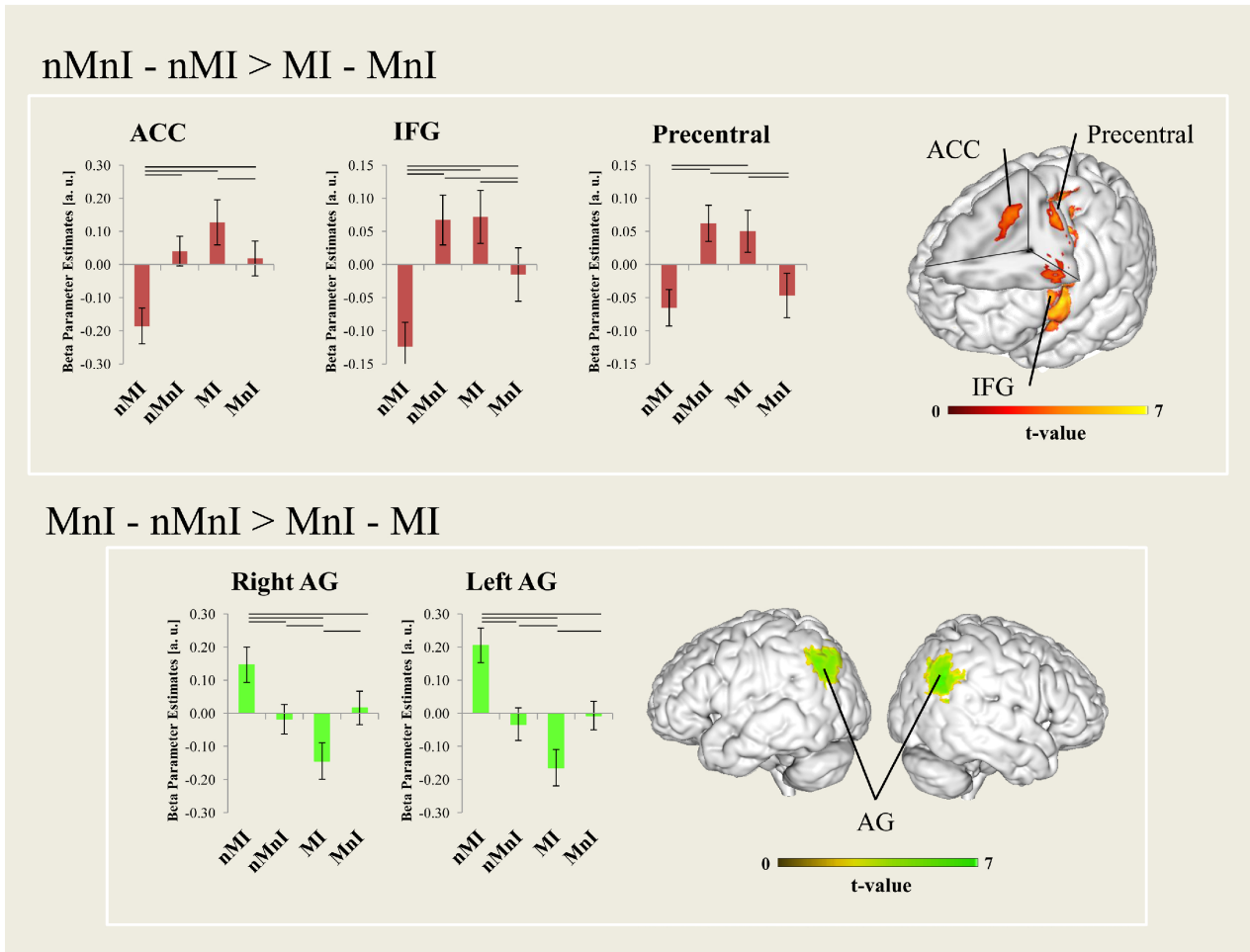



**Figure 2** t-maps showing the two interaction contrasts, and beta parameter estimates for each condition in each significant cluster. All t-maps are thresholded at $p<0.001$ peak level and are $p<0.05$ FWE cluster corrected ($k=37$). Significance in pairwise comparisons is represented by horizontal bars ($p<0.05$). Error bars represent the 90% confidence interval. (ACC: Anterior Cingulate Cortex, IFG: Inferior Frontal Gyrus, AG: Angular Gyrus).

### 3.2.2. Overall effect of integration

The effect of integration was calculated pooling congruent events with McGurk trials resulting in illusion (integrated percepts) and comparing them with incongruent trials and McGurk trials not resulting in illusion (not integrated percepts). Surprisingly no areas showed an effect of integration, in any direction ([nMI + MI] > [nMnI+MnI] or [nMnI+MnI] > [nMI + MI]), independently of the type of stimuli.

### 3.2.3. Interaction effect

The last, and critical, analysis step was to test for the interaction to reveal areas that responded differently to integration (integrated vs. no integrated) depending on the nature of the stimuli (McGurk vs. no McGurk).

For this, we used the [MI – MnI] > [nMI – nMnI] and [nMI - nMnI] > [MI - MnI]. To characterize the nature of the interaction in each of the relevant regions and check whether or not the pattern of interaction was consistent with our initial hypothesis we also computed pairwise comparisons between the 4 experimental conditions (see: Table 3).

#### [MI - MnI] > [nMI - nMnI]

First we tested for brain regions showing a larger effect of integration with McGurk ([MI > MnI]) than non-McGurk stimuli ([nMI > nMnI]). Three brain areas showed a significant pattern of interaction in this comparison, namely, the anterior cingulate cortex (ACC), the left inferior frontal gyrus (lIFG) and the superior left precentral cortex (lPC) (see: Figure 1). The ACC, the lIFG and the lPC generally showed higher activity during the conditions in which conflicting audiovisual

information was present. That is, activity on all McGurk trials and in the conflicting non-McGurk trials was higher than activity in the normal non-McGurk non-conflicting matching AV stimuli (with the exception of the lPC in which nMI and MnI were similar). Elaborating on this first set of areas, the ACC response when the McGurk illusion resulted in an illusory percept (MI) was higher, albeit only marginally, than when the illusion was not perceived (MnI) or just not possible (incongruent but non-McGurk stimulus, nMnI). In the lIFG there was no difference in response between MI and nMnI stimuli, but these two conditions showed a higher activation when compared to responses to MnI stimuli. The pattern in the lPC was similar to that in the lIFG with the exception that there was no difference between the nMI and MnI.

Based on previous literature a fair expectation on this interaction contrast, or in the overall effect of integration would have been to see activity in the STS, especially in the left hemisphere. Nonetheless, the pattern was only seen in a non-significant trend (uncorrected $p = 0.015$, size = 23, MNI coordinates of the cluster maxima in mm: -54, -37, 4) in this area.

### [nMI - nMnI] > [MI - MnI]

The reverse comparison ([nMI - nMnI] > [MI - MnI]; larger effects of integration for non-McGurk vs. McGurk stimuli) showed a significant interaction only in the angular gyrus bilaterally. Pairwise comparisons revealed that in this area showed the highest activity when the subject was presented with congruent audiovisual information (i.e. nMI was significantly higher than all the other conditions). Thus, the pattern of activation in the AG was reversed with respect to that in the ACC. In the AG, activity during the McGurk integrated trials was significantly

lower than in any of the conditions, and no difference was found between the conditions in which no integration occur regardless McGurk or non-McGurk stimuli.

# 4. Discussion

The main question addressed in this study was to assess to which extent the integration of AV signals thought to occur during the McGurk illusion (i.e. with incongruent AV input) can be conceptualized as analogous to that taking place under normal (congruent) AV conditions. We hypothesized that, if the McGurk illusion arises from the resolution of a conflict between the auditory and the visual information, then brain areas related to conflict processing, as seen in prior studies, would be strongly involved in its perception. Moreover if the perception of the illusion recruited these areas, we anticipated a differential activity depending on the perceptual outcome of the McGurk stimuli for otherwise physically identical trials.

We used fMRI to identify areas that showed a differential response to AV speech integration depending on whether the triggering stimuli were McGurk or non-McGurk. The main finding to arise from our study is that there exist major differences in the networks underlying the process of integration between conflicting stimuli, as the McGurk events, and naturally corresponding AV congruent stimuli. First, our results indicate that the well known conflict-related brain network comprising the ACC and the left IFG is activated when AV incongruency occurs, regardless of whether this incongruence originated from McGurk or non McGurk stimuli. Second, and equally critical, these two conflict-related areas activate differently depending on the illusory or non-illusory outcome of the McGurk stimuli, that is,

even when comparing trials containing exactly the same physical stimulus. This perception-dependent pattern was also found in the lPC. Thus, while the perception of the McGurk illusion may involve areas in common with normal congruent AV speech (Beauchamp, Nath & Pasalar (2010) and Nath & Beauchamp (2012); see also Section 4.2 below for further discussion on this) our data indicate that the McGurk illusion is additionally mediated by a separate network specifically involved in the detection and resolution of conflict, in this case the AV conflict. Below we discuss the significance of these results and propose a new interpretation of the neural correlates of the McGurk illusion.

## *4.1. Conflict detection and resolution in the McGurk illusion*

Our first question related to the possible role of the brain network related to conflict detection and resolution, in the detection of conflict during AV incongruent trials. This network encompasses the ACC and lIFG. The ACC has been related to conflict detection in a variety of conflict tasks not related to multisensory processing or speech such as the classical paradigms of Stroop, Go/NoGo or stimulus response compatibility (Nee et al., 2007; Roberts & Hall, 2008; Shenhav et al., 2013). Of course, this pattern has also been found in studies using speech when AV congruent and incongruent conditions are compared, in many cases, accompanied by the lIFG, (Benoit et al., 2010; Bernstein, Auer, et al., 2008; Miller & D'Esposito, 2005; Pekkola et al., 2006; Szycik et al., 2009), and discussed specifically in the multisensory context (Noppeney et al. 2008; Weissman, Warner, and Woldorff 2004; Zimmer et al. 2010; Morís Fernández et al. 2015). Our study generalizes the role of the ACC as an area that responds to

conflict in the perception of AV conflict in speech, which is a paradigmatic case encompassing the well known McGurk illusion. The response to conflict in AV speech is reflected by the increased responses of both the ACC and the lIFG to the three AV conflicting conditions used in this study (nMnI, MI and MnI, compared to nMI). In this respect, therefore, the data indicate that the presence of conflict in a McGurk stimulus engages the conflict detection network analogous to any other conflict stimulus.

The second question in this study was related to the specific role of the conflict processing brain network in the perception of the McGurk illusion in particular; and any differential activation with respect to the process of AV integration that unfolds for normally congruent speech events. For this, one must consider the differential pattern of activity found in the ACC for McGurk trials resulting in illusory percepts compared to McGurk trials not resulting in illusory percepts. We found that that ACC activity in the MI and MnI conditions reflects a differential response pattern depending on the final perceptual outcome, illusory or non-illusory, of the AV conflict in McGurk trials. According to previous literature, the role of the ACC is not restricted to the detection of conflict, but it has also been postulated to contribute in the recruitment of additional brain areas that would facilitate the resolution of the conflict at stake (Shenhav, Botvinick, and Cohen 2013). Therefore, based on the proposal by Shenhav and colleagues, a possible interpretation of our current results would be that the joint action of ACC and lIFG mediates the resolution of the AV conflict. The lIFG is an area previously found in studies comparing congruent vs. incongruent AV stimuli (see for example Ojanen et al. 2005; Pekkola et al. 2006; Miller and D'Esposito 2005; Szycik, Jansma, and Münte 2009

for speech stimuli or Noppeney et al. 2008 for incongruent sound and spoken words paired with pictures and written words), and, notably, also reported in some McGurk studies, although its role was not discussed in terms of conflict (Hasson et al., 2007).

Previous studies have suggested that the role of the lIFG, during AV conflict, is related to the mapping of speech inputs into motor representations of the articulatory gestures in Broca's area (Ojanen et al., 2005; Pekkola et al., 2006), consistent with motor-based theories of speech perception (Liberman & Mattingly, 1985; Skipper, Nusbaum, & Small, 2005; Wilson, Saygin, Sereno, & Iacoboni, 2004). Hasson et al. (2007) proposed that the IFG deals with abstract representations of the information and not with direct sensory representations (see also Noppeney et al. (2008), for a similar proposal). A different view was put forward by Miller and D'Esposito (2005), who related the activity in the lIFG with general processes dealing with conflicting or noisy representations. They also speculated about a possible dissociation between automatic AV processing in posterior cortical regions versus frontal regions reflecting more controlled processes.

Here we suggest the role of the lIFG is, at least in part, related to solving AV conflict when present. The lIFG showed low BOLD activity in response to congruent AV conditions (nMI), while activity increased selectively for the conditions including conflicting stimuli. We interpret this pattern as a general increase in processing needs for conflicting information, here comprising conflict between stimuli in different sensory modalities. Going a step further in this interpretation, our data also revealed a differential pattern of activity in the lIFG between the illusory and non-illusory trials from otherwise physically

identical McGurk conditions. This pattern suggests that the McGurk stimulus may lead to reduced processing when the conflict is resolved in favor of the auditory (non-illusion) representation, compared to when it leans toward an integrated (illusory) resolution. If we assume that the role of the ACC here is similar to its role in other types of conflict then a possible interpretation is that the non-illusory percept occurs when the ACC is not able to fully recruit the lIFG and therefore the conflicting inputs cannot be successfully reconciled in a perceptual compromise.

Together with the ACC and IFG, we also found the engagement of the lPC. Activation in the lPC due to McGurk stimuli has also been reported previously (Matchin et al., 2014; Skipper et al., 2007). Specifically a higher BOLD response during McGurk stimuli compared to congruent AV stimuli, an activation profile similar to ours. The pattern found in our study indicates that lPC responds to conflicting conditions[4]. This difference in the pattern may suggest that the lPC helps in processing AV conflicting information. Critically, this activation is differential between trials in which the illusion occurs and those in which doesn't, this reinforces its involvement in the perception of the illusion. The similarity between the lIFG and the lPC patterns of activity suggests that that the lPC may be the end point of a cascade of processes aimed at resolving the conflict, beginning in the ACC, followed by the lIFG and then the lPC. The similarity between the patterns of activity in the non-McGurk non-integrated and McGurk

[4] Although activity in the nMI condition was not significantly different from the MnI condition.

integrated trials may be explained by an attempt to resolve the conflict in both cases, but in the case of the McGurk stimulus the compatibility between the two syllables allows the resolution of the fused percept, whilst in the case of the non-McGurk conflict, the blatant incompatibility does not allow a compromise (resolution in favor of any integrated percept).

Together with the main findings concerning the activation of the ACC and lIFG (plus lPC) for the conflicting McGurk stimuli, our analyses also revealed activation of the angular gyrus (AG), selectively for the integration of non-McGurk stimuli. According to previous studies, the role of the AG has been found to spread across several domains, ranging from semantic processing, reading, number processing, conflict, attention, memory, cross-modal integration or forming part of the default mode network (Seghier 2013). In the context of cross-modal integration it has been postulated to be critical in the process of AV integration by Bernstein et al. (2008). Bernstein's study, using EEG, related the activity of the AG with the presentation of AV stimuli, particularly they found that activity in this area was different depending on the congruency of the stimuli, with incongruent stimuli associated with longer latencies. It may be the case also here that activity in the AG is related to the congruency of the stimuli presented and to the outcome of AV integration. What is more, the illusory outcome of the McGurk stimulus lead to lower activity in this area, compared to when the outcome is a non-illusory percept. Yet this interpretation must remain speculative for now, given its post-hoc nature.

## 4.2. Areas of integration

The main objective of this study was to find the brain areas that would reveal the interplay between audiovisual conflict and integration, and find out how each of these two processes contribute to the neural expression of the McGurk illusion. As noted in the introduction, a fair expectation would have been to find some effect (main effects or interactions) in a classical AV integration area like the superior temporal sulcus (STS), especially in the left hemisphere. Yet, when comparing McGurk integrated and the McGurk non-integrated conditions such pattern was not strongly expressed on our data, above and beyond a trend for larger activity in non-integrated vs. integrated trials in the non-McGurk condition. It is important to note that despite some studies, like Nath & Beauchamp (2012), report a correlation between the degree of activity of the STS and the amount of McGurk illusions perceived by participants, this pattern is not as general as often implied (see, Benoit et al. 2010; Hasson et al. 2007; Matchin, Groulx, and Hickok 2014). Nath and Beauchamp (2012) suggested that these differences may be due to the high anatomical variability of the multisensory locus in the STS, therefore group-wise analyses as the one performed here and in most of the previous studies (e.g. Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam (2010) and Hasson et al. (2007)) may be insensitive to these highly variable anatomical locations in the standard space (see Stevenson et al. 2010 for a similar finding regarding the left STS variability). Future studies may consider investigating the role of conflict in AV integration using some functional localizer to map multisensory STS at the single-subject level.

## 4.3. Summary and conclusions

The results of the current study lead us to conclude that the McGurk perception is, at least partially, mediated by a network different to that involved in the perception of regular AV congruent speech. The ACC and IFG regions highlighted in our study usually underlie the perception of incongruent AV speech and conflict, particularly AV conflict. Several authors have highlighted the potential for visual information to provide predictive information on upcoming acoustic input, given its earlier availability to the perceiver (Arnal, Wyart, & Giraud, 2011; Skipper et al., 2007; van Wassenhove, Grant, & Poeppel, 2005). In the case of incongruent AV input, the initial prediction based on visual information mismatches with the later upcoming auditory input leading to the detection of a conflict and activation of areas involved in conflict processing. The conflict processing and resolution leads to the McGurk illusion in the cases where resolution leads to integration. Therefore our data supports the hypothesis that the conflict detection and resolution between the auditory and visual modalities during a McGurk stimulus processing plays a role in the formation of the illusion. One tentatively picture is that the formation of the McGurk illusion might first involve the detection of the AV conflict and the subsequent allocation of resources to resolve this conflict, a role played by the ACC. The ACC allocates resources through the lIFG and the lPC which will help resolve the conflict either as an illusion or as purely auditory percept. We suggest that the processing of congruent AV speech, although involving integration mechanisms, it does not hinge upon these detection and resolution processes that instead are selective for trials entailing conflicting sensory input.

A related hypothesis has been recently proposed by Romero, Senkowski, and Keil (2015) in an EEG study. Romero and colleagues found two different beta modulations, at early and late timepoints, when comparing the McGurk illusion with congruent stimuli. They also hypothesized an initial stage of conflict detection and a later stage in which this conflict is resolved, each corresponding to the early and late modulations. Our study gives support to this initial idea and provides possible anatomical details about the brain areas involved in detecting and resolving the AV conflict, based on previous, independent conflict literature.

Several authors, including McGurk and MacDonald in their the original paper reporting the illusion, have pointed out the phenomenological observation that the incongruency in the McGurk stimuli often goes unnoticed to the observer (Möttönen, Krause, Tiippana, & Sams, 2002; Summerfield & McGrath, 1984). Yet, others have made the opposite point, that the subjective experience of the McGurk stimulus is different from that of a natural, AV congruent event, even when in both cases integration occurs (see van Wassenhove, Grant, and Poeppel 2007 for an initial hint and Soto-Faraco and Alsius 2009 for discussion of this point). It will be difficult to resolve the phenomenological debate, and it is not our intention to do so here. Yet, we believe this opens up at a possible, admittedly speculative, link between the neural expression of the McGurk illusion and mental phenomenon of its perceptual consequences. Some papers, such as Soto-Faraco and Alsius (2009), report that when participants are informally enquired about the (McGurk) stimuli after an experiment, they usually report a feeling of "oddness", even if they cannot not exactly pinpoint what was wrong. Our results here may offer a neural explanation of this subjective

feeling of "oddness" which may be the detection and resolution of the conflict between the auditory and visual modalities.

In conclusion, the McGurk effect has been regarded as one of the main examples of AV integration. Although this effect undeniably highlights an interaction between the two sensory modalities, several important differences exist between AV speech integration during the McGurk illusion compared to regular AV integration. These differences should be taken into account when using the McGurk illusion to infer more general properties of multisensory integration in general and, specially, of AV speech integration. Here we highlighted that the McGurk illusion is sub-served by the ACC and lIFG that characterize of conflict detection and resolution, and do not engage in AV integration with non-conflicting stimuli.

## 5. Acknowledgments

This research was supported by the Ministerio de Economía y Competitividad (PSI2013-42626-P), AGAUR Generalitat de Catalunya (2014SGR856 and 2012BE100392), and the European Research Council (StG-2010 263145). The Neuroimaging Laboratory of the Fondazione Santa Lucia is supported by the Italian Ministry of Health.

## 6. References

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

Andersen, T. 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., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. doi:10.1038/nn.2810

Basu Mallick, D., F Magnotti, J., & S Beauchamp, M. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0817-4

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(7), 2414–7. doi:10.1523/JNEUROSCI.4865-09.2010

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping*, *31*(4), 526–38. doi:10.1002/hbm.20884

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing.

*NeuroImage*, *39*(1), 423–35. doi:10.1016/j.neuroimage.2007.08.035

Bernstein, L. E., Lu, Z.-L., & Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, *1242*, 172–84. doi:10.1016/j.brainres.2008.04.018

Bhaganagarapu, K., Jackson, G. D., & Abbott, D. F. (2013). An automated method for identifying artifact in independent component analysis of resting-state FMRI. *Frontiers in Human Neuroscience*, *7*(July), 343. doi:10.3389/fnhum.2013.00343

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1001–10. doi:10.1098/rstb.2007.2155

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk – MacDonald effect : a phonetic representation within short-term memory, *113*, 495–506.

Dekle, D. J., Fowler, C. a., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics*, *51*(4), 355–362. doi:10.3758/BF03211629

Freeman, E. D., Ipser, A., Palmbaha, A., Paunoiu, D., Brown, P., Lambert, C., … Driver, J. (2013). Sight and sound out of synch: Fragmentation and renormalisation of audiovisual integration and

subjective timing. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 1–13. doi:10.1016/j.cortex.2013.03.006

Friston, K. J., Holmes, a, Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, *4*(3 Pt 1), 223–35. doi:10.1006/nimg.1996.0074

Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron*, *56*(6), 1116–26. doi:10.1016/j.neuron.2007.09.037

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–90. doi:10.1016/j.neuroimage.2011.09.015

Keil, J., Müller, N., Ihssen, N., & Weisz, N. (2012). On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(1), 221–31. doi:10.1093/cercor/bhr125

Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience*, *20*(12), 2175–84. doi:10.1162/jocn.2008.20152

L.E. Bernstein, E.T. Auer Jr., J. K. M. (2004). Audiovisual speech binding: Convergence or association. In B. E. S. G.A. Calvert, C.

Spence (Ed.), *Handbook of Multisensory Processes* (pp. 203–224). Cambridge MA: MIT press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. doi:10.1016/0010-0277(85)90021-6

Malfait, N., Fonlupt, P., Centelles, L., Nazarian, B., Brown, L. E., & Caclin, A. (2014). Different neural networks are involved in audiovisual speech perception depending on the context. *Journal of Cognitive Neuroscience*, *26*(7), 1572–86. doi:10.1162/jocn_a_00565

Massaro, D. (1987). *Speech Perception By Ear and Eye* (p. -1). Routledge. doi:doi:10.4324/9781315799742

Massaro, D., & Stork, D. (1998). Speech Recognition and Sensory Integration. *American Scientist*, *86*(3), 236. doi:10.1511/1998.3.236

Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience*, *26*(3), 606–20. doi:10.1162/jocn_a_00515

Matthew Brett, Jean-Luc Anton, Romain Valabregue, J.-B. P. (2002). Region of interest analysis using an SPM toolbox. In *8th International Conference on Functional Mapping of the Human*

*Brain.* Sendai, Japan. Retrieved from http://marsbar.sourceforge.net/

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1012311

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(25), 5884–93. doi:10.1523/JNEUROSCI.0896-05.2005

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S. (2015). Top-down attention regulates the neural expression of audiovisual integration. *NeuroImage*, *119*, 272–285. doi:10.1016/j.neuroimage.2015.06.052

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Research. Cognitive Brain Research*, *13*(3), 417–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11919005

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, *59*(1), 781–7. doi:10.1016/j.neuroimage.2011.07.024

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cognitive,*

*Affective & Behavioral Neuroscience*, *7*(1), 1–17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17598730

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex (New York, N.Y. : 1991)*, *18*(3), 598–609. doi:10.1093/cercor/bhm091

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*(2), 333–8. doi:10.1016/j.neuroimage.2004.12.001

Orr, J. M., & Weissman, D. H. (2009). Anterior cingulate cortex makes 2 contributions to minimizing distraction. *Cerebral Cortex (New York, N.Y. : 1991)*, *19*(3), 703–11. doi:10.1093/cercor/bhn119

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *NeuroImage*, *29*(3), 797–807. doi:10.1016/j.neuroimage.2005.09.069

Roa Romero, Y., Senkowski, D., & Keil, J. (2015). Early and Late Beta Band Power reflects Audiovisual Perception in the McGurk Illusion. *Journal of Neurophysiology*, jn.00783.2014. doi:10.1152/jn.00783.2014

Roberts, K. L., & Hall, D. a. (2008). Examining a supramodal network for conflict processing: a systematic review and novel functional

magnetic resonance imaging data for related visual and auditory stroop tasks. *Journal of Cognitive Neuroscience*, *20*(6), 1063–78. doi:10.1162/jocn.2008.20074

Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *19*(1), 43–61. doi:10.1177/1073858412440596

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–40. doi:10.1016/j.neuron.2013.07.007

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *NeuroImage*, *25*(1), 76–89. doi:10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(10), 2387–99. doi:10.1093/cercor/bhl147

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(2), 580–7. doi:10.1037/a0013483

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the

speeded classification task. *Cognition*, *92*(3), B13–23. doi:10.1016/j.cognition.2003.10.005

Stevenson, R. a, Altieri, N. a, Kim, S., Pisoni, D. B., & James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage*, *49*(4), 3308–18. doi:10.1016/j.neuroimage.2009.12.001

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology Section A*, *36*(1), 51–74. doi:10.1080/14640748408401503

Szycik, G. R., Jansma, H., & Münte, T. F. (2009). Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Human Brain Mapping*, *30*(7), 1990–9. doi:10.1002/hbm.20640

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*(4), 725. doi:10.3389/fpsyg.2014.00725

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457–472. doi:10.1080/09541440340000268

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–6. doi:10.1073/pnas.0408949102

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Weissman, D. H., Warner, L. M., & Woldorff, M. G. (2004). The neural mechanisms for minimizing cross-modal distraction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *24*(48), 10941–9. doi:10.1523/JNEUROSCI.3669-04.2004

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–2. doi:10.1038/nn1263

Zimmer, U., Roberts, K. C., Harshbarger, T. B., & Woldorff, M. G. (2010). Multisensory conflict modulates the spread of visual attention across a multisensory object. *NeuroImage*, *52*(2), 606–16. doi:10.1016/j.neuroimage.2010.04.245

## 2.3 The role of conflict during the perception of the McGurk illusion

Morís Fernández, L., Torralba M. & Soto-Faraco, S.

The role of conflict during the perception of the McGurk illusion

# The role of conflict during the perception of the McGurk illusion

Luis Morís Fernández [a]

Mireia Torralba Fernández [a]

Salvador Soto-Faraco [a, b]

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

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

Corresponding author: Luis Morís Fernández

Dept. de Tecnologies de la Informació i les Comunicacions
Universitat Pompeu Fabra, Office 55.128.
Roc Boronat, 138
08018 Barcelona
Spain

+34 686 17 30 58

luis.moris.fernandez@gmail.com

## Abstract

The McGurk effect has been used very often in the literature to infer properties of AV speech integration in general. Nonetheless, one obvious distinctive feature in the McGurk case is the mismatch between the auditory and visual modalities. In this study we propose that, because of this reason, the processing of a McGurk stimulus is essentially different from that of regular (congruent) AV speech and generalizations to natural speech perception are complicated. We propose that the cross-modal conflict present in the McGurk effect engages conflict mechanisms in the brain, akin those at work in other classic conflict paradigms (e.g., Stroop). To test this hypothesis we used a well-known conflict marker in EEG, a theta power increase in midfrontocentral electrodes, found previously in conflict tasks. We found the hypothesized increase in the theta band in the midcentral electrodes in the expected band when a McGurk stimulus was presented as compared to when congruent AV speech is presented. We conclude that the McGurk effect is processed differently from congruent AV speech, and that the McGurk effect is mediated by conflict processes.

## Keywords

McGurk; EEG; Conflict; Speech perception; Multisensory Integration; Audiovisual

# 1. Introduction

Forty years ago, in 1976, Harry McGurk and John McDonald discovered that by dubbing an auditory syllable (e.g., /ba/) with a different visual syllable (e.g., [ga]), the resulting auditory percept could be dramatically altered into a completely different syllable (e.g., /da/[1]; see Massaro & Stork, 1998 for a description of how this discovery was made). This effect pushed the boundaries of previous works by demonstrating that the influence of visual information on auditory speech perception went beyond supplementing the acoustic signal under noise (e.g., Sumby & Pollack, 1954).

This effect has been widely used in the literature due to its intertrial variability. While in some trials the observer's perception corresponds to an illusory sound (e.g., /da/) and it is usually interpreted as resulting from successful integration of sensory modalities, in other trials the observer's perception corresponds to the actual acoustic stimulus (e.g., /ba/), in which case the interpretation is of a failure in cross-modal integration. This allows researchers to separate integrated trials from non-integrated ones, or measure on average how often the stimuli are perceptually integrated, while the physical stimulation remains constant. Due to this easy and accessible measure, the McGurk effect in various forms has been used in countless studies to infer properties of multisensory integration in general and for AV speech integration in particular (e.g., Alsius et al. 2005; van Wassenhove, Grant, and Poeppel 2007; Skipper et al. 2007; Tiippana, Andersen, and Sams 2004; Bernstein et al. 2008; Andersen et al. 2009).

[1] Throughout this manuscript the visual part of a syllable will be written between brackets (i.e. [ba]) while the auditory part will be written between slashes (i.e. /ba/).

However one fundamental difference between regular AV integration (i.e. day by the day experience of the world) and the McGurk illusions that, while in the former case the AV stimulation is congruent in the latter there is a conflict between the auditory and the visual information. That is, the shape and movements of the lips do not correspond to the sound, as they normally do in the observer's experience. This rather obvious fact implies that properties derived from the study of the McGurk effect may not be fully generalizable to how multisensory integration happens in natural circumstances and, when generalized this must be done with care. Surprisingly, this relatively obvious argument has been rarely considered in the literature. The reason probably lies under the assumption that the McGurk effects, like other cross-modal integration phenomena are rather automatic and unavoidable, and hence the observer is rarely aware of the fact that there is a conflict at all. Several studies in the last few years, however, have questioned this strong version of automaticity in cross-modal integration (Alsius et al., 2005; Alsius, Navarra, & Soto-Faraco, 2007; Andersen et al., 2009; Nahorna, Berthommier, & Schwartz, 2012).

The claim in the present study is that while, undeniably, AV integration (i.e. an interaction between auditory and visual information) takes place in AV congruency and in the McGurk illusion, the process whereby this integration occurs may be very different. We suggest that AV integration in the case of the McGurk effect is a consequence of the detection and resolution of the conflict between the auditory and visual information. More specifically we hypothesize that this conflict arises due to the mismatch between the prediction based on the visual, speech reading, information and the upcoming auditory input (Arnal, Wyart, & Giraud, 2011; Morís Fernández, Visser, Ventura-Campos, Ávila, &

Soto-Faraco, 2015; Skipper et al., 2007; van Wassenhove, Grant, & Poeppel, 2005). In this respect, the McGurk effect might not be different from other cases of perceptual conflict. Following on this hypothesis, we predicted that the McGurk effect will display EEG correlates that are similar in power and scalp distribution as those seen for other forms of conflict.

The effect of conflict have been studied using different paradigms, such as the Stroop task (Stroop, 1935), Eriksen Flanker task (Eriksen & Eriksen, 1974) or the Cued Go/No-Go task (Fillmore & Weafer, 2004). Two big neural correlates of conflict in these classical tasks have been found. First, fMRI studies have revealed that conflict activates frontal areas in particular the anterior cingulate cortex (ACC) (Nee, Wager, & Jonides, 2007; Shenhav, Botvinick, & Cohen, 2013) and, second, EEG recordings revealed an increase in theta power (4-8 Hz) over the midfronto-central electrodes in the conflicting condition when compared with the non-conflicting condition (Cavanagh & Frank, 2014; Cohen, 2014; Ergen et al., 2014; Hanslmayr et al., 2008). Activity in the ACC was also found when we move away from classical conflict tasks and into AV conflict (i.e. incongruent AV stimulation)(Noppeney, Josephs, Hocking, Price, & Friston, 2008; Orr & Weissman, 2009; Weissman, Warner, & Woldorff, 2004; Zimmer, Roberts, Harshbarger, & Woldorff, 2010). This ACC is found as well in the domain of speech when comparing AV congruent with AV incongruent stimuli (Miller & D'Esposito, 2005; Morís Fernández et al., 2015; Ojanen et al., 2005; Pekkola et al., 2006; Szycik, Jansma, & Münte, 2009). Miller & D'Esposito, (2005) also speculated about a possible dissociation between automatic AV speech processing in posterior cortical regions versus a more controlled processing in frontal

regions in the case of incongruent AV speech information. Remarkably, ACC activity has also been shown in many studies involving the McGurk effect (Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam, 2010; Bernstein, Lu, & Jiang, 2008), albeit up to our best knowledge, the role of this brain area has never been interpreted in the McGurk context.

In the present study we asses if a McGurk stimulus is perceived and treated as a conflict. For this we will we measure theta power (4-8 Hz) over the midfronto-central electrodes as a probe for the presence of conflict mechanisms (Cavanagh & Frank, 2014; Cohen, 2014; Ergen et al., 2014; Hanslmayr et al., 2008). We will use a paradigm in which participants will be presented with three different kinds of stimuli: Congruent, Incongruent and McGurk stimuli. To reinforce our point participants EEG will also be measured in a classical Stroop task for comparison with the effects found during the McGurk effect. During the rest of the paper we will refer to the first task as Speech task and the second as Stroop task.

The main result we anticipate is that if the McGurk effect is perceived as a conflict we should see an increase in non-phase locked theta power in the central electrodes when compared with a Congruent stimulus, we expect to see a similar effect following the same logic when comparing Congruent with Incongruent stimuli. We further expect that the topographical distribution and spectral peak within the frequency range of interest will be comparable to that produced when comparing congruent vs. incongruent of the Stoop stimuli. Given the very different time course of information integration, the rise of conflict and its resolution between McGurk and Stroop, we do not make claims about any correlation between terms of latencies. Yet, in the particular case of

the McGurk stimuli, we expect the correlates of conflict to appear closely after the onset of the auditory stimulus, as this is the moment of the conflict between the visual prediction and the auditory part of the stimuli.

# 2. Methods

## *2.1. Stimuli*

The McGurk videos used in this study were borrowed from Basu Mallick, Magnotti and Beauchamp (2015), labeled as 2.5[2] in that study (we refer the readers to that reference for particular details on how they were recorded). Three different videos were used to build our materials, two audiovisually congruent (AVc condition: [ba] + /ba/ and [da] + /da/), and one with the McGurk combination (MC condition: [ga] + /ba/). The videos, originally two seconds long, were extended to five seconds duration following this process: first we found the frame preceding the first lip movement, and the frame following the last lip movement after the speaker closed the mouth. Then we replicated these two frames, at the beginning and the end of the video respectively, so that the auditory onset occurred ~2.5 seconds after the beginning of the video for all videos. These manipulations gave our Congruent and McGurk conditions. To create the Incongruent condition, we just reversed the video of the Congruent video clips. We also created Auditory Only stimuli by substituting the video with a white fixation cross on black background.

[2] We selected this particular set of stimuli from the eight available from Basu Mallick's study after running an informal pilot. This set was the one that showed the most unambiguous auditory stimulation and seemed to induce the McGurk effect more often.

## 2.2. Participants

The data of this study is based on responses from 17 participants, pre-selected from a larger inclusion behavioral study (n=64), so that we used participants who perceived the McGurk illusion with this particular set of stimuli[3]. From the initial set of 24 participants that consistently perceived the McGurk illusion and were included in the EEG experiment, 7 were discarded (5 of them did not show the McGurk illusion during the EEG experiment, and 2 were discarded due to excessive movements, artifacts, muscular and ocular). One participant was excluded from the Stroop analysis as his accuracy was very low during the Incongruent condition in the Stroop task (less than 1%). All participant selection was done based on criteria independent of the main analysis and was done before this analysis took place.

## 2.3. Procedure

The presentation protocol was programmed using E prime 2.0.10.242. EEG data was preprocessed and analyzed using fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 2011).

### 2.3.1. Task

#### 2.3.1.1. Speech

Participants performed the same task in both the inclusion study and the EEG study, albeit a single block was run during the pilot. In a given trial they were sequentially presented with a fixation cross (1s), an AV stimulus (5s) and a prompt screen until response. When the response screen was presented they engaged a three alternative forced choice

[3] This percentage of participants, as well as the behavior explained in the next section, for a given set of stimuli is very similar to that found in one of the few massive McGurk studies made (see Basu Mallick et al., 2015). We decided to select participants to maximize the effectiveness during the EEG experiment given this high variability.

task in which they had to identify which syllable they have perceived, BA, DA or GA. The three options appeared randomized in three different screen positions left, center and right; this way no motor anticipation could occur during the EEG experiment. Participants were presented with 100 Congruent , 100 Incongruent trials and 200 McGurk trials divided in five blocks where stimuli were proportional to the main set and randomized. Afterwards participants performed the Auditory Only (AO) block with 30 /ba/ trials and 30/ga/ trials with the same protocol to asses that auditory only stimuli could be identified properly in the absence of visual information and therefore control that the McGurk effect was produced by the influence of the visual information and not because of ambiguous auditory stimuli.

#### 2.3.1.2. Stroop

For this task we used the three Spanish words ROJO (red), AZUL (blue) and VERDE (green). We did not try to equate low level stimulus features with the McGurk stimuli simply because the two tasks are very different and have different requirements (e.g., temporal course of information presentation). Instead, we used a prototypical Stroop paradigm (Hanslmayr et al., 2008). In the Congruent condition participants were presented with a color word printed in the corresponding ink color while in the Incongruent condition they were presented with a color word printed in one of the two other colors (balanced and equiprobable across the incongruent stimuli). In a given trial participants were presented with a fixation cross (1s), the word stimuli (1s) and a black screen (1s). Participants were asked to emit a speeded response to the color of the ink while the word was in the screen and as soon as possible while trying to keep their accuracy as high as possible. Before starting they performed a training phase to

learn the association between three keys and the three colors as in this case the mapping between the color and the button was constant (counterbalanced across participants). Participants were presented with 100 Congruent trials and 100 Incongruent trials divided in two blocks.

### 2.3.2. EEG

Electrophysiological data was recorded at a rate of 500 Hz from 59 active electrodes (impedance was kept below 10 kΩ) placed according to the 10–20 convention (Fp1, Fp2, AF7, AF3, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT9, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, FT10, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO3, POz, PO4, O1, Oz, O2) (ActiCap, Brain Vision Recorder, Brain Products). Four extra electrodes were located on the left/right mastoids, and below and to the outer canthus of the right eye. An additional electrode placed at the participant's tip of the nose was used as reference during recording. Ground electrode was located at the AFz location.

### 2.3.3. EEG preprocessing

Data was re-referenced offline to the average of the mastoids. Three different filters were applied: a notch filter at 50 Hz, a 0.5Hz high-pass second order Butterworth filter, and a 50Hz low-pass eighth order Butterworth filter. The data set was segmented into 4 seconds epochs (from 2 seconds before the auditory onset to 2 seconds after the auditory onset). All epochs with amplitude peaks exceeding ±150μV were marked as artifacts. Epochs were visually screened for visual (blinks and eye movement) artifacts.

## *2.4. Analysis*

### 2.4.1. Behavior

Proportion correct responses were calculated for the congruent and incongruent conditions, for the McGurk condition the proportion of trials in which the illusion was perceived (/ga/ or /da/) was also calculated. Accuracy and reaction times were calculated for both Stroop conditions, only correct trials were included in the reaction times analysis.

### 2.4.2. Selection of trials for EEG analysis

#### 2.4.2.1. Speech

For the congruent and incongruent condition only those trials in which the response was correct were included in the analysis. For the McGurk trials only those trials in which the McGurk illusion was perceived were included in the analysis.

#### 2.4.2.2. Stroop

Only correct Stroop trials were included in the analysis.

### 2.4.3. Power analysis

#### 2.4.3.1. Speech

Given our a priori hypothesis we focused our analysis on the Cz electrode, on the theta band (5 to 7 Hz), on the post stimulus period. We first estimated the power of oscillatory activity for each participant and condition and frequency band of interest (5 to 7 Hz, 1 Hz steps), using a short Fourier transform and a Hanning taper (500 ms duration). The power estimate was calculated from -250 to 750 ms at 20 ms steps with respect to the auditory onset. These data was then baseline corrected by calculating the relative change with respect to a pre-stimulus baseline (-1.5 s to -0.5s with respect to the auditory onset) in dB. Data was then averaged through frequencies to obtain a single time series per participant and condition.

In a second level stage we ran a paired t-test across all time points, comparing: McGurk vs. Congruent conditions; and, Incongruent vs. Congruent conditions. The critical contrast of interest was the one comparing the Congruent and the (illusory) McGurk condition. In addition we also calculated the difference between the Congruent and Incongruent conditions as we also expected to find a similar effect as an AV conflict was also present in the incongruent condition. Correction for multiple comparisons for the time series was performed using the table included in the work by Guthrie & Buchwald (1991), with the following parameters: length of interval (T)=50, graphical threshold(Θ)=0.05, autocorrelation parameter (φ)=0.9 (estimated from our dataset) and number of subjects (N) = 15 as it was the closest to our dataset[4], therefore, the length of the sequence needed to achieve a level of significance of 0.05 was 9 consecutive data points. The time of interest was selected based on the duration of the auditory stimuli ~500 ms and the length of the Hanning tapper ~500 ms. Note that, in these conditions, the first point in which the post-stimulus auditory activity may have any influence is 250 ms before the auditory onset.

#### 2.4.3.2. Stroop

The same procedure was applied to the Stroop trials, with the difference that the baseline in this case was calculated from -1 s to 0 s with respect to the presentation of the stimuli. Also in this case the time of interest was from 0s to 1s with respect to the presentation of the stimuli.

It is worth noting that all the analyses described above, both McGurk and Stroop analyses, were decided prior to data collection according to

[4] The same number of consecutive points were needed in case of rounding up to N=20.

the hypothesis, and we did not ran other tests besides the ones specified above expect for one exception: After the data collection, we decided to repeat our analysis procedure for a wider range of frequencies (2 Hz to 50 Hz in 1 Hz steps) and for all electrodes, in order to produce a topology figure and a frequency map for both the McGurk and the Stroop tasks.

# 3. Results

## *3.1. Behavior*

### 3.1.1. Speech

As seen in Table 1A and 1B, the McGurk illusion occurred to a high degree with our stimuli and (pre-selected) participant group, and the auditory alone versions of the stimuli were clearly identifiable. This pattern ensures that the McGurk effect was genuinely due to visual influence and its interaction with the auditory signal, not due to decision effects over ambiguous auditory stimuli. Therefore, any possible neural correlates of conflict seen in the subsequent EEG analyses cannot be explained by extra cognitive effort due to an ambiguous auditory stimulus.

### 3.1.2. Stroop

As expected, in the Stroop task participants' performance was poorer in the Incongruent condition than in the Congruent condition, accompanied by an increase in reaction time, indicating that the Stroop effect was present (Table 1C).

### A. MCGURK

| Congruent | Incongruent | McGurk | |
|---|---|---|---|
| **Correct** | **Correct** | **Illusory(DA/GA)** | **Non-Illusory(BA)** |
| 0.99 (±0.003) | 0.86 (±0.04) | 0.86 (±0.043) | 0.14 (±0.043) |

### B. AUDITORY ONLY

| BA | GA |
|---|---|
| 0.94 (±0.031) | 0.99 (±0.006) |

### C. STROOP

| | Congruent | Incongruent |
|---|---|---|
| RT | 637 ms (±11) | 705 ms (±13) |
| ACC | 0.96 (±0.011) | 0.86 (±0.021) |

**Table 1 A.** Behavioral data corresponding to the experimental task performed during the EEG recordings. Proportion of trials in which the syllable was correctly identified (e.g. heard /ba/ responded /ba/) is presented for the Congruent and Incongruent conditions. For the McGurk condition the proportion of responses for illusory and non-illusory percepts is presented. All participants at individual level perceived DA consistently as the dominant illusory percept. DA percepts accounted on average for 0.98 of all illusory trials while GA only accounted for the other 0.02. **B.** Behavioral data corresponding to the auditory only task performed after the EEG experiment. Proportion of trials in which the syllable was correctly identified (e.g. heard /ba/ responded /ba/) is presented for the two possible auditory syllables that appeared in the EEG experiment. **C.** Reaction time and accuracy are presented for each of the conditions during the Stroop task. Congruent and incongruent trials were significantly different on both measures ($p<10^{-4}$, paired t-test). Standard error of the mean is presented in parenthesis in all tables.

## 3.2. EEG Power Analysis

### 3.2.1. McGurk

By hypothesis, our analysis focused on the theta band (5-7 Hz). We observed an increase in the post stimulus power between - 60 ms to 100 ms with respect to the auditory onset when the McGurk illusory condition was compared with the congruent condition. A power increase over theta was found when we compared the Incongruent with the Congruent condition although it was slightly later and more prolonged in time (50 ms to 300 ms).

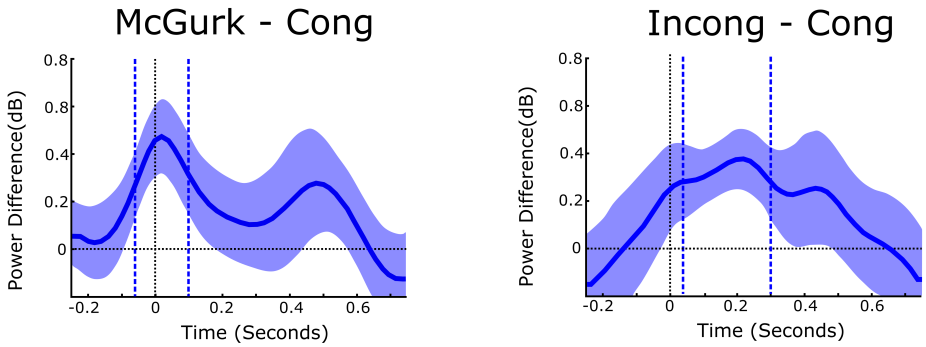



**Figure 1** Theta power (5-7 Hz) evolution in Cz electrode. Lines represent the evolution of the difference in power relative change in dB between the McGurk illusory and Congruent conditions (left panel) and the Incongruent and Congruent conditions (right panel), shaded areas around the lines represent the standard error of the mean. Bold dashed lines indicate the significant period after multiple comparisons corrections.

### 3.2.2. Stroop

When comparing theta power of the Incongruent with the Congruent Stroop conditions we found a similar modulation in power at Cz, though in this case its time window was from approximately 450 ms to 750 ms, a result that corresponds very well with those found previously in Stroop literature (see for example Ergen et al., 2014; Hanslmayr et al., 2008).

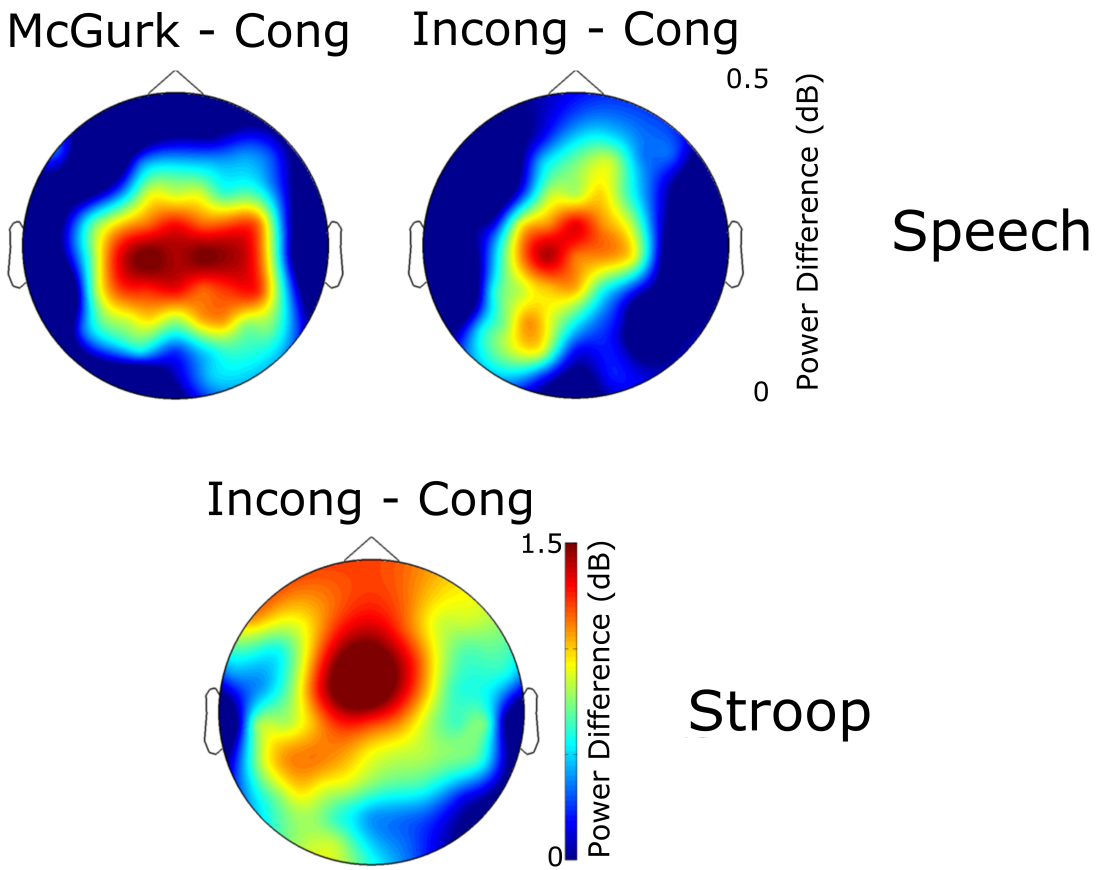



**Figure 2** Topographies displaying theta (5-7 Hz) power across the scalp at the peak times for each task and comparison of interest. In the top row the comparisons in the speech task between the McGurk and Congruent condition 20 ms after the auditory onset and between the Incongruent and Congruent condition 200 ms after the auditory onset. In the bottom row the difference between Incongruent and Congruent condition in the Stroop task 650 ms after the stimuli onset is presented. In all cases difference in power in dB with respect to the baseline is presented. Please notice the difference in scale between the Speech and Stroop topographies.

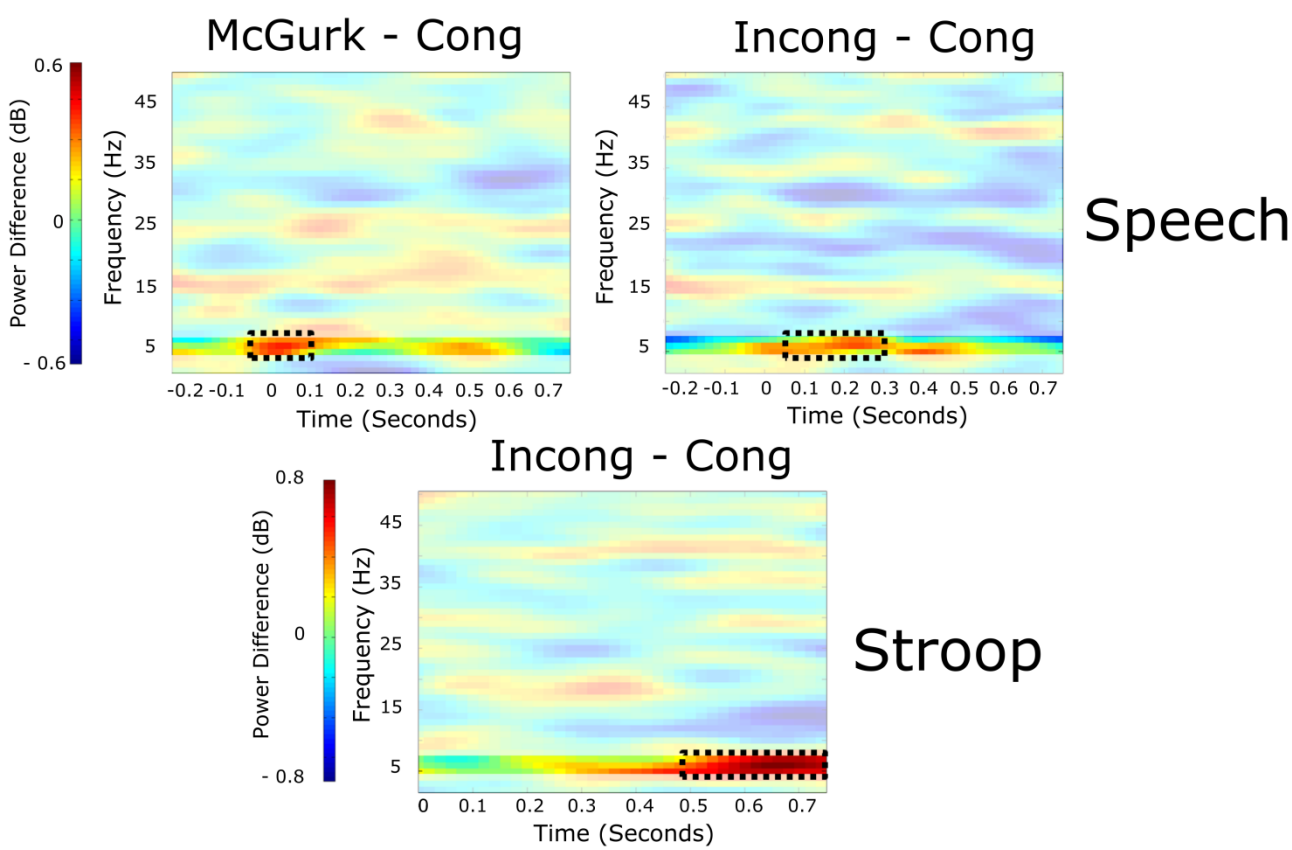



**Figure 3** Power distribution across frequencies (2 – 50 Hz) and time in the Cz electrode for each task and comparison of interest. In the top row the comparisons in the speech task between the McGurk and Congruent condition and between the Incongruent and Congruent from -250 ms to 750 ms after the auditory onset. In the bottom row the difference between Incongruent and Congruent condition in the Stroop from 0 ms to 750 ms after the stimulus onset. In all cases difference in power in dB with respect to the baseline is presented, frequencies of interest are highlighted; significant periods are marked with a dotted square box. Please note the difference in time and power scale between the Speech and Stroop graphs.

# 4. Discussion

The main question addressed during this study was if the McGurk effect is processed as a conflict by the brain. If this were true, we hypothesized, then we should find a post-stimulus increment in Theta power, as this increment is one of the most common markers of conflict in EEG for classical conflict tasks (Cavanagh & Frank, 2014; Cohen, 2014; Ergen et al., 2014; Hanslmayr et al., 2008). To reinforce our result we ran a Stroop task while recording EEG and compared the results of the Stroop task in the same group of participants.

## *4.1. Behavior*

Our behavioral results in the speech task align well with previous studies such as for example those by Basu Mallick and colleagues (2015). As in their results we found that in the selection procedure there was an almost binary distribution between participants who perceived the illusion almost always, those that were selected, and those who never perceived the illusion, those that were not selected. This distribution justifies the need of a selection process in the McGurk studies to ensure that the illusion is effective in all participants.

The behavioral data obtained during the Stroop task are in line with those obtained in previous studies, see for example (Ergen et al., 2014; Hanslmayr et al., 2008; MacLeod, 1991).

## *4.2. EEG*

Regarding the Stroop effect the results are well in line with those found in previous studies (Ergen et al., 2014; Hanslmayr et al., 2008), indicating mid-frontocentral activity when comparing the Incongruent with the Congruent condition. Particularly we found the expected

modulation in the Cz electrode in the Theta power band in consonance with the findings of previous studies.

Critical to our initial hypothesis, we obtained a similar result in the audiovisual speech EEG, when comparing the McGurk condition with the Congruent condition. An obvious difference between the speech and Stroop incongruency results is the timing profile in the Theta differences, which appears much earlier in the case of speech and much later in the Stroop. This was expected due to the dramatic difference in the nature of the stimuli, in the time needed to process each of the two sources of information, and the point at which the conflict can be detected in each case. In the case of the Stroop task, the latency results are well in line with prior observations. The early effects seen in the case of the speech conflict can be attributed to the fact that, at the moment of the auditory onset, the visual part of the stimulus has been already partially processed and had created a strong prediction that, when violated by the upcoming sound, would trigger the conflict processing network. This interpretation is in line with previous works that have highlighted the role played by visual speech to provide predictive information on upcoming acoustic input, given its earlier availability to the perceiver (Arnal et al., 2011; Morís Fernández et al., 2015; Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011; Sánchez-García, 2013; Skipper et al., 2007; van Wassenhove et al., 2005). Therefore, as this prediction is already prepared is not surprising to see that the peak occurs soon after the auditory stimulus is presented, nonetheless, we are cautious with respect to the timing of the peak, as the window used for the analysis is very wide (500 ms).

In line with the hypothesis that the McGurk effect engages the conflict detection network, the comparison between Incongruent vs. Congruent speech, without McGurk effect, reveals a pattern very similar in terms of spectral peak and topography to the comparison between McGurk vs. Congruent conditions in the speech task. The effects in these two contrasts do also overlap in time, albeit the effect in the Incongruent-Congruent comparison is delayed and prolonged in time. This latency shift may be due to the nature of the stimuli, given that an incongruency produced by a reversed video could be larger than in the subtle case of the McGurk effect, or due to the impossibility of reconciling the two modalities.

An interesting finding is that this Theta band activity peaked at 6 Hz for all comparisons, irrespective of the task (Speech or Stroop), indicating some overlap in the processing mechanisms engaged. In what refers to the topographical distribution during the peak activities, theta increases in all kinds of incongruence seem to be rather focused on the central electrodes.

This set of results supports our initial hypotheses. First, that the AV incongruency is perceived as a conflict and it engages the processes previously found in classical conflict tasks. This is in line with the hypothesis presented in another study by this same group (Morís Fernández et al., 2015), in which we highlight that AV incongruency in speech may engage classical areas related to conflict perception such as the anterior cingulate cortex. Second and critically, we not only found that conflict processes are engaged by incongruent AV speech but also that the McGurk effect, a particular case of AV incongruency, also engages these conflict processes as signaled by EEG. This result

establishes a clear difference between the processing of a regular AV congruent stimulus and that of a McGurk stimulus. This implies that care must be taken when generalizing results obtained from the McGurk effect to AV speech integration in general. Moreover, we go further and suggest that the possible origin of the McGurk illusion may in fact lie on the resolution of the AV conflict.

The McGurk effect has been regarded as automatic, fast and pre-attentive (Colin et al., 2002; McGurk & MacDonald, 1976; Rosenblum & Saldaña, 1996; Soto-Faraco, Navarra, & Alsius, 2004; Summerfield & McGrath, 1984). Nonetheless recent evidence have challenged this point of view and started highlighting the role of the inner state of the participant, specially attention, in the perception of the McGurk illusion (Alsius et al., 2005, 2007; Andersen et al., 2009; Nahorna et al., 2012). In this case we do not challenge the general view of multisensory integration, particularly AV speech integration and automatic process when we are confronted with congruent AV stimulation. In this case, our point is that in the face of the AV conflict and the impossibility of integrating these two sources of information general conflict mechanisms are invoked. These general conflict mechanisms are therefore only activated in the case of AV conflict may not participate in the regular congruent AV integration that occurs as suggested by previous studies in a fast and automatic manner (see Morís Fernández et al., 2015 for a similar hypothesis). Roa Romero, Senkowski, & Keil, (2015) in light of their results has hypothesized a three stage process in the perception of the McGurk illusion: early AV integration, detection of the conflict and allocation of extra resources to resolve the conflict, and finally resolution of the conflict and creation of the illusory percept. While in their case they base their hypothesis mainly on two

changes, early and late, in the beta band, we add to this previous literature by relating the perception of the McGurk effect with mechanisms previously related to conflict detection and resolution.

Summing up, our data indicates that Incongruent AV speech is perceived as a conflict. More importantly the McGurk effect is also perceived as a conflict and we suggest that the McGurk illusion is the outcome of the resolution of the conflict between the visual and auditory parts of the AV stimuli.

## 5. Acknowledgments

This research was supported by the Ministerio de Economía y Competitividad (PSI2013-42626-P), AGAUR Generalitat de Catalunya (2014SGR856 and 2012BE100392), and the European Research Council (StG-2010 263145).

## 6. References

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

Andersen, T. 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., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. doi:10.1038/nn.2810

Basu Mallick, D., F Magnotti, J., & S Beauchamp, M. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0817-4

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping*, *31*(4), 526–38. doi:10.1002/hbm.20884

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage*, *39*(1), 423–35. doi:10.1016/j.neuroimage.2007.08.035

Bernstein, L. E., Lu, Z.-L., & Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, *1242*, 172–84. doi:10.1016/j.brainres.2008.04.018

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. doi:10.1016/j.tics.2014.04.012

Cohen, M. X. (2014). A neural microcircuit for cognitive conflict detection and signaling. *Trends in Neurosciences*, *37*(9), 480–490. doi:10.1016/j.tins.2014.06.004

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk – MacDonald effect : a phonetic representation within short-term memory, *113*, 495–506.

Ergen, M., Saban, S., Kirmizi-Alsan, E., Uslu, A., Keskin-Ergen, Y., & Demiralp, T. (2014). Time-frequency analysis of the event-related potentials associated with the Stroop test. *International Journal of Psychophysiology : Official Journal of the International*

*Organization of Psychophysiology*, *94*(3), 463–72. doi:10.1016/j.ijpsycho.2014.08.177

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. doi:10.3758/BF03203267

Fillmore, M. T., & Weafer, J. (2004). Alcohol impairment of behavior in men and women. *Addiction (Abingdon, England)*, *99*(10), 1237–46. doi:10.1111/j.1360-0443.2004.00805.x

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*(2), 240–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1946890

Hanslmayr, S., Pastötter, B., Bäuml, K.-H., Gruber, S., Wimber, M., & Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *Journal of Cognitive Neuroscience*, *20*(2), 215–25. doi:10.1162/jocn.2008.20020

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163–203. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2034749

Massaro, D., & Stork, D. (1998). Speech Recognition and Sensory Integration. *American Scientist*, *86*(3), 236. doi:10.1511/1998.3.236

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1012311

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(25), 5884–93. doi:10.1523/JNEUROSCI.0896-05.2005

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S. (2015). Top-down attention regulates the neural

expression of audiovisual integration. *NeuroImage*, *119*, 272–285. doi:10.1016/j.neuroimage.2015.06.052

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America*, *132*(2), 1061–1077. doi:10.1121/1.4728187

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective & Behavioral Neuroscience*, *7*(1), 1–17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17598730

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex (New York, N.Y. : 1991)*, *18*(3), 598–609. doi:10.1093/cercor/bhm091

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*(2), 333–8. doi:10.1016/j.neuroimage.2004.12.001

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. doi:10.1155/2011/156869

Orr, J. M., & Weissman, D. H. (2009). Anterior cingulate cortex makes 2 contributions to minimizing distraction. *Cerebral Cortex (New York, N.Y. : 1991)*, *19*(3), 703–11. doi:10.1093/cercor/bhn119

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *NeuroImage*, *29*(3), 797–807. doi:10.1016/j.neuroimage.2005.09.069

Roa Romero, Y., Senkowski, D., & Keil, J. (2015). Early and Late Beta Band Power reflects Audiovisual Perception in the McGurk

Illusion. *Journal of Neurophysiology*, jn.00783.2014. doi:10.1152/jn.00783.2014

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(2), 318–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8934846

Sánchez-García, C. (2013). *Cross-modal predictive mechanisms during speech perception*. Universitat Pompeu Fabra.

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PloS One*, *6*(10), e25198. doi:10.1371/journal.pone.0025198

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–40. doi:10.1016/j.neuron.2013.07.007

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(10), 2387–99. doi:10.1093/cercor/bhl147

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, *92*(3), B13–23. doi:10.1016/j.cognition.2003.10.005

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi:10.1037/h0054651

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212. doi:dx.doi.org/10.1121/1.1907309

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The*

*Quarterly Journal of Experimental Psychology Section A*, *36*(1), 51–74. doi:10.1080/14640748408401503

Szycik, G. R., Jansma, H., & Münte, T. F. (2009). Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Human Brain Mapping*, *30*(7), 1990–9. doi:10.1002/hbm.20640

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457–472. doi:10.1080/09541440340000268

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–6. doi:10.1073/pnas.0408949102

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Weissman, D. H., Warner, L. M., & Woldorff, M. G. (2004). The neural mechanisms for minimizing cross-modal distraction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *24*(48), 10941–9. doi:10.1523/JNEUROSCI.3669-04.2004

Zimmer, U., Roberts, K. C., Harshbarger, T. B., & Woldorff, M. G. (2010). Multisensory conflict modulates the spread of visual attention across a multisensory object. *NeuroImage*, *52*(2), 606–16. doi:10.1016/j.neuroimage.2010.04.245

# 3 Discussion

This dissertation proposed a framework based on three hypotheses:

1.Integration in AV speech will occur only if both modalities of the stimulus, the auditory and visual inputs, are attended.
2.If both modalities of the AV stimulus are attended an attempt to fuse them together will be made, independently of the congruency between the auditory and visual inputs.
3.If there exists an AV conflict (i.e. auditory and visual inputs are incongruent) conflict resolution processes are engaged to reduce the impact of this conflict in the final percept.

This chapter will focus on the interpretation of the results reported in this thesis with respect to these three concrete hypotheses, and discuss their implications within the current literature.

## 3.1 About the role of attention on AV speech integration

In the study presented in section 2.1, the aim of the experiment was to test if attention was necessary for multisensory integration to occur. For this, a prime example of multisensory integration was used: AV speech.

### *3.1.1 Behavior*

In behavior, it was found that attention must be placed in both stimulus modalities, auditory and visual, for multisensory

integration to occur. This could be seen in the fact that participants' performance was higher when their attention was focused in the congruent auditory and visual streams that contained the target words if compared to when they focused their attention in the auditory stream that was incongruent with the visual stream. Moreover, no indication of multisensory integration capturing participants' attention was found, as no increase in performance was found when the target word was located in the unattended auditory stream that was congruent with the visual information. Therefore the results support a non-automatic attention-dependent view of AV integration.

These results contrast strongly with some of previous literature dealing with this same question. If multisensory integration would have happened in an automatic way (Bertelson et al., 2000; Vroomen et al., 2001), then an increment in the performance when target words were located in a congruent unattended stream should have been found. Even more, if multisensory integration could capture attention as argued by, for example, Van der Burg and colleagues (2008) or Driver (1996), a decrease in performance should have been found when participants tried to attend an incongruent AV stream in the presence of an unattended congruent AV stream. However, none of this was found in the behavioral data.

However, a more established consensus on the need of attention for multisensory integration to take place is found in those studies specifically focused on AV speech. Although these two different

sets of studies, those in favor of the need of attention and those against the need of attention, seem to be irreconcilable, several factors can explain these differences. First, the nature of stimulation seems to be decisive in the need of attention for multisensory integration to occur. While in *beep and flash* studies automaticity seems to be more predominant, in AV speech integration, attention seems to be necessary. In fact, multisensory integration is probably a manifold process that expresses differently depending on the nature of the stimulation or the task at hand. Second, previous studies suggest that probably we need to deplete the resources of the participant before attentional processes start having effect, something recently studied in the context of cocktail-party paradigms. This can be framed within the Perceptual Load Theory described by Lavie (1995), considered further in the following discussion.

One of the main findings arguing against the need of attention and in favor of automaticity in the context of AV speech is the one published by Jon Driver (1996) (see also Soto-Faraco et al., 2004 for another example), discussed in the introduction. In a series of three experiments, the automatism of AV speech integration was shown through the ventriloquist effect by illusorily displacing, closer or apart, the perceived source of two sounds using a video of the speaker in different locations. In his critical second experiment (see right panel of Figure 1), Driver argued that the small decrease in performance found in behavior was due to the ventriloquist effect that brought the apparent location of the unattended sound stream closer to the attended one. However,

one could also argue that this interference could be due to the overall reduction of information available when the lips were occluded. This might have freed more cognitive resources to process and segregate the two auditory streams. That is, one could see the lips as a source of interference in the control condition rather than as a source of ventriloquism in the experimental condition. Therefore, a control condition showing the video of a speaker not matching any of the two auditory streams could, in fact, have had a similar detrimental effect to that of the unoccluded lips.

In conclusion, an improvement in the word recognition task was only found when both the auditory and visual congruent modalities were attended, and no effect was observed due to congruency out of the focus of attention. Therefore, these results support that for AV integration to occur both modalities of the AV stimulus must be attended.

### *3.1.2 fMRI*

In the study presented in section 2.1, participants also had to perform the same task in an fMRI scanner. The main result, aligned with the hypothesis, was that the modulation found in areas previously related to multisensory integration–such as the STS–depended on the participants' focus of attention. This modulation in the BOLD response was not only found in heteromodal areas (STS) (Beauchamp, Nath, & Pasalar, 2010; G. a Calvert et al., 2000; Miller & D'Esposito, 2005; Nath & Beauchamp, 2012; Stevenson et al., 2010, 2011), but extended to

unimodal areas, such as the auditory and visual cortices (Driver & Noesselt, 2008; Macaluso & Driver, 2005; Schroeder & Foxe, 2005), as well as the motor cortex.

As discussed in the Introduction (section 1.1.2), the STS has been studied as one of the main locus of AV integration, particularly in AV speech. This area has proved to be more active when the observer is presented with AV stimulation than when the responses to auditory and visual stimulation alone are summed offline (Calvert et al., 2000). It has also been demonstrated that its activity is dependent on the congruency or incongruency of the presented stimuli (Ojanen et al., 2005; Pekkola et al., 2006; Szycik et al., 2009), their synchrony (Miller & D'Esposito, 2005), and that said activity relates to the perception of the McGurk illusion (Beauchamp et al., 2010; Nath & Beauchamp, 2012; Szycik et al., 2012). Even more, BOLD activity in the STS has previously been used as a probe for AV speech integration in attention manipulation studies (Fairhall & Macaluso, 2009). The pattern found in the STS in this study supports the hypothesis that attention on both modalities of the AV stimulus was needed for multisensory integration to occur, as the STS was more active when participants directed their attention towards a congruent AV stimulus than when they directed attention towards an incongruent AV stimulus. Critically, under inattention no differential activity depending on AV congruency was observed if compared to AV incongruent stimuli, indicating that the presence of congruent AV stimulation was not enough to observe an increase in activity, as compared to incongruent AV stimulation.

The conclusion drawn from this is that, for this increase to occur the congruent AV stimulus had to be attended.

In addition to the STS, a similar pattern of activation was found in the sensory-motor areas, close to the mouth region. This pattern can be interpreted within the framework of motor theory of speech (Liberman & Mattingly, 1985) that relates motor activity with AV speech perception (Skipper et al., 2005, 2007; Wilson, Saygin, Sereno, & Iacoboni, 2004), more particularly, with the translation of sounds into their articulatory movements. Particularly, activity in these areas has been proved to increase when the auditory input is accompanied by its visual counterpart (articulatory information).

An increase in activity was also found in the auditory cortex (superior temporal gyrus). This increase has previously been found in the context of AV speech, and has been related to the presence of AV information also in lip-reading (or speechreading) (Calvert et al., 2000; Miller & D'Esposito, 2005; Pekkola et al., 2005).

Very interestingly, a modulation in the visual cortex depending on the focus of attention of the participant was also found. A similar pattern to that in previous areas was found: high BOLD signal when attending AV congruent information and low signal when attending an AV incongruent stimulus. This was interpreted as a deeper processing of the visual information that proved to be beneficial for the task, as demonstrated by the behavioral data. Nonetheless, the interpretation of this

modulation may not be restricted just to an up-regulation of visual processing in the cases were attention was directed towards a congruent AV stimulus. Instead–or in addition, perhaps–it could be associated to a down-regulation of visual processing when attention was directed towards an incongruent AV stimulus. The latter becomes the more relevant possibility if we try to explain the role of another set of areas in which BOLD activity increased when attention was directed towards incongruent AV stimulation.

Two areas showed an increased pattern of activity when participants attended an AV incongruent stimulus: the ACC and the insula. These areas–specially the ACC–as explained in the introduction (section 1.2.2) are related to conflict perception and resolution, more particularly, to the moment when an automatic behavior is overridden and a non-automatic one must take control (Shenhav et al., 2013). This set of areas and interpretation is not unknown in the AV context (Noppeney et al., 2008; Orr & Weissman, 2009; Weissman et al., 2004; Zimmer et al., 2010), particularly in the AV speech (Miller & D'Esposito, 2005; Pekkola et al., 2006; Szycik et al., 2009).

### *3.1.3 Conclusion*

Data from behavior and fMRI converge in supporting the first of the hypotheses: that for integration in AV speech to occur both modalities of the stimulus, the auditory and visual, must be attended.

Nonetheless, I propose a framework that encompasses not only the increase in activation in multisensory areas when AV congruent stimulus are processed, but also the down-regulation of the visual areas accompanied with an increased activity in conflict areas when processing (attending) an incongruent AV stimulus.

From an introspective point of view, multisensory integration seems hard to access and modulate at will, as opposed to top-down attention, for example. Therefore, if we cannot choose to integrate then one feasible course of events is that once both sensory modalities have been processed to a certain extent, an attempt to integrate their contents will occur, regardless of congruency. If this attempt is unsuccessful, that is, it is impossible to fuse these two sources of information, the conflict mechanism will detect the mismatch and try to reduce its impact.

In the case of the study described in section 2.1, when participants attend a congruent stimulus, an attempt to integrate AV information is made and, as both pieces of information are congruent, this attempt is successful. This successful integration is reflected in the increase of activity in areas previously described as multisensory or in the improvement in behavior. If, on the other hand, the AV information is incongruent, the attempt is unsuccessful and conflict areas are engaged instead. As we assume that the multisensory integration process cannot be regulated itself, the way of reducing the impact of the mismatch in this case is the down-regulation of the least informative, or

useful, of the two modalities given the task at hand–in this case, the visual one.

Recent studies have connected the predictive coding framework with speech perception (Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011; Sánchez-García, Enns, & Soto-Faraco, 2013; van Wassenhove et al., 2005). Pickering & Garrod (2006) proposed a parallelism between action and speech production on the one hand, and action and speech perception on the other hand. They proposed that, in the case of perception, listeners constantly update predictions about the likely upcoming content (based on phonology, words, grammatical category…) to facilitate (speed up) processing. The articulatory (motor) production system is essential to generate these predictions which are, in turn, compared with the actual incoming input. This comparison produces a corresponding error signal as described in the predictive coding framework. In my proposal, a possible interpretation within a predictive coding framework can easily be made by linking the error signal with the difficulty to fuse AV speech and the activity in conflict areas and linking the update of the internal model with the down-regulation of the visual input (see Skipper et al., 2007 for a similar approach).

This hypothesis can also be framed into a Bayesian point of view. If one considers how rarely we find incongruent AV speech in the natural environment, it is logical to think that we have a strong prior to fuse AV speech information, as it has proven to be beneficial in most of the previous situations.

Although this strong prior, or tendency, to fuse AV information may resemble the pre-attentive AV integration hypothesis defended in other studies (Bertelson et al., 2000; Driver, 1996), in this thesis, a clear distinction between the two is made. While the preattentive AV speech integration hypothesis states that this integration occurs prior to attention, the hypothesis presented in this section sets as requisite that both modalities have to be processed under attention to a certain degree for a fusion attempt to occur. Previous results defending a preattentive integration process can be explained if we consider that these studies may have failed to deplete attentional resources, and therefore both stimulus were processed using the remaining attentional resources and then integrated (Lavie, 1995).

Based on these results, a new testable hypothesis was proposed: when an AV mismatch in speech is detected as a conflict, an attempt to minimize its impact will be made. In the next section I will discuss the implications of this conflict hypothesis in taking into account the results of this thesis.

## 3.2 About the role of conflict during incongruent AV speech perception

After the results of the study in section 2.1, one logical follow up was to test the role of conflict in AV integration for incongruent stimuli, especially during the McGurk illusion. The studies in section 2.2 and section 2.3 used similar paradigms trying to answer if conflict areas were active during incongruent AV

speech stimulation and if these areas were involved in the resolution of this AV conflict.

First, it was addressed if brain areas previously related to conflict in other paradigms (mainly the ACC and the IFG; Nee et al., 2007; Shenhav et al., 2013) were also involved in the perception of incongruent AV speech. For this, a simple paradigm in which participants were presented with congruent and incongruent stimuli while recording BOLD signal using fMRI was used. Second, a similar study was carried out but, instead of BOLD signal, EEG signal was recorded. In this second case, an increase in power of the theta band was expected when comparing incongruent and congruent AV speech, as this particular band has previously been related to conflict processing (Cavanagh & Frank, 2014; Cohen, 2014; Hanslmayr et al., 2008).

In both cases the results matched the hypothesis that incongruent AV speech engages conflict mechanisms, the aforementioned activity in the ACC and IFG, and the modulation of power in the theta band was found.

Moreover, the critical point in this set of studies was not only to see if incongruent AV speech engaged conflict mechanisms but also if one of the most prevalent effects used to study AV speech integration, the McGurk effect, also engaged this same set of mechanisms. As explained during the introduction (see section 1.2) the McGurk effect has been used many times as a model for AV integration (e.g., Alsius et al. 2005; van Wassenhove, Grant, and Poeppel 2007; Skipper et al. 2007; Tiippana, Andersen, and

Sams 2004; Bernstein et al. 2008; Andersen et al. 2009) when in fact it differs from regular AV integration in a fundamental aspect: the congruency between the visual and auditory inputs.

Critically, the study revealed a pattern of activation in the conflict-related areas when measuring BOLD activity during the McGurk effect, as well as an increase in the theta band power, compared to a regular AV congruent stimulus. Even more, according to the fMRI data, the McGurk effect did not only engage conflict brain areas, but these areas also showed differential activity depending on the outcome of the McGurk illusion. This result suggests that the role of conflict mechanisms is not reduced to detecting AV conflict, but that they are involved in the perception of the McGurk illusion.

### *3.2.1 fMRI*

As mentioned above, one of the main findings of this thesis is that a network of brain areas similar to the one responding in studies of conflict (ACC and lIFG) was found when participants were presented with an incongruent AV stimulus, regardless of the nature of the stimuli–McGurk or Non McGurk. This supports the results found in the previous study suggesting that an incongruent AV stimulus is perceived as a conflict and helps to generalize the role of these conflict areas to incongruent AV speech. As previously mentioned, activity in conflict areas is not unknown in the context of multisensory integration, both outside the domain of speech (Noppeney et al., 2008; Orr & Weissman, 2009; Weissman et al., 2004; Zimmer et al., 2010) and inside the

domain of speech, although it has rarely been discussed in this last context (Miller & D'Esposito, 2005; Morís Fernández et al., 2015; Ojanen et al., 2005; Pekkola et al., 2006; Szycik et al., 2009).

*Anterior Cingulate Cortex*

The role of the ACC has been described many times in the context of conflict, and has usually been associated with its detection and resolution (Nee et al., 2007; Shenhav et al., 2013). Particularly, it has been suggested that the ACC does not resolve the conflict by itself but would be in charge of recruiting additional brain areas to try and solve the detected conflict (Shenhav et al., 2013). The pattern of BOLD activity within the study matches very well the possible conflict-detection role of the ACC, as this area was found to be more active in all conditions that showed AV conflict. Even more, this area also showed different BOLD signal depending on the outcome of the McGurk illusion, thus indicating that the level of activity in this area may be determinant to the outcome of the McGurk illusion. Based on the pattern of activity of the lIFG and the lPC, I suggest that these may be the two areas recruited by the ACC to resolve this AV conflict.

*Inferior Frontal Gyrus and Left Precentral Cortex*

It has been proposed that the IFG plays different roles in AV speech stimuli processing. For example, mapping the speech inputs (visual and auditory) into motor representations of the articulations, in line with the motor theory of speech, this common representational space will allow its combination. The pattern of activity found in this study–higher activation during

incongruent AV stimulation–has been found in previous studies and interpreted as an increased activity due to the need of processing two different inputs during incongruent stimulation (one auditory and one visual) against only one in the congruent condition (the same input auditory and visual) (Ojanen et al., 2005). Hasson and colleagues (2007) claimed that the IFG and premotor regions were in charge of generating an initial hypothesis about the possible speech input. In fact, they suggested an information flow similar to that of a closed-control circuit, in which the goal is to minimize the discrepancy between the initial hypothesis generated in the IFG and the actual sensory information. Along a similar line, Miller & D'Esposito (2005) proposed that activity in the IFG deals with conflicting or noisy representations, and speculated about a possible differentiation between automatic processing (under regular conditions) in posterior regions against high-order controlled processes (under noisy conditions) in more frontal regions.

Present data indicates that the IFG and lPC are involved in the processing of incongruent AV speech, as shown by their increased response to AV incongruent stimulation. Moreover, as in the case of ACC, activity in this area was different depending on the outcome of the McGurk illusion. This suggests the implication of these areas is not limited to the detection of the conflict but also in the resolution of the AV conflict particularly in the McGurk illusion. This enhanced activity is interpreted as a general increase in the processing needs when a conflict between the two sensory modalities is detected (Hasson et al., 2007; Miller &

D'Esposito, 2005; Pekkola et al., 2005). In particular, as explained before, I propose that this area could be in charge of solving the discrepancy between the prediction produced by the visual information and the following auditory information. This role is not far from that proposed by Ojanen (2005), Miller & D'Esposito (2005) or Hasson et al. (2007). The main difference between their proposals and the one made in this thesis is that here activity in the lIFG and the lPC is driven by the ACC, which would detect the conflict and then recruit the lIFG and the lPC to solve it.

*Angular Gyrus*

A higher activity in the angular gyrus when comparing the congruent condition with the incongruent conditions was found. This area has been related to many different domains (Seghier, 2013), and particularly to the perception of AV speech (Bernstein, Auer, et al., 2008). In the scope of this thesis it showed a differential pattern depending on the congruency, and also depending on the outcome of the McGurk stimulus. It may be partially related to the proposed switch from posterior areas to more frontal areas depending on an automatic or more controlled processing of the stimulus, as proposed by Miller & D'Esposito (2005), albeit their posterior areas were located in the intraparietal sulcus. Nonetheless, given the post-hoc nature of this interpretation, the role of the angular gyrus remains speculative in this context.

*Superior Temporal Sulcus*

During this study it was expected to find activity in the superior temporal sulcus, as its involvement in AV speech integration and the perception of the McGurk illusion has frequently been found in previous literature. Nonetheless this pattern is not always consistently found in all studies. One possible explanation, suggested by Nath & Beauchamp (2012) or Stevenson et al. (2011) is that this area location is highly variable across subjects, and therefore group analysis may not be sensitive enough to detect changes in its activity.

Altogether, this pattern shows that a network of areas previously related to conflict may play a role in conflict detection (ACC) and conflict resolution (ACC, IFG, IPC). Above and beyond the differential response to AV conflict, the response of these three areas showed a differential pattern depending on the outcome of the McGurk stimulus, illusory or non-illusory. That is, to the same physical stimuli, the BOLD responses were stronger when the outcome was illusory than when the illusion did not take place. This suggests that the perception of the illusion from a behavioral point of view was dependent on the level of activity in these areas. Particularly, when the illusion is not perceived, I speculate that the ACC fails to recruit the lIFG gyrus in the case of the non-illusory McGurk, although probably a partial recruitment might still occur as activity is higher than in the case of congruent

stimuli. Because of this, no integration between both percepts would take place.

In conclusion, areas previously related to conflict–ACC, lIFG and lPC–were more active when an AV conflict was present. Even more, these areas showed a differential activity depending on the outcome of the McGurk illusion, which indicated that they were not only involved in the detection but also in the resolution of the AV conflict. These results support the last hypothesis of this thesis, that in order to reduce the impact of AV conflict, conflict processes are activated.

### *3.2.2 EEG*

The last hypothesis of this thesis, namely, the engagement of conflict processes to reduce the impact of AV conflict, was tested again using EEG in the last study of this thesis, found in section 2.3. This study used a well-known marker of conflict in EEG, a power increase in the theta band in the midfrontocentral electrodes. For this, a paradigm similar to that in the previous study was used. In this case it was hypothesized that an increase in the theta power band should occur once the visual prediction is violated by the auditory component of the AV syllable. Therefore the analysis was time locked to the auditory onset. If incongruent AV speech (especially in the case of the McGurk effect) is perceived as a conflict, a modulation in the theta band when contrasting McGurk trials with congruent AV speech trials would be expected.

As predicted, the study showed a power modulation in the theta band peaking at ~20 ms after the auditory onset when comparing the McGurk condition against the congruent condition, although, given the limitations of this analysis that uses a 500 ms window, caution must be taken if interpreting this timing very strongly. Nonetheless, said timing coincides with the hypothesis, as the theta modulation occurs as soon as the visual prediction is violated at the onset of the auditory stimulus.

A similar effect when we compare the incongruent condition with the congruent one occurs, albeit this effect is a bit delayed and more extended in time, maybe due to the inability to reconcile both modalities, or to the larger incongruency produced by a reversed video that required a longer processing of the stimulus.

Critically, in this EEG study similar results were found when comparing AV congruency/incongruency with the results of the Stroop task run in the same group of participants. When comparing the Incongruent and Congruent conditions during the Stroop task, the effect generally coincided with previous findings (Hanslmayr et al., 2008; MacLeod, 1991), and peaked at the same frequency (6 Hz) as the effect with speech, reinforcing that we are addressing a similar process, albeit the timing is very different as the two tasks differ widely in their processing needs.

### *3.2.3 Conclusions*

The results in the last two studies of this thesis (sections 2.2 and 2.3) support that incongruent AV speech is perceived and

processed as a conflict. This result is interpreted within the predictive coding framework, in which the prediction created by the visual input is violated by the auditory information (Arnal, Wyart, & Giraud, 2011; Morís Fernández et al., 2015; Pickering & Garrod, 2006; Sánchez-García et al., 2011, 2013; Skipper et al., 2007; van Wassenhove et al., 2005). This mismatch engages conflict processing areas that try to reduce its influence or correct it if possible.

Present results extend the activity in conflict solving areas to the McGurk effect, in which the mismatch between the auditory and visual modalities also engaged the ACC and lIFG. Even more the ACC and the lIFG also showed different patterns of activity depending on the outcome of the McGurk illusion indicating that they were not only involved in the detection of the AV conflict but also played a role in its resolution.

## 3.3 About the automaticity of AV speech processing

The second hypothesis of the framework introduced in this thesis proposed that, whenever we perceive and attend AV speech information, an attempt to integrate both modalities occurs independently of the congruency between them. This hypothesis is proposed based on the activity of the ACC and the IFG, which is higher when an incongruent AV stimulus is presented if compared to when an AV congruent stimulus is presented. One possible interpretation of these data is that, once the attempt to integrate incongruent AV information fails, conflict processes are engaged to solve this conflict. This interpretation fits within a

Bayesian framework if we think about how often we perceive and attend AV incongruent speech information (which is very unlikely), and how often we perceive and attend AV congruent speech (which is very likely). Therefore, I propose that it is more parsimonious and effective to have and automatic pathway working in most of everyday situations.

Nonetheless, this hypothesis was not directly tested during this thesis, and alternative proposals also exist. For example, Nahorna and colleagues (2012, 2015) propose a two-stage model in which a binding-unbinding stage occurs before the integration stage. This binding-unbinding state is in charge of deciding if the auditory and visual information should be bound together and therefore integrated. This decision to bind information together or not is based on the previous history of congruency between the auditory and visual information. If previous AV speech information is congruent, there is a higher tendency to bind–and in a following stage integrate–information, as compared to when previous AV speech information is incongruent. Tentatively this approach can be interpreted within the framework postulated in this thesis if the binding-unbinding stage proposed by Nahorna et al. is a consequence of the failure to integrate the AV information and a subsequent readjustment of the system based on this conflict produced by the AV mismatch. However, further research is needed to directly test the second hypothesis of the framework proposed in the thesis.

## 3.4 General Discussion

I will now discuss the impact of the findings of this thesis framed within the current state of the art in integration of AV speech, presented during the Introduction (Chapter 1), and particularly in the debate regarding its automaticity/non-automaticity.

The main core of results defending the automaticity of AV speech integration comes from the McGurk effect. For example, participants could not break the illusion even when they were informed of the nature of the stimulus, nor they could not distinguish a McGurk trial from a regular (AV matching) one (McGurk & MacDonald, 1976). Moreover, it was shown that infants perceived this illusion (Rosenblum et al., 1997), that it elicited a mismatch negativity effect (Colin et al., 2002) and that even cognitive incongruency could not diminish it (Green et al., 1991). All these pieces of evidence suggest, albeit indirectly, the automatic nature of the illusion, as it seems independent of various forms of previous knowledge. These studies were used as proof of the McGurk effect and, by extension of the AV speech integration process, being low level preattentive automatic processes. Nonetheless they did not directly address the automaticity or its preattentive nature. In fact, studies directly addressing this independence from attention in AV integration challenged it being automatic and pre-attentive and found that the actual focus of attention was determinant for this AV integration, and therefore the illusion, to occur.

This thesis has presented a tentative framework that may explain this apparent disparity of results. First, I propose that attention is needed for this integration to occur. It may be the case that previous studies finding automatic AV speech integration did not deplete completely the attentional resources and therefore stimuli were processed under the remaining resources of attention.

The second hypothesis of the framework states that anytime an AV speech stimulus is perceived and sufficiently processed, an attempt to integrate the auditory and visual inputs–independently of the congruency between them–will be made. This second hypothesis is based on a Bayesian point of view of our perception, in which, although our senses are exposed to incongruent sources of information (i.e., any two spurious crossmodal inputs in a natural environment), we rarely attend both of them at the same time. In fact, most of the time we deploy our attention on objects providing coherent information across modalities (i.e. the face and the voice of our speaker in the presence of the voice of a third person, as seen in section 2.1). Therefore, I speculate that we have a strong prior to fuse AV information by default given that, on the majority of situations found in our daily life, it provides an advantage. It is this strong prior that may have been confused with automaticity if combined with the insufficient depletion of attentional resources.

The third piece of this puzzle is what happens when this automatic attempt to fuse two attended sources of AV speech information fails, for example, due to lips and speech sounds

being incongruent as seen in the experiment in section 2.1 or the McGurk illusion. I hypothesize that the attempt to integrate two attended mismatching pieces of information activates conflict resolution processes to reduce the mismatch impact. The results of the studies presented in this thesis suggest the involvement of conflict detection and resolution mechanisms, also found in other conflict tasks such as the Stroop task. The usual role of conflict processes is trying to diminish the impact of conflict to achieve the task at hand. In this thesis, two different ways to reduce the impact of the AV conflict were found. In the first study (section 2.1), the AV conflict was detected (indexed by the activity in the ACC) and solved by a down-regulation of the visual areas that was interpreted as a reduced processing of the visual modality. Introspectively speaking, AV integration seems not easily accessible by volition and, as hypothesized, the attempt to integrate AV information occurs automatically if attention is deployed to the inputs. Therefore, the way to stop attempting AV integration is by modulating the processing of the inputs. A different situation occurs during the McGurk illusion, in which an increased activity in areas previously related to language (lIFG), as well as in the ACC, was found. In this case, the same logic is applied to the detection of the conflict, but now the resolution is achieved by an extra processing of the AV input, maybe reconciling both modalities as in this case the incongruency is less pronounced, and producing the illusion (for further insight refer to the previous section framing it in predictive coding).

This framework offers a new approach to the AV integration process by including high level executive processes, namely, conflict detection and resolution mechanisms in AV integration. Previous explanations described the AV integration as a low level, early process based on its automatic nature. The suggestion in this thesis is that the AV congruent speech integration may occur as a default strategy (though not automatic in the classical sense), and only when the AV information is incongruent, conflict processes come into play. This is of special relevance in the case of the McGurk effect, as this thesis does not support its categorization as an automatic low level effect but suggests that it is mediated by executive processes and high level areas.

The state of affairs described above grants some discussion about the possible routes for AV integration in the brain. In general, direct anatomical connections allowing an influence of the visual cortex over the auditory cortex had already been described (Cappe & Barone, 2005; Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003). Though this direct route has mostly been described in anatomical studies in animals, its impact in human multisensory integration has been largely speculated, albeit not concluded (Driver & Noesselt, 2008). In addition to these direct connections, indirect connections between different sensory inputs have also been described through supramodal areas, such as the STS. This route has been described many times, specifically in the processing of AV speech (Beauchamp, Lee, Argall, & Martin, 2004; Ghazanfar, Maier, Hoffman, & Logothetis, 2005). These two different pathways have been

studied and distinguished recently by Arnal and colleagues (2009).

Arnal and colleagues (2009) described an initial influence of the visual information in the form of a fast prediction. This fast prediction is not constrained by AV congruency and occurs independently of it. The second indirect route through the STS is described as sensitive to congruency: If both signals are congruent, then a fine-tuning and a speed-up of the AV processing occurs. However, if both signals are incongruent, a higher activity in the STS occurs due to the need of processing both signals separately, and then a feedback signal is sent to the auditory cortex, either producing a successful AV tuning in the form of the McGurk illusion or a failure and subsequent perception of a mismatch.

In a study from the same group (Arnal et al., 2011) based on the predictive coding framework, using magnetoencephalography and speech stimuli, a switch in the neural dynamics depending on the congruency of the stimuli was shown. Particularly, they concluded that when the intermodal prediction (the prediction about the upcoming auditory stimulus created by the visual part of the stimuli) was violated, an increase in the coordination between beta (focalized in the STS) and high-gamma (focalized in the low level sensory areas) oscillatory activity occurred. They interpreted this finding as an updating of the prediction in the STS that was conveyed to low level areas and generated new prediction errors. This thesis adds to this literature by suggesting

that in the case of AV mismatch, there also exist an involvement of high level areas (ACC and IFG) in the resolution of this AV conflict, in addition to the STS and low level sensory areas. Whether the result of the activity in the ACC and IFG is conveyed directly to low level areas or to the STS is an interesting question for future research, but it is not answered within the scope of this thesis.

The implication of high level areas in the perception of the McGurk illusion does not only challenge its automaticity but also indicates that the McGurk illusion is not equivalent to the integration during a regular AV congruent speech stimulus. As explained in the introduction, the McGurk effect has very often been used to infer properties of the general process of AV integration. The results of this thesis suggest that the McGurk illusion is at least partially mediated by areas previously related to conflict, which indicates that the process of integration in the McGurk illusion may differ from that of regular AV congruent integration. These findings have implications for previous and future studies using the McGurk effect as a probe for the AV integration process in speech, as they suggest that generalizing properties from the McGurk effect to the general process of AV integration should be done with care.

Even more, as mentioned at the beginning of this section, several studies reported that the incongruent nature of the McGurk stimuli is unperceived by the observer (Möttönen, Krause, Tiippana, & Sams, 2002; Summerfield & McGrath, 1984).

Nonetheless, others have challenged this vision and defended that the experience of a McGurk stimulus is different from that of an AV congruent stimulus from a subjective point of view (Soto-Faraco & Alsius, 2009; van Wassenhove et al., 2007). In fact, when presented for the first time with a McGurk stimulus, very often participants report an odd feeling, even when they cannot report exactly why. One possible speculation over this odd feeling is that it is related with the detection and resolution of the AV conflict during the McGurk effect.

Summing up, within this section I have presented a feasible explanation of the AV speech integration process reconciling these findings consistent with the automaticity of this integration and those against within a common framework.

# 4 CONCLUSIONS

The studies included in this thesis advance two main conclusions in the investigation of the mechanisms underlying the perception of AV speech.

First, this thesis shows that the inner state of the observer is relevant to the multisensory integration process. In particular, for AV speech at least, both modalities must be under the focus of attention to result in integration.

Second, this thesis shows that conflict processes are engaged when incongruent AV speech is presented, and that these conflict

processes are not only involved in the detection of the mismatch between the auditory and visual information but also extend into the resolution of this AV conflict by readjusting the integration process.

Therefore, I propose that the AV speech integration process is not automatic, and in the case of incongruent AV speech, it implicates the recruiting of high level areas. Said recruitment of high level areas becomes very relevant when considering the McGurk effect, which has been used as a probe for AV integration widely in literature. Present results suggest that this effect is not purely the outcome of the AV integration process, but also conflict processing is involved in this illusion.

# 5 Questions for Future Research

To this point, evidence suggests that conflict mechanisms are involved in the perception of the McGurk illusion. Nonetheless, evidence presented only shows a correlation between these mechanisms and the McGurk illusion. One possible way of demonstrating causality between conflict and the perception of the McGurk illusion would be to interfere with activity in this conflict related areas and measure the influence of this interference on the behavior. This could be operationalized by interfering with the activity in the ACC by means of transcraneal magnetic stimulation, or transcraneal direct current stimulation. A modulation of the participant's response due to this interference with activity in the anterior cingulate cortex modulates the behavior of participants, would strengthen our point and highlight a possible causal role between the McGurk illusion and conflict mechanisms.

# 6 References

Alais, D., Newell, F. N., & Mamassian, P. (2010). *Multisensory processing in review: from physiology to behaviour. Seeing and perceiving* (Vol. 23, pp. 3–38). doi:10.1163/187847510X488603

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, T. 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

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., Morillon, B., Kell, C. a, & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(43), 13445–53. doi:10.1523/JNEUROSCI.3194-09.2009

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. doi:10.1038/nn.2810

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus. *Neuron*, *41*(5), 809–823. doi:10.1016/S0896-6273(04)00070-4

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The*

*Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(7), 2414–7. doi:10.1523/JNEUROSCI.4865-09.2010

Benoit, M. M., Raij, T., Lin, F.-H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping*, *31*(4), 526–38. doi:10.1002/hbm.20884

Bermant, R. I., & Welch, R. B. (1976). EFFECT OF DEGREE OF SEPARATION OF VISUAL-AUDITORY STIMULUS AND EYE POSITION UPON SPATIAL INTERACTION OF VISION AND AUDITION. *Perceptual and Motor Skills*, *43*(2), 487–493. doi:10.2466/pms.1976.43.2.487

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage*, *39*(1), 423–35. doi:10.1016/j.neuroimage.2007.08.035

Bernstein, L. E., Lu, Z.-L., & Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, *1242*, 172–84. doi:10.1016/j.brainres.2008.04.018

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

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–52. doi:10.1016/j.bandl.2012.10.008

Botvinick, M., & Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature*, *391*(6669), 756. doi:10.1038/35784

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive, Affective & Behavioral Neuroscience*, *7*(4), 356–66. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18189009

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–46. doi:10.1016/j.tics.2004.10.003

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

Calvert, G. a, Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649–657. doi:10.1016/S0960-9822(00)00513-3

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

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *The European Journal of Neuroscience*, *22*(11), 2886–902. doi:10.1111/j.1460-9568.2005.04462.x

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. doi:10.1016/j.tics.2014.04.012

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. doi:10.1371/journal.pcbi.1000436

Cohen, M. X. (2014). A neural microcircuit for cognitive conflict detection and signaling. *Trends in Neurosciences*, *37*(9), 480–490. doi:10.1016/j.tins.2014.06.004

Colin, C., Radeau, M., Soquet, a, & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *115*(9), 1989–2000. doi:10.1016/j.clinph.2004.03.027

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk – MacDonald effect : a phonetic representation within short-term memory, *113*, 495–506.

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

Ergen, M., Saban, S., Kirmizi-Alsan, E., Uslu, A., Keskin-Ergen, Y., & Demiralp, T. (2014). Time-frequency analysis of the event-related potentials associated with the Stroop test.

*International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *94*(3), 463–72. doi:10.1016/j.ijpsycho.2014.08.177

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. doi:10.3758/BF03203267

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

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical Evidence of Multimodal Integration in Primate Striate Cortex, *22*(13), 5749–5759.

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

Ghazanfar, A. a, Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(20), 5004–12. doi:10.1523/JNEUROSCI.0799-05.2005

Ghazanfar, A. a, & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–85. doi:10.1016/j.tics.2006.04.008

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection, (June 1999), 1197–1208.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677. doi:10.1121/1.422788

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*(6), 524–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1780200

Hanslmayr, S., Pastötter, B., Bäuml, K.-H., Gruber, S., Wimber, M., & Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *Journal of Cognitive Neuroscience*, *20*(2), 215–25. doi:10.1162/jocn.2008.20020

Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory

representation. *Neuron*, *56*(6), 1116–26. doi:10.1016/j.neuron.2007.09.037

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure & Function*, *212*(2), 121–32. doi:10.1007/s00429-007-0154-0

Keil, J., Müller, N., Ihssen, N., & Weisz, N. (2012). On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(1), 221–31. doi:10.1093/cercor/bhr125

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

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

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. doi:10.1016/0010-0277(85)90021-6

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the

human brain. *Trends in Neurosciences*, *28*(5), 264–71. doi:10.1016/j.tins.2005.03.008

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163–203. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2034749

Malfait, N., Fonlupt, P., Centelles, L., Nazarian, B., Brown, L. E., & Caclin, A. (2014). Different neural networks are involved in audiovisual speech perception depending on the context. *Journal of Cognitive Neuroscience*, *26*(7), 1572–86. doi:10.1162/jocn_a_00565

Massaro, D., & Stork, D. (1998). Speech Recognition and Sensory Integration. *American Scientist*, *86*(3), 236. doi:10.1511/1998.3.236

Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience*, *26*(3), 606–20. doi:10.1162/jocn_a_00515

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1012311

McNeill, D. (1992). *No Title* (p. 423). Chicago: The University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought* (p. 328). Chicago: The University of Chicago Press.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(25), 5884–93. doi:10.1523/JNEUROSCI.0896-05.2005

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S. (2015). Top-down attention regulates the neural expression of audiovisual integration. *NeuroImage*, *119*, 272–285. doi:10.1016/j.neuroimage.2015.06.052

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Research. Cognitive Brain Research*, *13*(3), 417–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11919005

Munhall, K. G., ten Hove, M. W., Brammer, M., & Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Current Biology : CB*, *19*(9), 735–9. doi:10.1016/j.cub.2009.03.019

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America*, *132*(2), 1061–1077. doi:10.1121/1.4728187

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2015). Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America*, *137*(1), 362–77. doi:10.1121/1.4904536

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, *59*(1), 781–7. doi:10.1016/j.neuroimage.2011.07.024

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, *11*(1), 4–11. doi:10.1016/j.inffus.2009.04.001

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84–93. doi:10.1016/j.brainres.2010.01.059

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective & Behavioral Neuroscience*, *7*(1), 1–

17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17598730

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex (New York, N.Y. : 1991)*, *18*(3), 598–609. doi:10.1093/cercor/bhm091

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*(2), 333–8. doi:10.1016/j.neuroimage.2004.12.001

Orr, J. M., & Weissman, D. H. (2009). Anterior cingulate cortex makes 2 contributions to minimizing distraction. *Cerebral Cortex (New York, N.Y. : 1991)*, *19*(3), 703–11. doi:10.1093/cercor/bhn119

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *NeuroImage*, *29*(3), 797–807. doi:10.1016/j.neuroimage.2005.09.069

Pekkola, J., Ojanen, C. A. V., Autti, T., Jìiskelìinen, I. P., Riikka, M., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech : an fMRI study at 3 T, *16*(2), 5–8.

Pickering, M. J., & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, *4*(2-3), 203–228. doi:10.1007/s11168-006-9004-0

Roa Romero, Y., Senkowski, D., & Keil, J. (2015). Early and Late Beta Band Power reflects Audiovisual Perception in the McGurk Illusion. *Journal of Neurophysiology*, jn.00783.2014. doi:10.1152/jn.00783.2014

Roberts, K. L., & Hall, D. a. (2008). Examining a supramodal network for conflict processing: a systematic review and novel functional magnetic resonance imaging data for related visual and auditory stroop tasks. *Journal of Cognitive Neuroscience*, *20*(6), 1063–78. doi:10.1162/jocn.2008.20074

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, *50*(1-2), 19–26. doi:10.1016/S0167-8760(03)00121-1

Rosenblum, L. D., Schmuckler, M. a, & Johnson, J. a. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9136265

Ross, L. a, Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy

environments. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(5), 1147–53. doi:10.1093/cercor/bhl024

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–145. doi:10.1016/0304-3940(91)90914-F

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PloS One*, *6*(10), e25198. doi:10.1371/journal.pone.0025198

Sánchez-García, C., Enns, J. T., & Soto-Faraco, S. (2013). Cross-modal prediction in speech depends on prior linguistic experience. *Experimental Brain Research*, *225*(4), 499–511. doi:10.1007/s00221-012-3390-3

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, "unisensory" processing. *Current Opinion in Neurobiology*, *15*(4), 454–8. doi:10.1016/j.conb.2005.06.008

Searle, J. R. (2000). Consciousness. *Annual Review of Neuroscience*, *23*, 557–78. doi:10.1146/annurev.neuro.23.1.557

Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *19*(1), 43–61. doi:10.1177/1073858412440596

Senkowski, D., Talsma, D., Herrmann, C. S., & Woldorff, M. G. (2005). Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *166*(3-4), 411–26. doi:10.1007/s00221-005-2381-z

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147–152. doi:10.1016/S0926-6410(02)00069-1

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–40. doi:10.1016/j.neuron.2013.07.007

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. doi:10.1037/h0020586

Simons, D. J., & Chabris, C. F. (2000). Gorillas in our midst : sustained inattentional blindness for dynamic events, *28*, 1059–1074.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *NeuroImage*, *25*(1), 76–89. doi:10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(10), 2387–99. doi:10.1093/cercor/bhl147

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport*, *18*(4), 347–50. doi:10.1097/WNR.0b013e32801776f9

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(2), 580–7. doi:10.1037/a0013483

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, *92*(3), B13–23. doi:10.1016/j.cognition.2003.10.005

Soto-Faraco, S., Sinnett, S., Alsius, A., & Kingstone, A. (2005). Spatial orienting of tactile attention induced by social cues. *Psychonomic Bulletin & Review*, *12*(6), 1024–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16615323

Spence, C., & Driver, J. (1996). Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology*.

*Human Perception and Performance*, *22*(4), 1005–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8756965

Spence, C., & Driver, J. (2004). *Crossmodal Space and Crossmodal Attention*. Oxford University Press. doi:10.1093/acprof:oso/9780198524861.001.0001

Stein, B. E. (2012). *The New Handbook of Multisensory Processing* (p. 840). The MIT Press.

Stein, B. E., & Meredith, M. A. (1993). *The Merging of the Senses* (1st ed.). Cambridge: MIT Press.

Stevenson, R. a, Altieri, N. a, Kim, S., Pisoni, D. B., & James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage*, *49*(4), 3308–18. doi:10.1016/j.neuroimage.2009.12.001

Stevenson, R. a, VanDerKlok, R. M., Pisoni, D. B., & James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *NeuroImage*, *55*(3), 1339–45. doi:10.1016/j.neuroimage.2010.12.063

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi:10.1037/h0054651

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212. doi:dx.doi.org/10.1121/1.1907309

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology Section A*, *36*(1), 51–74. doi:10.1080/14640748408401503

Szycik, G. R., Jansma, H., & Münte, T. F. (2009). Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Human Brain Mapping*, *30*(7), 1990–9. doi:10.1002/hbm.20640

Szycik, G. R., Stadler, J., Tempelmann, C., & Münte, T. F. (2012). Examining the McGurk illusion using high-field 7 Tesla functional MRI. *Frontiers in Human Neuroscience*, *6*(April), 95. doi:10.3389/fnhum.2012.00095

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral Cortex (New York, N.Y. : 1991)*, *17*(3), 679–90. doi:10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20675182

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the

evoked brain activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–114. doi:10.1162/0898929054475172

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457–472. doi:10.1080/09541440340000268

Tiippana, K., Puharinen, H., Möttönen, R., & Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing and Perceiving*, *24*(1), 67–90. doi:10.1163/187847511X557308

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(5), 1053–65. doi:10.1037/0096-1523.34.5.1053

Van Ee, R., van Boxtel, J. J. a, Parker, A. L., & Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(37), 11641–9. doi:10.1523/JNEUROSCI.0873-09.2009

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United*

*States of America*, *102*(4), 1181–6. doi:10.1073/pnas.0408949102

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vroomen, J., Bertelson, P., & de Gelder, B. (2001a). Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta Psychologica*, *108*(1), 21–33. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11485191

Vroomen, J., Bertelson, P., & de Gelder, B. (2001b). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, *63*(4), 651–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11436735

Weissman, D. H., Warner, L. M., & Woldorff, M. G. (2004). The neural mechanisms for minimizing cross-modal distraction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *24*(48), 10941–9. doi:10.1523/JNEUROSCI.3669-04.2004

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech

production. *Nature Neuroscience*, *7*(7), 701–2. doi:10.1038/nn1263

Zimmer, U., Roberts, K. C., Harshbarger, T. B., & Woldorff, M. G. (2010). Multisensory conflict modulates the spread of visual attention across a multisensory object. *NeuroImage*, *52*(2), 606–16. doi:10.1016/j.neuroimage.2010.04.245

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." *Neuron*, *77*(5), 980–991. doi:10.1016/j.neuron.2012.12.037