CROSS-MODAL PREDICTIVE MECHANISMS DURING SPEECH PERCEPTION

Carolina Sánchez García

DOCTORAL THESIS UPF / 2013

SUPERVISOR

Dr. Salvador Soto-Faraco

Dept. de Tecnologies de la Informació i les
Comunicacions

UNIVERSITAT
POMPEU FABRA

A mi gente.

Acknowledgments

Esta tesis es el resultado de un trabajo hecho durante los últimos cuatro años.

Siento la tentación de nombrar a toda la gente que me ha ayudado, y que ha estado ahí día a día, uno por uno. A todos los que han estado presentes en mi vida aquí, que han hecho de Barcelona mi hogar.

Me siento muy afortunada de haber conocido a tanta gente maravillosa. Cuando llegué a Barcelona no me imaginaba que iba a conocer a tanta buena gente, ni que al marcharme echaría de menos a tantos buenos amigos.

Me gustaría dar las gracias a todas esas personas con las que he compartido estos años, dentro y fuera de la universidad, de las que he aprendido tantas y tantas cosas y con los que tan buenos momentos he pasado. A mis amigos, dentro y fuera de la universidad. Los que sois y los que fueron. Imposible dar nombres, sois demasiados. Vosotros habéis sido los que me habéis soportado día a día, y es sin duda gracias a vosotros, que he llegado a escribir esta tesis.

A toda la gente del MRG, y también de otros grupos, que he sentido como del mío. A mis compañeros de despacho, con los que he compartido trabajo y a la vez ratos inolvidables.

No habría podido encontrar compañeros y amigos mejores que vosotros. Me habéis hecho sentir tan a gusto que ha sido imposible distinguir la amistad del compañerismo y habéis hecho que ir al trabajo no fuera un esfuerzo, sino un placer.

Gracias también a todas las personas que me han hecho, con su paciencia y amabilidad, mi día a día más fácil. A Xavi y Silvia, por buscar soluciones a problemas técnicos irresolubles con tanto sentido del humor, gracias por haberme ayudado tantísimo. A Mireia, por estar siempre disponible para ayudarme a encontrar el pathway en matlab. A Santa, a Cristina, a Flore, por recibirme siempre con una sonrisa y a Bea, por no perder su simpatía a pesar de todas las tarjetas que he perdido en estos años.

A Sonia y a Jim, por haberme acogido en sus laboratorios y en sus familias y haberme hecho guardar tan buenos recuerdos de mis visitas a Grenoble y Vancouver. Y a Jona, por hacerme sentir tan bien en Hamburgo.

Especialmente le agradezco a Salva toda su paciencia y comprensión en tantos momentos. Por darme la oportunidad de trabajar a su lado, que es un privilegio, tanto a nivel científico como humano. Por todo el esfuerzo que ha puesto en esta tesis, y por decirme justo lo que debía en cada momento. Esta tesis no habría sido posible sin él.

Gracias a Agnès, por ser tan sabia persona y tan buena amiga, por haber estado tan presente todo este tiempo. Desde los primeros momentos de mi doctorado, gracias por tu paciencia, y por iluminarme en tantos momentos. Incluso estando tan lejos, siempre he sentido que estabas cerca.

A Jordi y a Alexis, entre otros millones de cosas, por estar conmigo el día de mi primera charla en una conferencia. A Melània, por repetirla conmigo tantas veces que se la acabó sabiendo de memoria.

To Phil, for being such a good friend. A Antonia, por su alegría, por darle perspectiva a la vida. A Nara, por su paciencia y su continua sonrisa, por estar siempre dispuesta a ayudarme. A mi Karla, por su lindo espíritu... no puedo nombraros a todos, pero os llevo conmigo.

A Sara, por su amistad y su lealtad. Por haber sido un apoyo tan importante durante estos años y haber compartido inquietudes y tantos buenos ratos. Por ser parte de mi familia, por haberme cuidado tan bien.

A Lauri y Carmen, por estar constantemente en mi vida, y por vuestra paciencia y cariño.

A mi familia, por su apoyo incondicional, por creer en mí. Por su amor. Esta tesis es para ellos.

Finalement, un très gross merci à Olivier, d'être à ma côté chaque jour, dans le bonnes et le mauvaises moments, en me soutenant pendant l'écriture de cette thése. Pour ses conseilles et ses mots, "allez, c'est la dernière ligne droite, courage!".

Todos vosotros habéis hecho que mí doctorado no solo haya sido eso, sino también un período maravilloso de mi vida.

Muchas gracias. Un beso.

Carol.

Abstract

The present dissertation addresses the predictive mechanisms operating online during audiovisual speech perception. The idea that prediction mechanisms operate during the perception of speech at several linguistic levels (i.e. syntactic, semantic, phonological….) has received increasing support in recent literature. Yet, most evidence concerns prediction phenomena within a single sensory modality, i.e., visual, or auditory. In this thesis, I explore if online prediction during speech perception can occur across sensory modalities. The results of this work provide evidence that visual articulatory information can be used to predict the subsequent auditory input during speech processing. In addition, evidence for cross-modal prediction was observed only in the observer's native language but not in unfamiliar languages. This led to the conclusion that well established phonological representations are paramount for online cross-modal prediction to take place. The last study of this thesis, using ERPs, revealed that visual articulatory information can have an influence beyond phonological stages. In particular, the visual saliency of word onsets has an influence at the stage of lexical selection, interacting with the semantic processes during sentence comprehension. By demonstrating the existence of online cross-modal predictive mechanisms based on articulatory visual information, our results shed new lights on how multisensory cues are used to speed up speech processing.

Resumen

El objetivo de esta tesis es investigar los mecanismos predictivos que operan de forma *online* durante la percepción audiovisual de una lengua. La idea de que existen mecanismos predictivos que actúan a distintos niveles lingüísticos (sintáctico, semántico, fonológico...) durante la percepción de una lengua ha sido ampliamente apoyada recientemente por literatura. Sin embargo, casi toda la literatura está relacionada con los fenómenos predictivos dentro de la misma modalidad sensorial (visual o auditiva). En esta tesis, investigamos si la predicción *online* durante la percepción del habla puede ocurrir a través de distintas modalidades sensoriales. Los resultados de este trabajo aportan evidencias de que la información visual articulatoria puede ser utilizada para predecir la subsiguiente información auditiva durante el procesamiento de una lengua. Además, los efectos de la predicción intermodal se observaron únicamente en la lengua nativa de los participantes pero no en una lengua con la que no estaban familiarizados. Esto nos lleva a concluir que representaciones fonológicas bien establecidas son esenciales para que ocurra una predicción *online* a través de modalidades. El último estudio de esta tesis reveló, mediante el uso de ERPs, que la información visual articulatoria puede ejercer una influencia más allá de las etapas fonológicas. En concreto, la saliencia visual de la primera sílaba de una palabra influye durante la etapa de selección léxica, interaccionando con los procesos semánticos durante la comprensión de frases. Los resultados obtenidos en esta tesis demuestran la existencia de

mecanismos predictivos a través de distintas modalidades sensoriales, basados en información articulatoria visual. Estos mecanismos actúan de forma *online*, haciendo uso de la información multisensorial disponible durante la percepción de una lengua, para optimizar su procesamiento.

Preface

Humans are extremely social beings, and us such, we are continuously interacting amongst each other, exchanging information about our needs, desires, opinions, thoughts, factual knowledge...

When involved in this kind of interactive process, the interlocutors share a common goal, which is to understand each other.

In fact, when having a conversation we have all, at some point, had the impression that we know what our partner is going to say before s/he even completes the word s/he is uttering. This happens because speech perception is not a passive process, but rather, we are engaged on it in an active manner. We use the information we are receiving at each moment and make predictions about what is coming next, constraining the interpretation of the message. These predictions allow for a faster and more effective interpretation of the incoming message.

Moreover, as spoken language has evolved as a communication system within the context of face-to-face communication, multiple cues from different sensory modalities (voice, face movements, lip movements, hand-gestures…) are potentially available to the perceiver. Therefore, in order to optimize the interaction, making the communicative process as fluid as possible, perceivers make use of all different kinds of information they are able to extract from the speaker.

With this in mind, the thought motivating this thesis is that as we have information from the speaker through different sensory modalities, it might be possible that information could be transferred from one modality to another to constrain the interpretation of the input. Along this line, it has been shown that, some information about the speaker's features (voice, face) can be transferred cross-modally, allowing for its identification.

Following the same logic it might be possible that a cross-modal transfer of information occurs online from one modality to another, as the perceiver is receiving speech. Furthermore, it might be possible that perceptual systems use this information as soon as it is available. This would allow the perceiver to make online predictions in order to anticipate information from a different sensory modality, which may naturally arrive later on, speeding up the processing of the message.

These cross-modal predictive mechanisms are the object of this thesis. In the first section, I will introduce some general aspects of audiovisual integration in speech and predictive mechanisms. I will describe previous related literature with the purpose of giving the reader the necessary context framework to understand the motivation of our studies. Afterwards, in the Experimental section, I will present three studies addressing different aspects of the cross-modal predictive mechanisms operating during speech perception. Finally, in the last section, I will relate our findings to previous literature and describe its implications. I will then conclude with some general remarks about the significance of the present findings in natural communicative situations.

Table of Contents

1. INTRODUCTION

1.1 Audiovisual speech perception as a natural case of Multisensory Integration

Our experience of the world is profoundly multisensory. In order to create a coherent and adaptive representation of the surrounding environment, our brain must integrate the information from different senses into a unique percept. Classically, each sensory modality - sight, sound, touch, smell, and taste - were thought to mediate perception by operating as independent and separated modules up until late stages in the information processing hierarchy. However, several recent studies have demonstrated that the different sensory modalities are in fact strongly interconnected and interact with each other earlier on in the chain of perception. Indeed, these studies argue that most, if not all, processes supporting perception are rather multisensory in nature (Stein & Meredith, 1993; Driver & Spence, 2000; Spence & Driver, 2004; Calvert, Spence, & Stein, 2004), occurring at many different levels of processing in the brain (Calvert & Thesen, 2005).

Speech processing is a compelling example of how multisensory integration is at the core of perception. Visual information on its own is often not enough for efficient speech recognition whereas auditory information is (i.e. we are perfectly able to maintain a conversation on the phone). However, under certain circumstances, when visual information accompanies the voice of a speaker, it leads to a substantial enhancement in the recognition and identification of speech (Grant & Seitz, 2000; Grant, 2001; Schwartz, Berthommmier, & Savariaux, 2004; Davis and Kim, 2006). Actually, as soon as the 1950's, Sumby & Pollack (1954) provided a classic demonstration on the importance of dynamic cues present in the visual speech signal, especially in noisy auditory environment. Notably, this is thought to occur because the speaker's visual articulatory gestures are in close relation with the movements of the vocal tract (Yehia, Rubin, & Vatikiotis-Bateson, 1998; Vatikiotis-Bateson & Yehia, 1996) and therefore, they correlate with the temporal envelope of the acoustic signal (Grant & Seitz, 2000; Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Moreover, in natural face-to-face communication, there is visual information not just from the speaker's lips, but from their head, jaw and eyebrow movements (Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Davis & Kim, 2006). In addition, there are often the speaker's hand gestures, which also contribute to speech processing (McNeill, 2005; Biau & Soto-Faraco, 2013).

1.1.1 Behavioural evidence for audiovisual integration in speech

It is relatively well agreed upon that humans are informational omnivores. In other words the perceiver's brain tends to make use of all the available sources of speech information, extracting as much cues as possible, in order to make communication more robust. This includes information sources of different sensory modalities, such as lip movements and speech sounds. An illustrative example of how visual information can enhance auditory speech perception relates to the famous Cocktail Party Effect (Cherry, 1953). This classical effect describes the ability to focus auditory attention on a single voice when embedded in a multi-speaker scenario, as in the case of speaking to someone in a crowded party. Different cues may facilitate this ability, such as voices coming from different directions, or acoustical differences between the voices (pitch, speed, gender, accent ...). Yet, in the original Cocktail Party Effect studies, researchers had not paid attention to the fact that the very example of a cocktail party usually occurs in a multisensory context. Therefore, in addition to the auditory input, visual information is available to the perceiver. Following up on this idea, a recent MEG study has shown that visual information from the speaker supports the focus of attention toward the relevant sound source in a cocktail party environment, facilitating the tracking of the auditory signal from the speaker within the auditory cortex (Zion Golumbic, Poeppel, & Schroeder, 2012).

The influence of visual information in face-to-face conversations, may become more evident in situations where the auditory information is strongly degraded by noise (Sumby & Pollack, 1954; MacLeod & Summerfield, 1987; Grant & Seitz, 2000; Kim & Davis, 2003; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Altieri & Townsend, 2011; Zion Golumbic et al.

2012), in hearing-impaired listeners (Grant, Walden, & Seitz, 1998; Rouger, Fraysse, Deguine, & Barone, 2007) and under circumstances when the audio signal is acoustically preserved but the message is difficult to understand. For instance, while interacting in a non-native language (e.g., Burnham, 1998; Navarra & Soto-Faraco, 2007) or when the message is semantically complex (Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987).

The enhancement of speech perception when visual information is present may be due, in part, to the complementary nature of the information provided by each of the two modalities (Summerfield, 1987). Whereas the most informative aspect from visual speech seems to be the place of articulation (e.g. if a phoneme is produced with the lips, frontal place of articulation, or at the back of the mouth, back place of articulation), the auditory modality transmits information about voicing and manner of articulation most efficiently (Smeele, 1994; Buschwald, Winters, & Pisoni, 2009; Jesse & Massaro, 2010; ten Oever, Sack, Wheat, Bien, & Van Atteveldt, 2013). The difference between the kind of information transmitted by both modalities makes it possible for visual cues to disambiguate two speech utterances which sound similar, but differ only in place of articulation (e.g. /ma/ and /na/).

Another interesting property of visual speech signals is that facial articulatory gestures can precede the corresponding auditory signal by as much as 100-300 ms (Grant & Seitz, 2000;

Chandrasekaran et al. 2009). Therefore, the sight of visual speech cues may support speech processing by predicting the auditory speech signal (Schwartz et al. 2004; Kim & Davis, 2003; Stekelenburg & Vroomen, 2007; Arnal, Morillon, Kell, & Giraud, 2009). More importantly, some of the information content carried by the visual speech can have an impact on acoustic processing (van Wassenhove, Grant, & Poeppel. 2005; Arnal et al. 2009; ten Oever et al. 2013).

Watching lip movements can have a dramatic impact on the perception of acoustically degraded signals and of perfectly audible speech. This latter point is well illustrated by the McGurk effect (McGurk & MacDonald, 1976), arguably one of the most notorious multisensory illusions. Furthermore, in the deaf population, visual speechreading may even be the sole basis of speech perception (Bernstein, Demorest, & Tucker, 2000), though only a few can master this ability and it varies greatly in both deaf and hearing populations (Auer & Bernstein, 2007; Bernstein et al. 2000; Mohammed et al., 2005). Finally, some studies have shown that visual speech can be enough to discriminate languages. As is the case of the perceiver who is familiar with at least one of the test languages (Soto-Faraco et al., 2007) and when s/he is bilingual in any other languages (Sebastián-Gallés, Albareda-Castellot, Weikum, & Werker, 2012), presenting this ability very early on in infancy (Weikum et al., 2007). This finding provides additional support for the fact that visual speech can be efficiently decoded to extract rich linguistic information.

1.1.2 Neural correlates of cross-modal audiovisual interactions

Many structures in the brain receive projections from more than one sensory modality. These regions include several structures in high-level associative or heteromodal cortices, such as the superior temporal sulcus (STS), intraparietal sulcus (IPS), inferior frontal gyrus (IFG), the insula and subcortical structures like the claustrum and the superior colliculus (SC). However, as for many other sensory integration processes, the specific neural substrates mediating audiovisual speech integration, underlying its behavioral benefits, remain somehow elusive.

The posterior superior temporal gyrus (pSTG) and STS have been pointed out as key regions during audiovisual integration of speech by a good deal of functional imaging studies. In addition to be responsive to auditory speech input (Scott, Blank, Rosen & Wise, 2000), the STS responds when presenting visible face movements, during speech reading and during audiovisual speech perception (Campbell et al., 2001; Calvert & Campbell, 2003, Callan et al., 2003; Miller & D'Esposito, 2005). This suggests that auditory and visual speech might interact at high levels of cortical processing (e.g., STS). However, it is less clear how early (or low) in the information processing hierarchy, multisensory interactions may occur.

Several lines of evidence suggest that audio-visual speech interactions may occur at the earliest functional-anatomic stages of cortical processing, such as unisensory auditory and visual cortices (Sams et al. 1991; Calvert et al. 1997; Möttönen, Krause, Tiippana, & Sams, 2002; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). For instance, using fMRI, Calvert et al. (1997) found that linguistic visual cues are sufficient to activate primary auditory cortex in normal hearing individuals in the absence of auditory speech sounds (see also Pekkola et al., 2005; but see Bernstein 2002). Moreover, congruent visual speech accompanying auditory input increases activity in the auditory cortex showing that the cross-modal activity modulation in the auditory cortex is not merely due to the imagery of speech sounds (Okada, Venezia, Matchin, Saberi & Hickok, 2013).

Taken together, these studies may indicate that after AV integration in higher order areas like STS, the auditory cortex is activated by visual speech through feedback connections from these areas (Calvert and Campbell, 2003).

However, this interpretation purely based on feedback is challenged by evidence showing that visual speech information can also converge as early as the primary auditory cortex in a direct (feed-forward) fashion. According to some authors, although audiovisual integration involves the STS, the actual integration process might not begin there. Indeed, electrophysiological data in the speech domain indicate that visual speech influences the early temporal stage of auditory processing (between 100–200 milliseconds after stimulation)

(Besle, Fort, Delpuech, & Giard, 2004; van Wassenhove et al. 2005). Van Wassenhove et al. (2005) focused on the timing of AV integration for both congruent (/ka/, /pa/ and /ta/) and incongruent (McGurk effect) speech syllables and found that the latencies of some components of the auditory evoked potential (i.e., N1/P2) were speeded-up when corresponding visual articulatory information was present. Interestingly, Van Wassenhove et al. also observed that the magnitude of this audiovisual temporal facilitation in ERPs correlated with the visual saliency of the syllables. Thus, findings showing reduced evoked responses mediated by saliency and redundancy during audiovisual speech interactions as early as the N1 stage, around 100 ms (Van Wassenhove, 2005; Arnal et al. 2009; Besle et al. 2004; 2008), together with the mounting evidence implicating the primary auditory cortex regions in visual speech processing (Besle et al. 2004; Pekkola et al. 2005; Van Wassenhove et al. 2005; Arnal et al. 2009; Okada et al. 2013) suggest that during audiovisual integration, an early representation of speech can be extracted while watching the lips of the speaker and used to speed up subsequent sound processing very early on in time.

In addition, during natural speech, facial movements sometimes precede the acoustic signal (Cathiard, Tiberghien, Tseva, Lallouache, & Escudier, 1991; Smeele, 1994; Grant & Seitz, 2000; van Wassenhove et al. 2005; Chandrasekaran et al., 2009) (but see Annex). It might be that earlier visual input supports speech perception by predicting when the auditory input will arrive (Stekelenburg & Vroomen, 2007; Schwartz et al., 2004; Kim & Davis, 2003). Accordingly, Schroeder et al.

(2008) argued for a modulatory effect of visual input on the auditory cortex, specifically by way of resetting the phase of neural firing so that the neural population would go through a high excitability phase, resulting in response amplification when the auditory signal arrives at an appropriate phase/time.

Taken together, electrophysiological studies showing audiovisual interactions early in time, support the notion that feedback from higher order areas such as STS seems to be complemented by that of direct modulation from other low level sensory areas (e.g., visual) to the auditory cortex. In fact, recent studies in primates have suggested that multisensory integration might occur at every level of cortical processing (Ghazanfar & Schroeder, 2006; Kayser, Petkov, Remedios, & Logothetis, 2012). According with this idea, recent studies using brain imaging have brought further support for cross-modal interactions at a variety of levels of representation.

Arnal et al. (2009) proposed two different functional pathways mediating audiovisual speech processing. One is a fast, direct cortico-cortical pathway from visual (motion-sensitive cortex) to the auditory cortex. Arnal et al. showed that functional connectivity measures of these two regions were positively correlated with the visual predictability of the input, but independent on the incongruency of the AV stimuli. According to Arnal et al. this might indicate that the direct pathway conveys mostly information about the amount and timing of the facial movements. The other pathway is an indirect, slower pathway, in which STS receives converging input from visual and auditory modalities, involving a fine-tuned phonological comparison between both signals, and sends feedback to auditory and visual cortices. When AV input does not match it might involve several visual cortex/STS/auditory cortex comparison loops. Arnal, Wyart, & Giraud (2011) more recently reported, in a MEG study, changes in oscillatory brain activity supporting the two different pathways. When audiovisual events were congruent, predictions were mediated by slow oscillations in high-order speech areas, whereas in the case of incongruent visual and auditory inputs, they were mediated by high-frequency oscillations in early unisensory cortex and the STS.

Adding to this picture, other findings have demonstrated that activity in motor areas is modulated during speech perception, either while listening to speech (Fadiga, Craighero, Buccino, & Rizzolati, 2002; Wilson, Pinar Saygin, Sereno, & Iacoboni, 2004) or when watching speech-related lip-movements (Watkins, Strafella, & Paus, 2003; Skipper, Nusbaum, & Small, 2005). For instance, activation of the left pre-motor cortex (PMC) (Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007) seems to be somehow mediated by Broca's area (posterior part of the left IFG), tightly functionally connected with the PMC (Watkins & Paus, 2004). These pre-motor areas have been shown to form a network with the STS, supporting phonemic categorization after the acoustic analysis of the speech sound is done in the STS (Wilson & Iacoboni, 2006; Iacoboni, 2008). The role of the motor system during speech perception will be address in greater detail below (**Section 1.2.2**).

1.2 Theoretical approaches to audiovisual speech perception

Some models of speech perception have considered speech as a unimodal process, taking into account uniquely the acoustic input (e.g. Klatt, 1980; McClelland & Elmand, 1986; Hickok, Holt, & Lotto, 2009; Holt, & Lotto, 2008*)*. However, some theoretical developments addressing speech perception as an audiovisual process were fuelled by the discovery of the McGurk effect (McGurk & MacDonald, 1976). These models consider the visual input as an informative source, which integrates at some point with the acoustic input during the perception process. As Quentin Summerfield put it: *"any comprehensive account of how speech is perceived should encompass audio-visual speech perception"* (Summerfield, 1987).

The Visual Place Auditory Manner account (VPAM; MacDonald & McGurk, 1978) stated that place of articulation is mostly identified using the visual signal whilst manner of articulation is transmitted by the auditory modality. Following this hypothesis, audiovisual integration occurs only after both signals are evaluated independently in a parallel fashion. The biggest problem of the VPAM view is that the dichotomy about the features transmitted by one of the modalities is not as absolute as the theory assumes. For this reason, it has been argued that taking into account the information from both visual and auditory modalities for each phonetic feature seems more suitable (Robert-Ribes, Piquemal, Schwartz & Escudier, 1996). Opposite to the VPAM, Summerfield (1987) proposed an early audiovisual integration hypothesis according to which integration would take place before phonetic categorization occurs (see Schwartz, Robert-Ribes, & Escudier, 1998). The main proposal of this hypothesis becomes the basis for later theories arguing for early integration. Below, I will describe in detail some of the recently more influencial models accounting for the multimodal nature of communicative processes.

1.2.1 The FLMP

The Fuzzy-Logical Model of Perception (FMLP, Massaro, 1987; 1998) is a relevant example of late integration models. It describes the process of speech perception in three sequential but temporally overlapping stages: evaluation, integration and decision. These stages make use of prototypes stored in long-term memory; each prototype is defined by a set of features, which are available from the auditory and visual sources during speech. These prototypes are established in memory after repeated experience with visual and acoustic speech. During the first stage, *evaluation*, specific speech features are extracted from the incoming visual and acoustic signals and independently evaluated in a parallel manner, against the pre-existing prototypes, providing information about to which degree each feature of the auditory and visual signal matches to the corresponding feature value of the prototype. In the stage of *integration*, the degrees of match from each feature of a prototype are combined, following a Bayesian analysis, to provide a degree of support for each alternative. In the final *decision* stage, the total support value for each prototype is evaluated against the sum of the support for all relevant alternatives, obtaining a "relative goodness of match", that reflects the degree to which a stimulus matches the category of a given prototype. See Figure 1 below.

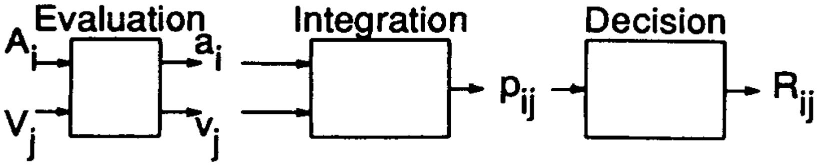


Figure 1. The fuzzy logical model of perception (FLMP). In the figure, auditory information is represented by Ai and visual information by Vj (uppercase letters). The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters ai and vj). These sources are then integrated to give an overall degree of support for a given alternative pij. The decision operation maps the outputs of integration into some response, such as a discrete function or a rating, Rij. Adapted from Massaro and Cohen, 1993.

The FLMP is a late integration model, because auditory and visual features are only integrated in the second stage, namely "*integration stage*", following independent processing (evaluation stage) based on each modality separately. In contrast to the predictions of this model, there is empirical evidence for audiovisual fusion at a pre-phonological level. For instance, the presence of visual cues correlated to auditory speech enhances the detection (Kim & Davis, 2003; Bernstein, Auer, & Takayanagi, 2004) and identification (Schwartz et al. 2004) of speech in noise. Moreover, the rate of visual speech affects the classification of a heard token (Green & Miler, 1985) and visual coarticulation affects auditory categorical perception (Green, 1998). These evidences are incompatible with a late audiovisual integration. See also Schwartz, Robert-Ribes, & Escudier (1998) and Altieri & Townsend (2011) for a more recent discussion.

Even if the FLMP is maybe the model that describes with more detail the possible mechanisms involved in audiovisual integration of speech, it makes little reference to any visual parameter, considering enough the acoustic changes during speech perception. Since visual cues are an integral part of speech, alternative theories include the motor system in the process of speech perception, assuming the primitives of speech as gestural in nature. These theories are discussed below.

1.2.2 Motor theories of speech perception and the perception and production relationship

The "Motor Theory of Speech Perception (MTSP)" (Liberman & Mattingley, 1985) proposed that speech is represented as the gestures of the articulators from the speaker, and is decoded by recruiting the perceiver's motor system. This theory claims that the translation of speech sounds into a phonological code is not necessary, because the articulatory gestures are phonological in nature. This claim is shared by the Direct-realistic-theory or the ecological approach to speech perception (Fowler & Rosenblum, 1989; 1991), but unlike the motor theory, they rely on purely perceptual processes during the recognition of articulatory speech gestures.

According to the MTSP, speech perception implies the activation of motor commands in the perceiver's brain that would be involved in producing the gestures observed in the speaker (Stevens & Halle, 1967; Liberman & Mattingly, 1985). According to several current views, this internal speech simulation might even allow the listener to anticipate some aspects of the spoken message (Watkins & Paus, 2004; Pickering & Garrod, 2006; 2013; Skipper et al. 2005, Skipper, van Wassenhove, Nusbaum, & Small, 2007).

In line with the basis of the MTSP, a theoretical model of AV speech perception based on the analysis by synthesis approach (Stevens & Halle, 1967) has been proposed (van Wassenhove et al., 2005; Skipper et al., 2007; Pickering & Garrod, 2006; 2013). According to this model, while watching someone speaking, visual and auditory cues are integrated as early representations about what the speaker is producing. These early perceptual representations are mapped onto motor commands, involved in speech production, that send an efferent copy (forward model) to the sensory cortices about what the consequences of articulating these phonemes would be, based on prior knowledge in speech production. This forward model is then compared with the actual speech input and adjusted online in order to reduce the disparity between predicted and actual input (i.e., error signal), thereby constraining the interpretation of the spoken signal (Skipper et al, 2005; 2007; Pickering & Garrod, 2006; 2013; See Figure 2 below). This model assumes the existence of a close link between perception and production systems, as originally suggested in the MTSP (Liberman & Mattingley, 1985).

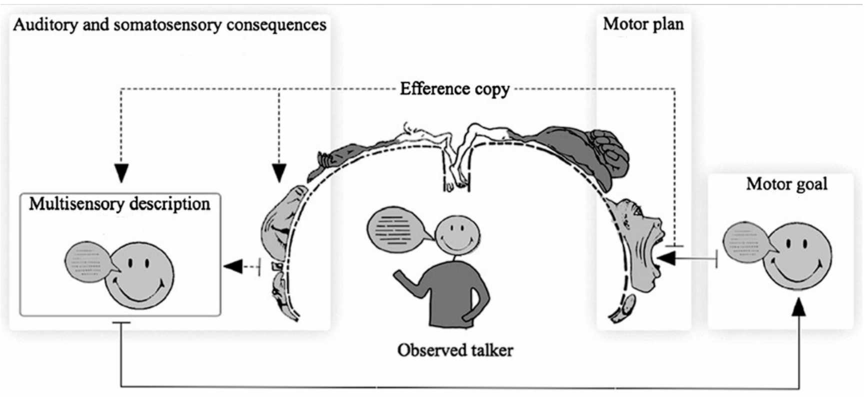


Figure 2. Model of AV speech perception in a predictive coding framework. A multisensory description in the form of a hypothesis about the observed talker's mouth movements and speech sounds results in the specification (solid lines) of the motor goals of that hypothesis. These motor goals are mapped to a motor plan that can be used to reach that goal, resulting in the prediction through efference copy (dashed lines) of the auditory and somatosensory states associated with executing those motor commands. Auditory and somatosensory predictions are compared with the current description of the sensory state of the listener. The result is an improvement in speech perception in AV contexts due to a reduction in ambiguity of the intended message of the observed talker. Adapted from Skipper et al., 2007.

The interest on the perception-production link was revived by the discovery, early in the 90's, of the "mirror neurons" in the pre-motor cortex of the macaque. These neurons respond (engage in action vigorous discharge) when the animal is producing a particular goal-directed action, as well as when that animal is observing someone else performing that same kind of action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). More recent, indirect evidence, pointing to the existence of a similar mirror system in humans (e.g., Rizzolatti & Craighero, 2004), has revived the hypothesis that the motor structures can be involved in perceptual processes, setting the possible physiological basis for motor models of speech perception.

In the last decade, the results from TMS experiments have supported the predictions from these models. For instance, some studies have shown that during speech listening, motor-evoked potentials recorded from the listeners' muscles increase when the subject perceives words whose pronunciation involves this articulator, i.e. tongue movements (Fadiga et al., 2002) or lip-movements (Watkins et al., 2003). In addition, in an fMRI study, Pulvermuller (2006) showed that the areas in the motor cortex representing the lips and the tongue are differentially activated by the perception of sounds articulated using the lips and the tongue, respectively. All these evidences might indicate that the increased motor excitability may reflect an internal simulation of the perceived speech, implicating the same articulators as if the perceiver's would produce it.

Brain motor areas have also been shown to be involved in phoneme categorization processes. For example, in an fMRI study, Wilson & Iacoboni (2006) showed that the activation of the pre-motor cortex was higher while listening to non-native than to native phonemes. Wilson & Iacoboni interpreted that during non-native phonemes perception, the motor system attempt to model different phonemes, because the perceived one is not known, resulting in a higher motor activation. A direct influence of the motor cortex in speech perception has been shown during discrimination of specific phonemes. Meister et al. (2007) managed to disrupt, with repetitive TMS over the left pre-motor cortex, participants' performance on a phonetic discrimination task. Similarly, Mötönnen & Watkins (2009) used rTMS over the lip representations in the left M1 disrupting the ability to discriminate sounds lip-articulated but not sounds not articulated by the lips (/ba/ vs. /da/ and /pa/ vs. /ta/). Using a different paradigm, D'Ausilio et al. (2009) used TMS to enhance activity in the areas of the primary motor cortex corresponding to lips or tongue while participants performed a phoneme discrimination task of sounds whose production involved one of those articulators. D'Ausilio showed that when TMS was applied to the area in the motor cortex representing the lips, the discrimination of sounds whose production involved the lips (/b/ or /p/) was faster than sounds produced with the tongue (/d/ or /t/) while the opposite was true when stimulating the tongue cortex.

Despite these evidences, it has been argued that the motor system is not necessary for speech perception per se, but rather it is recruited only in the case listening conditions are compromised (McGettigan, Agnew & Scott, 2010; Lotto, Hickok, & Holt, 2009). Supporting this view, normal auditory perception has been shown in patients with strongly impaired speech production abilities (see Hickok, Costanzo, Capasso, & Miceli, 2011; Bishop, Brown & Robson, 1990) and in aphasic patients with fronto-parietal lesions (Rogalsky, Love, Driscoll, Anderson & Hickok, 2011). However, a recent developmental study (Yeung & Werker, 2013) shows that perception in infants as little as 4.5 month-old is already affected by the production system. This might leave open the possibility that the motor system plays a major role during the acquisition of speech (perception, and obviously, production).

The concrete role of the motor system in audiovisual speech perception is, just like its role in speech perception in general, still controversial. A possibility might be that visual phonemic information modulates activity in motor areas during speech perception (Skipper et al. 2005), and these areas might contribute to the integration of the auditory and visual speech signals (Callan et al. 2003; 2004; Jaaskelainen, 2010; Ojanen et al. 2005). In fact, the motor system is not only activated when listening to speech (Fadiga et al. 2002; Wilson et al. 2004) (but see Lotto et al., 2009 for a discussion) but also while viewing the articulatory gestures associated with speech (Campbell et al. 2001; Murakami, Restle, & Ziemann, 2011; Watkins et al. 2003;

Skipper et al. 2005). Furthermore, a modulation of the activity of the cerebellum and cortical motor areas has been observed depending of the visual saliency of the perceived phonemes (Skipper et al. 2005; 2007), (but see Wilson et al. 2004; Wilson & Iacoboni, 2006). Another possibility is that the motor system might play a role in predictive coding, by helping generate forward models about the sensory consequences of the production of the phonemes one is hearing (Wilson & Iacoboni, 2006) or lip-reading (Möttönen, Järveläinen, Sams & Hari, 2005; Sams, Möttönen, & Sihvonen, 2005; van Wassenhove et al., 2005; Skipper et al. 2007; Kauramäki et al. 2010).

1.3 Audiovisual speech perception and predictive mechanisms

As it has been discussed above, the excitability of the motor system can be modulated during speech perception, supporting the possibility of an internal representation of the perceived speech by the production system. This internal representation might improve the listener's ability to understand and even anticipate the heard speech. In the next section we introduce the concept of predictive coding and then present some evidences of predictive mechanisms operating during perception.

1.3.1 Predictive coding

In 1890, William James used the concept "*pre-perception*" to refer to the sensory anticipation of an event. The concept of "*pre-perception*" reflected the pre-activation of relevant brain structures in situations of high expectation of a particular event, which then reduced the need for very elaborate processing following the actual appearance of the event. Nowadays, it is increasingly clear that perception is not a passive, receptive, process in which information from (possibly a variety of) sensory systems is integrated in a bottom-up fashion to create a final percept. In recent years the idea of perception as an active process has been applied in many different perceptual domains.

Throughout our life, due to our interactive experience with the world, we acquire knowledge about the events surrounding us. It is thought that this past experience helps to actively form expectations in an attempt to anticipate information during perception. Therefore, if the incoming information matches the expectations, our perceptual system will be ready to integrate the new events immediately. On the contrary, events that break expectations call for a re-evaluation of the expectations, and take longer to integrate. A good illustration of how anticipatory processes are present in everyday life is the one used by Enns & Lleras (2008): "*Imagine you are in your office and you hear familiar footsteps outside. By the time the person knocks to your door and you see her, if it was who you were expecting, you will start immediately a conversation. However, if the person in front of you* *is not who you were expecting, you will be surprised, and your visual identification will be momentarily frozen because of the mismatch between your expectation and the visual information you are receiving*". As illustrated by the example, our knowledge about the world makes us perceive any event in a subjective way, as a weighted combination of prior knowledge and current sensory information. Predictions very often lead to an advantage when one faces a new situation, because they allow to constraint interpretations, leading to a faster recognition/reaction to the upcoming events, sparing resources and time.

The mechanisms through which these expectations are created incorporate predictions supported by recurrent neural processing (Friston, Kilner, & Harrison, 2006; Bar, 2007). In this sense, we could speak not only about a "proactive" (Bar, 2007) but also about a "predictive brain" (see Bubic, von Cramon, & Schubotz, 2010 for a review). Predictive models were first described with regard to sensory-motor integration (Wolpert, 1997) but later on these models have been applied to a wide variety of perceptive systems. Amongst others, visual processing (Bar et al., 2006; Bar, 2007; Enns & Lleras, 2008), music (Keller & Koch, 2008), emotional processing (Gilbert & Wilson, 2009) and language (DeLong, Urbach, & Kutas, 2005; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Skipper et al. 2007; Pickering & Garrod, 2006). The general idea underlying these models is that sensory information in the brain flows in a forward fashion that, at one or more stages, is compared with top–down "predictions" or "internal models" endogenously generated based on previous experience, projected back from higher levels of information processing. These feedback predictions help reducing ambiguity among potential interpretations of sensory input, refining and enhancing perception.

1.3.2 Predictive coding in speech perception

In real-time language processing, anticipation is present too. In the sentence "*Yesterday, I went to the library to borrow a... "*, the reader most likely may be anticipating the word *"…book"* even before it actually appears. For instance, when people stutter, listeners often feel the urge to help finishing the sentence, because they already know what their partner wants to say. The same reasoning can be applied when, while having a conversation, sometimes we know what our partner wants to say even before he finishes the sentence (see Clark & Wilkes-Gibbs, 1986).

As in the library example above, prediction can be strongly grounded on high level aspects of speech such as the paralinguistic context, semantics or syntax. However, the predicted information itself can be rather detailed in terms of the phonological or even acoustic features of the expected word form. Following up on with the library example, we do not only expect to hear the word "*book"*, but also the phoneme /b/, and perhaps even pronounced in the particular accent or pitch of the voice of the person we are talking with). Experimentally, ERP studies have shown the consequences of prediction of upcoming words while listening to a sentence (Van Berkum et al. 2005) or when reading it (DeLong et al. 2005; Dambacher, Rolfs, Gollner, Kliegl, & Jacobs, 2009), in sufficiently constraining natural discourses. For example, DeLong et al. took advantage from the fact that in English language, the realization of the indefinite article changes depending on the phonological context, so that 'an' precedes nouns beginning with vowel sounds, whereas 'a' precedes nouns beginning with consonant sounds. DeLong created sentences where a particular article + noun ending was highly expected, such as "The boy went out to fly …"), and tested the ERP response to indefinite articles which were congruent / incongruent with the most expected (yet not presented) noun. So, in the example, the sentence could continue with the article 'a' (as in "…a kite", the most likely continuation) or with 'an' (as in "… an airplane", a semantically acceptable but unlikely continuation). De Long et al. observed a strong N400 component modulation by the article, so that the unlikely article produced the largest N400. This was interpreted as evidence of violation of the expectations from online predictions during visual (written) word recognition. Furthermore, these predictions seemed to express at the phonological level, because the grammatical, syntactic and semantic aspects of the two possible realizations of the indefinite article were, otherwise, equivalent. Moreover, ERPs differences between predictable and not predictable words while reading a sentence arise no later than 90 ms after stimulus onset (Dambacher et al. 2009), meaning that the comparison between top-down predictions and stimulus-driven bottom-up processes occurs very early in time.

In the framework of analysis-by-synthesis based models (Stevens & Halle, 1967) such as the predictive coding one (Skipper et al. 2005; 2007; van Wassenhove et al. 2005), the perceiver extracts cues from speech and use them to make predictions about the subsequent information in an online manner. According to this framework, this process is mediated, at least in part, by the cortical machinery that has been developed through experience producing speech (i.e. motor system). Prior visual and auditory information is constantly used to interpret the subsequent speech signal while watching or listening to someone speaking (speaker's lip-movements, semantic context* , prosody*...). The consequence is that conversation becomes more fluid (Pickering & Garrod, 2006; Skipper et al. 2007; see Scott, McGettigan, & Eisner (2009), for a role of the motor system in the timing of turn taking during a conversation).

* See Glossary

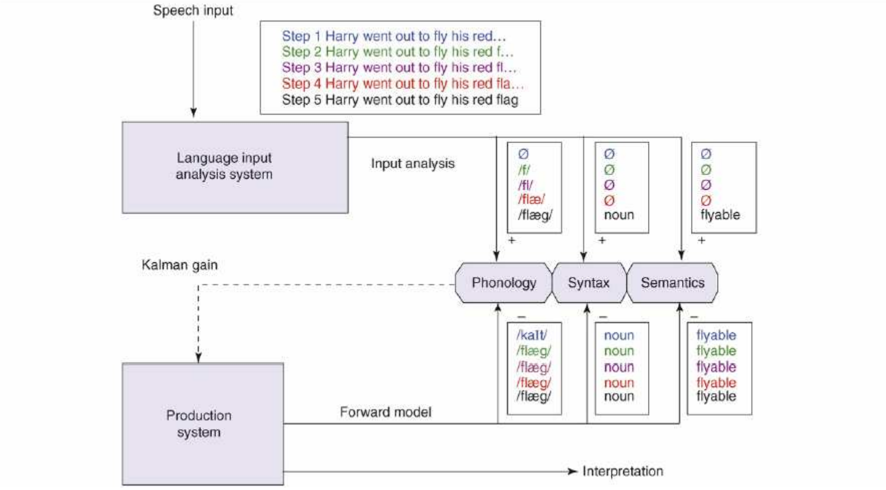

Speech input
Step 1 Harry went out to fly his red...
Step 2 Harry went out to fly his red f...
Step 3 Harry went out to fly his red fl...
Step 4 Harry went out to fly his red fla...
Step 5 Harry went out to fly his red flag
Language input analysis system
Input analysis
Ø
/f/
/fl/
/flæ/
/flæg/
Ø
Ø
Ø
Ø
noun
Ø
Ø
Ø
Ø
flyable
+
+
+
Kalman gain
Phonology
Syntax
Semantics
−
−
−
/kalt/
/flæg/
/flæg/
/flæg/
/flæg/
noun
noun
noun
noun
noun
flyable
flyable
flyable
flyable
flyable
Production system
Forward model
Interpretation

Figure 3. Schematic representation of a forward model of speech perception. At each moment, predictions are weighted against analysis of the input at each step. If the prediction is strong and the input noisy, there is a strong top-down influence on interpretation); if the prediction is poor and the input clear, there is strong bottom-up influence. Illustrated are the five steps in comprehending the end of the sentence 'Harry went out to fly his red flag.' At each step, the input analysis and the forward prediction are shown in the same colour for three different levels of prediction (phonology, syntax and semantics). Adapted from Pickering and Garrod, 2006.

A main objective of this thesis is to test the possibility that predictive coding might operate not only using information within the same sensory modality, but whether it does so across sensory modalities too. We explore the possibility that cross-modal prediction might operate as an online mechanism mediating the audiovisual enhancement during speech perception. If such online cross-modal prediction exists, then one would expect that it is based on the possibility of sharing information across modalities. For this reason, demonstrations of cross-modal transfer are relevant. In the next section I will describe some evidences about the transfer of information between visual and auditory modalities during speech, before presenting a brief summary of the scope of the studies included in this thesis.

1.3.3 Cross-modal transfer of information during speech perception

Due to the mechanics of speech production, when looking at a speaker's face, what we see is the direct reflect from his/her vocal tract. Thus, there is a strong correspondence between a wide range of acoustic parameters and visible movements from the articulatory apparatus (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulus, 1996; Yehia et al., 1998; Chandrasekaran et al. 2009) as well as from head movements (Yehia et al. 2002; Munhall et al. 2004; Davis & Kim, 2006). This implies that some linguistic and paralinguistic cues can be available from both modalities, supporting the basis for a transfer of information. Perceivers are sensitive to this cross-modal relationship, and make the most of it by extracting information from lip's to head movements during the perception of speech (Summerfield, 1987; Benoît & LeGoff, 1998). This leads to a perceptual benefit, reflected in an enhancement in recognition and identification of speech (Grant & Seitz, 2000; Grant, 2001; Schwartz et al. 2004; Davis & Kim, 2006) when auditory input is accompanied with the view of the speaker, as described in Section 1.1 of this Introduction. In addition, when perceiving a non-native language, observation of the speaker's articulatory movements improves learning of non-native phonetic contrasts (Hardison, 2005; Hazan, Sennema, Iba & Faulkner, 2005; Hirata & Kelly, 2010) and facilitates phonetic discrimination of phonemic contrasts which non-native speakers would not be able to make based on auditory cues alone (Navarra & Soto-Faraco, 2007).

Face and voice not only produce correlated information about what is being said (the message), but they also seem to inform about who is saying it (the talker). Indeed, a few studies have shown that some kind of amodal speech information is shared by visual and auditory modalities, allowing the identification (or matching) of talkers across modalities. Experience with a speaker in one modality (seeing the face speaking or listening to his/her speech) allows listeners to match his/her identity thereafter when exposed to information to the opposite modality (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003) even in an unknown language (Lander, Hill, Kamachi, & Vatikiotis-Bateson, 2007). The information transferred seems to be based on general spatio-temporal features that are specific of an individual's speech (Kamachi et al. 2003; Munhall & Buchan, 2004; Lander et al. 2007).

Interestingly, exposure to visual-only speech information from a given speaker, improves speech recognition of subsequent auditory-only speech in noise from that person (Rosenblum, Miller, & Sánchez, 2007). It seems that amodal information extracted from the articulatory gestures is used to facilitate phonetic* recovery of the same talker's speech (Rosenblum, et al. 2007; Kamachi et al. 2003). Phonetic amodal information seems therefore to be transferred cross-modally, allowing for the recognition of the speaker (Kamachi et al. 2003; Lander et al. 2007) and recovering speech from the same speaker across modalities (Rosenblum et al. 2007). More specifically, the articulatory style particular from a talker, named *idiolect**, seems to underlies some cross-modal correspondences in these studies.

Speech alignment is defined as the tendency of speakers to converge at a myriad of speech levels such as phonological, lexical or syntactic while having a conversation (Pickering & Garrod, 2004). Supporting the idea of cross-modal transfer of idiolectic information, Miller, Sánchez, & Rosenblum, (2010) showed that speech alignment can be based not only on auditory speech but also on visual speech, and transferred across modalities. Miller et al. compared participant's shadowed words (i.e. repeated after a speaker) when lip-read from the speaker or when heard. The shadowed words were rated by independent judges in comparison with the original speaker pronunciation, and with the same words when read aloud by the participants, instead of shadowed. The subject's utterances* turned to be rated as more similar to the speaker's utterances than they were to their own read, non-shadowed, utterances. Moreover, an alignment across modalities was perceived by the judges as they

*See Glossary

were able to match a subject's voice to the visible articulating face of the speaker they shadowed.

In sum, according to the studies discussed above (Kamachi et al. 2003; Lander et al. 2007; Rosenblum et al. 2007; Miller et al. 2010) the amodal or modality-neutral information shared by visual and auditory sources (see Rosenblum, 2005) may be at the basis of the relationship between perception and production systems, as supported by the close link between voice and face resulting from the mechanics of speech production (as it is the surface of the vocal tract) (Yehia et al. 1998; Vatikiotis-Bateson & Yehia, 1996). This may also forms a basis for cross-modality correlations. The relationship between perception-production might in fact support cross-modal predictive mechanisms operating online during speech perception.

Supported by all the findings described, the main aim of this thesis is to explore the possibility of online predictive mechanisms acting cross-modally during speech perception. This aim is described in more detail in the next section.

1.4 Scope of this thesis

Many previous studies support the idea that predictive mechanisms operate during perception and in the particular case of perceiving speech, these mechanisms allow perceivers to anticipate information within the same sensory modality (DeLong et al. 2005; Van Berkum et al. 2005; Dambacher et al. 2009). In addition, some studies support the existence of cross-modal transfer of information in an off-line manner (Rosenblum et al. 2007; Kamachi et al. 2003; Lander et al. 2007). Speech is audiovisual in nature and the addition of visual cues speeds up speech processing (van Wassenhove et al. 2005). Since visible articulations are sometimes temporally advanced to its acoustic correlates (Chandrasekaran et al. 2009), predictive mechanisms might take advantage of this fact, anticipating subsequent information, and operating at different stages online during speech processing. Presumably, this might depend on how informative visual information is (i.e. degree of saliency). Moreover, as perception and production are closely linked, one possibility to implement this predictive function would be that anticipated visual information facilitates the simulation of its auditory correlates even before acoustic information arrives. In such a case, prediction would be most effective when the perceiver knows the consequences of the articulatory gestures (as for known language but not for unknown ones).

1.4.1 Goals and hypothesis

The first objective of this thesis is to find out whether predictive mechanisms might operate online during speech perception in a cross-modal fashion, mediating the audiovisual enhancement observed when visual information accompanies auditory speech. A related question is whether this possible cross-modal prediction is a two-way mechanism (i.e. operating from visual-to-auditory and from auditory-to-visual modality) or else there is an asymmetry. We hypothesize that predictive mechanisms might operate cross-modally, but in a unidirectional manner, from visual-to-auditory modality, taking advantage of the natural visual anticipation with respect to its acoustic correlates.

The second goal of this thesis is to address the nature of the putative cross-modal predictive mechanism. We hypothesize that given the close relationship between production and perception, prediction is likely to operate at a phonological level. In this case, one could think that the effects of prediction will be strongest when talker and perceiver share a common phonological code. This is based on the fact that the listeners' simulations will be most efficient when s/he can produce the phonemes being heard, thus being able to simulate the visual articulatory gestures. The other possibility is that prediction might operate at a pre-phonological level, taking advantage of general predictive mechanisms based on visual information acting as a timing cue predicting the arrival of auditory input.

Finally, we aim to address how prediction based on the visual modality (varying in saliency across different phonemes) interacts with predictive mechanisms at higher processing levels, based on the semantic context, during sentence processing. This study constitutes a novelty with respect to previous ERPs studies, which typically have looked either at the effect of expectations created from the semantic context of sentences or to predictions at the phonological level based on the visual saliency of the phonemes using syllables, independently. We think that looking at a possible interaction between both levels of prediction is interesting because in natural speech, syllables and words are usually embedded in sentences, giving rise to the possibility of an interaction between various levels of prediction (i.e. here, phonological and semantic). We looked at this possibility in the third study of this thesis, in which we measured ERPs during sentence presentation.

In the next section of this thesis, I describe the experimental work carried out in order to test the hypotheses stated above. These experiments are presented in the form of three published papers.

2. EXPERIMENTAL SECTION

2.1 Overview of the experiments

The **Experimental section** of this thesis consists of three research articles published in international journals representative of this area of research. Each article addresses a different but related issue, connected with the goals and hypothesis described at the end of the introductory chapter. Below, I will briefly introduce each of them.

As discussed in the **Introduction**, predictive mechanisms operate in a wide variety of perceptual domains, constraining the interpretation of the incoming input. As a multisensory phenomenon, speech perception is enhanced when visual information accompanies, and correlates with, auditory speech. I contend that this enhancement may be partly mediated by predictive mechanisms. Thus, a first objective on this thesis is to address if predictive mechanisms might operate during online speech perception in a cross-modal fashion. Supporting this possibility, cross-modal transfer of information has been shown to facilitate speaker (Kamachi et al. 2003; Lander et al. 2007) and speech recognition (Rosenblum et al. 2007). In the first study presented in **Section 2.2 (Sánchez-García et al., 2011)**, we explore online predictive mechanisms within and across modalities during speech processing. In this study, aside of intramodal prediction (from vision-to-vision and audition-to-audition), we found online cross-modal prediction operating uniquely in one way, from visual-to-auditory modality. But, what is the nature of this cross-modal predictive mechanism? Some important features of the predictive mechanisms are addressed in **Section 2.3 (Sánchez-García et al., 2013),** by investigating the level of processing at which it may operate**.** On one hand, the presence of visual information might speed up speech processing because it might act as a temporal cue, priming* the subsequent auditory input (Schwartz et al. 2004; Stekelenburg & Vroomen, 2007; Schroeder et al. 2008). This possibility is supported by the natural temporal anticipation of visual cues associated to auditory information during speech (Grant & Seitz, 2000; Chandrasekaran et al. 2009). On the other hand, higher-level information (i.e. phonological) might be extracted from the speaker's articulations, and used to anticipate its consequences in the auditory modality, enhancing speech perception at a higher level of processing (van Wassenhove et al. 2005; Arnal et al. 2009). In fact, the amount of content in information carried by the visual speech has been shown to bias the final percept (ten Oever et al. 2013), depending on the visual saliency of the presented phoneme. To disentangle between these two possibilities, in the study presented in Section 2.3., we compared native and non-native speech. Native phonological categories are well established during the first months of life and difficult to change thereafter, affecting the perception of non-native languages (Werker & Tees, 1984; 2002; Sebastián-Galles & Soto-Faraco, 1999; Navarra et al. 2005). Therefore, the logic of this comparison is that if cross-modal prediction makes use of the general spatio-temporal cues shared by the visual and the auditory modalities (pre-phonological level), the predictive effects might be observed in any language, regardless of the previous knowledge about it. However, if pre-phonological information is not enough, but rather, online predictive mechanisms rely on higher level information, such as phonological, prediction effects will be observed only (or most efficiently) if perceiver and speaker share a common phonological code. Visual predictive cues will be informative not only about when auditory information is arriving but also about the articulatory consequences of the seen gesture. We addressed the role of these visual predictive cues during sentence processing, looking at the possibility that visual phonemic cues might interact with expectations created from the semantic context in Section 2.4 (Brunellière et al. 2013). Early benefits at pre-lexical level* based on the amount of content in information carried by the visual speech (i.e. saliency of the presented phoneme) have been shown by previous ERP studies. This effect is reflected as an amplitude reduction and temporal facilitation of the auditory N100 component, when visual information is simultaneously presented with the corresponding auditory syllables (Besle et al. 2004; Stekelenburg & Vroomen, 2007; van Wassenhove et al. 2005). The temporal facilitation correlated with the degree of visual saliency of the presented syllables (van Wassenhove et al. 2005; Arnal et al. 2009). In Section 2.4 (Brunellière et al. 2013) we partially replicated these results. When looking at word level, a visually highly salient phoneme* at word onset can start lexical processes, possibly taking advantage of its earlier temporal arrival (Fort et al. 2012). But in natural speech, words are usually not isolated, but embedded in contextual sentences with a semantic meaning. At sentence level, larger amplitude of the N400 component reflects the violation of expectations about a word, created from the previous semantic context of the sentence. Therefore, in natural speech contexts, semantic information carried by the context, brings into play a higher level of prediction (De Long et al. 2005; Van Berkum et al. 2005; Dambacher et al. 2009) during speech processing, which might interact with a lower level of prediction carried by the phonemes' visual saliency (van Wassenhove et al. 2005; ten Oever et al. 2013). To the best of our knowledge, however, the role of visual speech in sentential context has never been explored by previous electrophysiological studies. In the study described in **Section 2.4 (Brunellière et al. 2013)** we look at this possibility by using event relate potentials (ERPs) in order to address possible interactions between predictions from visual articulatory information (naturally in advance of its acoustic correlates, and more or less informative depending on its degree of visual saliency) and predictions based on the semantic context.

2.2 Sánchez-García, C., Alsius, A., Enns, J. T., and Soto-Faraco, S.

Cross-modal prediction in speech perception. PLoS One. 2011 Oct 05;6(10): e25198 DOI 10.1371/journal.pone.0025198

Cross-Modal Prediction in Speech Perception

Carolina Sánchez-García[1], Agnès Alsius[2], James T. Enns[3], Salvador Soto-Faraco[1,4]*

[1] Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain, [2] Department of Psychology, Queen's University, Kingston, Canada, [3] Department of Psychology, University of British Columbia, Vancouver, Canada, [4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract

Speech perception often benefits from vision of the speaker's lip movements when they are available. One potential mechanism underlying this reported gain in perception arising from audio-visual integration is on-line prediction. In this study we address whether the preceding speech context in a single modality can improve audiovisual processing and whether this improvement is based on on-line information-transfer across sensory modalities. In the experiments presented here, during each trial, a speech fragment (context) presented in a single sensory modality (voice or lips) was immediately continued by an audiovisual target fragment. Participants made speeded judgments about whether voice and lips were in agreement in the target fragment. The leading single sensory context and the subsequent audiovisual target fragment could be continuous in either one modality only, both (context in one modality continues into both modalities in the target fragment) or neither modalities (i.e., discontinuous). The results showed quicker audiovisual matching responses when context was continuous with the target within either the visual or auditory channel (Experiment 1). Critically, prior visual context also provided an advantage when it was cross-modally continuous (with the auditory channel in the target), but auditory to visual cross-modal continuity resulted in no advantage (Experiment 2). This suggests that visual speech information can provide an on-line benefit for processing the upcoming auditory input through the use of predictive mechanisms. We hypothesize that this benefit is expressed at an early level of speech analysis.

Citation: Sánchez-García C, Alsius A, Enns JT, Soto-Faraco S (2011) Cross-Modal Prediction in Speech Perception. PLoS ONE 6(10): e25198. doi:10.1371/journal.pone.0025198

Editor: Stefan J. Kiebel, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

Received April 4, 2011; **Accepted** August 30, 2011; **Published** October 5, 2011

Funding: This research was supported by the Spanish Ministry of Science and Innovation (PSI2010-15426 and Consolider INGENIO CSD2007-00012), Comissionat per a Universitats i Recerca del DIUE-Generalitat de Catalunya (SRG2009-092 and PIV2009-00122), and the European Research Council (StG-2010 263145). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Salvador.Soto@icrea.es

Introduction

Perceptual information from different sensory systems is often combined to achieve a robust representation of events in the external world [1]. Research during the past two decades has documented numerous instances of multisensory interactions at neuronal and behavioral levels (see [2]). These interactions are demonstrated, for example, in the McGurk effect, such that listening to the spoken syllable /ba/ while simultaneously watching the lip movements corresponding to the syllable /ga/ often results in the illusory perception of /da/ [3]. When visual and acoustic speech signals are correlated, the benefits of multisensory integration in speech perception are also well documented (e.g., [4], [5]). This multisensory advantage is strongest at moderate to high acoustic noise levels [5], [6], when the message is semantically complex [6], [7], or when it involves processing second language sounds [8]. However, the mechanisms that enable this cross-modal benefit are still not well understood.

We hypothesize that one mechanism that could potentially contribute to multisensory speech enhancement is that of predictive coding, operating both within each sensory modality and possibly even between modalities. The principle of predictive coding has been successfully applied, with some variations, to explain information processing in many domains (e.g., [9], [10], [11], [12]), including motor control [13], object identification [14], shape perception [15], [16], music perception [17], visual masking [18], visual search [19], visual spatial orienting [20], [21], and speech perception [22], [23], [24], [25], [26]. What all these proposals have in common is the idea that information in the brain not only flows forward through a hierarchy of processing levels, but that at some stage/s of processing it is also met by a top-down 'prediction', projected back from higher levels in the functional hierarchy. These feedback predictions help to reduce ambiguity among potential interpretations of sensory input and to provide finer spatial and temporal parsing of the incoming signals.

In the case of speech, there are several levels of linguistic analysis where on-line predictions might contribute to parse the signal, including phonology, lexical access, syntactic parsing, and semantics. For instance, when listening to a sentence like "I went to a library and borrowed a …", the expectation to hear "book" is strongly driven by a semantic prior context, but it is likely to constrain lower levels of input analysis including that of phonology and the lexicon (i.e., a strong expectation to hear the phoneme /b/, from the word *book*). Supporting evidence for this has been reported in spoken [27] and written language perception [28], [29]. In Van Berkum's as in DeLong's study, increases in the amplitude of the N400 ERP component were evoked by words that were grammatically incongruent with the most likely continuation in a contextually biasing sentence, even though the remainder of the sentence was never presented. For example, in DeLong et al., the sentence fragment (i.e., "The boy went out to fly …") could continue with the article *a* (as in "…a kite", the most likely continuation) or with *an* (as in "… an airplane", an unlikely continuation). The finding that the unlikely article produced the largest N400 effect was interpreted as evidence for on-line predictions guiding visual (written) word recognition. Furthermore, these predictions seemed to express at the phonological level, because the grammatical, syntactic and semantic aspects of the two possible realizations of the indefinite article were, otherwise, equivalent.

An important question that still remains unexplored is whether the predictions made during speech perception can cross from one sensory modality to the other. If so, such predictions may occur at phonological or even pre-phonological levels of processing. For instance, phonology has been proposed as a common representational code for various aspects of speech perception (visual and acoustic) as well as production [22], [24], [30], [31], [32], [33]. Some evidence for a link between auditory and visual speech representations comes from Rosenblum et al. [31], who exposed participants, previously inexperienced in lip reading, to silent video-clips of an actor producing speech. In a subsequent task, the same participants performed auditory word recognition in noise, being more accurate when the words were spoken by the same speaker they had previously experienced visually (but not heard). Another example of cross-modal transfer in speech comes from Kamachi et al. [33], who reported that people are able to match the identity of speakers across face and voice (i.e. cross-modally), according to the authors based on the link between perception and production of speech.

Results such as these demonstrate the potential for cross modal transfer of information in speech perception. The basis for such transfer during off-line tasks could be phonological or pre-phonological, given the putative relation at these early representation levels between speech perception and production. However, what has not been established to date is a clear demonstration that such transfer is possible in an on-line task, akin the type of processing engaged during normal speech perception. Some hints about this possibility do, however, exist. For example, indirect support for on-line transfer can be drawn from the finding that facial articulatory movements typically precede (and strongly correlate with) the corresponding acoustic signal. The lead time of the facial movements over the corresponding sound is on the order of a few tenths to a few hundredths of milliseconds (e.g., [34]). Further indirect support comes from Van Wassenhove et al. [26], who reported a significant speed up of the ERP components N1 and P2 when they were evoked by audiovisual syllable presentations as compared to audio presentations alone. Interestingly, the size of this latency shift in the auditory evoked components was proportional to the visual saliency of the phoneme, but no correlate of a behavioral benefit was tested. These cross-modal effects on ERP latency, may not necessarily be based on speech-specific mechanisms, as shown by Stekelenburg and Vroomen [35], but abide to a more general mechanism from which speech processing can capitalize.

The present study was conducted in an effort to test for possible on-line cross-modal benefits during speech perception. In Experiment 1 we began by asking whether performance in an audiovisual matching task would benefit from prior unimodal contextual information (speech fragment in one sensory modality) that was continuous with one of the channels in the audiovisual target clip. As indicated in Figure 1, participants made speeded responses during the presentation of the target clip, to whether or not the speaker's face talked in agreement with the concurrent auditory stream. The critical manipulation was whether a preceding unimodal sentence context (auditory or visual) was continuous with the target clip or whether no such context was provided. When we found that the context provided a benefit in this task, we were ready, in Experiment 2, to compare the benefits of a sentence context that was continuous within a single sensory channel to a context that was continuous across sensory channels. This manipulation allowed us to directly compare potential benefits of on-line predictions unimodally and cross-modally, again testing in both directions, from vision to audition and vice-versa.

Results

Experiment 1: Benefits of prior visual and auditory information

We included four types of trials, depending on the information content of the context (unimodal speech or no speech) and the matching nature of the target (audiovisual matching or mismatching). In this experiment, when available, the context was always continuous with the corresponding modality channel in the target fragment. In the auditory version of the experiment, the informative context was auditory, and in the visual version the context was given visually alone. In both cases, the context in the baseline trials (no informative context) contained no speech information. Figure 2 shows the mean correct response times in Experiment 1. In both the visual and the auditory versions, participants detected audiovisual mismatch in the target more rapidly following a leading informative context than no context. This supports the hypothesis that on-line speech perception benefits from advance information in both the visual and auditory modality.

An ANOVA of correct RTs (filtered 2SDs above and below the mean for each participant and condition) indicated faster responses following leading informative context as compared to no context (visual: $F(1,15) = 10.42$, $p<0.05$; auditory: $F(1,17) = 13.8$, $p<0.05$) and faster responses to matching audiovisual targets than to mismatching ones (visual: $F(1,15) = 114.5$, $p<0.05$; auditory: $F(1,17) = 368.9$, $p<0.05$). In general, participants were always faster responding after a visual leading context than after an auditory context ($t(32) = 2.33$; $p<0.05$). A significant interaction between presence of informative context and target congruency (visual: $F(1,15) = 17.8$, $p<0.05$; auditory: $F(1,17) = 9.25$, $p<0.05$), reflected that the benefit of context was significant for mismatch trials (visual: $t(15) = 5.60$, $p<0.05$; auditory: $t(17) = 4.33$, $p<0.05$), but not for match trials (visual: $t(15) = 0.44$, $p = 0.66$; auditory: $t(17) = 1.18$; $p = 0.25$).

Accuracy was high overall (visual = 88%, auditory = 90%), and did not reflect speed-accuracy trade-offs. We analysed the signal detection parameter d' (hits = match responses on matching trials; false alarms = match responses on mismatching trials) and the *criterion*, C, as a measure for response bias. In the auditory version, d' was higher in presence of leading context ($d' = 2.99$ vs. 2.64; $t(17) = 3.28$; $p<0.05$), in keeping with the RT pattern. No differences in sensitivity were found in the visual version ($d' = 2.57$ vs. 2.67; $t(15) = 0.91$; $p = 0.37$). In terms of criterion, both the auditory and the visual versions revealed a stronger bias towards a matching response in the informative context condition as compared to the no context one (auditory, $C = -0.38$ vs. -0.20, $t(17) = -3.76$; $p<0.05$; visual, $C = -0.37$ vs. -0.05, $t(15) = -4.95$; $p<0.01$).

Experiment 1 provided evidence that audiovisual processing can benefit from information present a few hundred milliseconds earlier in either a visual or an auditory channel. This can reflect the consequences of forming on-line predictions in a cross-modal speech perception task. However, from this result alone one cannot tell whether the leading channel benefits the perception of

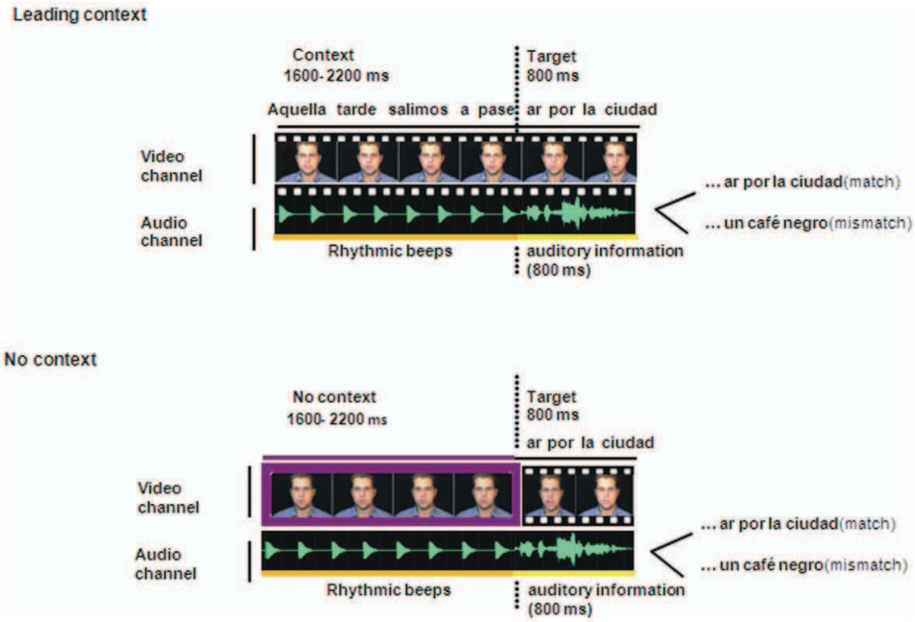

Leading context
Context
1600- 2200 ms
Target
800 ms
Aquella tarde salimos a pase ar por la ciudad
Video channel
Audio channel
... ar por la ciudad(match)
... un café negro(mismatch)
Rhythmic beeps
auditory information (800 ms)
No context
No context
1600- 2200 ms
Target
800 ms
ar por la ciudad
Video channel
Audio channel
... ar por la ciudad(match)
... un café negro(mismatch)
Rhythmic beeps
auditory information (800 ms)

Figure 1. Illustration of the stimulus sequences in Experiment 1. In the example is shown the visual version of the experiment. For the leading context condition, a video clip of the moving lips of the speaker, presented in conjunction with rhythmic beeps, preceded the combined audio and visual target of the sentence. In the no context condition, the leading context consisted of the still video frame of the speaker and rhythmic beeps. In the auditory version (not shown here), the context in the leading context condition consisted of a still video frame and the original audio channel of the spoken sentence. The no context condition was exactly the same to the one shown in the figure for the visual version. English translation of the sentences: That afternoon we went out to walk… around the town/ a black coffee.
doi:10.1371/journal.pone.0025198.g001

subsequent speech in the same sensory modality as the leading context, or whether the information in the leading channel can be used to constrain processing in the other sensory modality as well. Experiment 2 was designed to isolate potential cross-modal effects.

Experiment 2: Cross-modal vs. intra-modal predictions

This experiment also had visual and auditory versions, each including three main types of trials (see Figure 3). *Intra-modal continuous* (akin to the informative context condition of Experiment

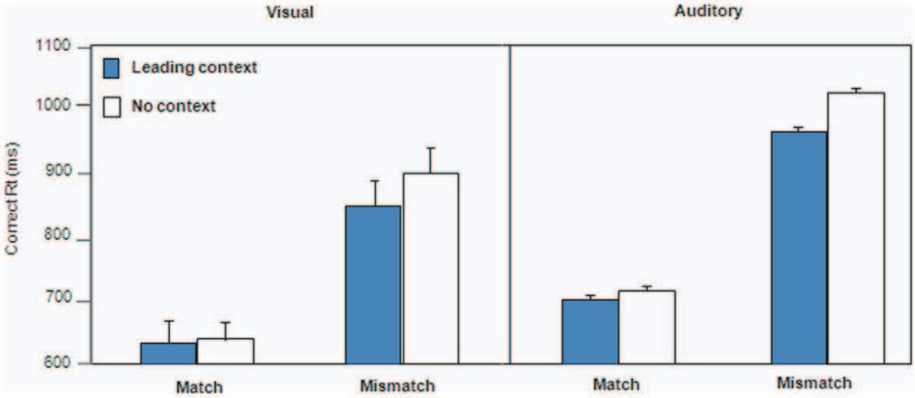


Figure 2. Mean correct RT (in milliseconds) in Experiment 1. Visual (left panel) and auditory (right panel) versions. Error bars represent one standard error of the mean.
doi:10.1371/journal.pone.0025198.g002

1, where the context continued onto the same sensory modality in the target); *cross-modal continuous* (where the context fragment was continuous only with the opposite sensory modality in the target), and *discontinuous* (where there was no continuity from context to target). In this experiment all trials contained speech information in the context. The intra-modal continuous and the discontinuous trials could have audio-visually matching or mismatching target fragments, but the cross-modal continuation could only have mismatching targets (as a necessary design limitation, see the Methods section for details). Thus, the critical conditions in Experiment 2 for testing prediction across modalities involved the three comparable types of mismatching trials, as illustrated in Figure 3. It is critical to note that the comparison of greatest interest in this experiment is between the discontinuous and the cross-modal continuous conditions, both of which involve an identical video splice (or audio splice) between context and target fragments. Because the discontinuity from context to target portions of the sentences is identical in these cases, it cannot lead to differences in attentional capture at the splice point.

Figure 4 shows the mean correct response times in Experiment 2. In the visual version (left side), participants were able to detect audiovisual matches more rapidly following a continuous versus a discontinuous leading context. They were also able to detect mismatches more rapidly following both an intra-modal and a cross-modal continuation, as compared to the discontinuous condition. The auditory version (right side) revealed the same pattern of results, with one exception. Although the data showed an advantage for intra-modal continuity over discontinuity on matching trials and mismatching trials, there was no evidence of a benefit when the continuity was cross-modal.

An ANOVA including the factors of context continuity (intra-modal continuous vs. discontinuous) and target congruence (match vs. mismatch), revealed faster responses when the context was continuous intramodally than discontinuous (visual: F(1,15) = 15.3, $p<0.05$; auditory: F(1,15) = 26.99, $p<0.05$), and when the target fragment was matching rather than mismatching (visual: F(1,15) = 186.16; $p<0.05$; auditory: F(1,15) = 115.63; $p<0.05$). This result supports the within modality continuous context advantage found in Experiment 1, with a different baseline (discontinuous context, rather than no context). The interaction between context and congruence was not significant in the visual version, F(1,15) = 2.38, p = 0.14, but it was in the auditory version, F(1,15) = 26.19, $p<0.05$.

A second ANOVA included all three types of context continuity (but only mismatching trials, given the design constrain discussed in the Methods section). This was the critical analysis to test the hypothesis of cross-modal prediction. The main effect of type of context was significant in the visual version, F(2,30) = 7.72, $p<0.01$, but not in the auditory version, F(2,30) = 0.412, p = 0.66. Follow-up tests in the visual version showed that RTs in both the intra- and cross-modal continuation conditions were faster than the discontinuous condition (t(15) = 3.24, p = 0.05; t(15) = 3.26, $p<0.05$, respectively), and not different from one another, t(15) = 0.83, p = 0.41. Equivalent tests in the auditory version failed to reach significance, all $|t|<1$. Overall, participants were slightly faster responding after a visual leading context than after an auditory context, but the difference was not significant (t(30) = 1.38; p = 0.17).

Like in Experiment 1, response accuracy was high (visual = 90%, auditory = 84%). In the visual version, intra-modal continuation performance ($d' = 2.84$) was superior to that of discontinuous ($d' = 2.60$), (t(15) = 2.76, $p<0.05$), and there were no significant differences between cross-modal continuation ($d' = 2.58$) and discontinuous, t(15) = 0.24, p = 0.81. In the auditory version, there were no significant differences, intra-modal continuous, $d' = 1.88$; discontinuous, $d' = 1.92$ and cross-modal, $d' = 1.86$, all $|t|<1$. The criterion was not significantly different from zero in any of the two versions (visual version: intra-modal continuous, $C = 0.03$, t(15) = 0.63, p = 0.53); cross-modal continuous, $C = -0.10$, t(15) = −1.26, p = 0.22; discontinuous, $C = -0.09$, t(15) = −1.28, p = 0.21. Auditory version: intra-modal continuous, $C = 0.02$, t(15) = 0.35, p = 0.73; cross-modal continuous, $C = 0.01$, t(15) = 0.118, p = 0.90; discontinuous, $C = 0.04$, t(15) = 0.54, p = 0.59)), indicating the absence of bias towards any kind of response.

Discussion

This study offers behavioral evidence that listeners can use speech information on-line to constrain the interpretation of the subsequent signal within and across sensory modalities, thereby benefiting performance in an audiovisual speech matching task. When the leading context fragment (auditory or visual) was continuous within the same modality in the audiovisual target fragment, there was a reduction in response time for the detection of audiovisual mismatch (Experiments 1 and 2). However, when the context and target fragments were continuous across different modalities, only visual continuity into auditory channel (but not the reverse) produced a benefit. We interpret these results as indicating that at least under some conditions, immediately preceding speech context can be used to form predictions about the upcoming input, facilitating the detection of a mismatch between audio and visual channels. And in the case of visual to auditory transfer, the information can even be transferred within the time limits of the modality switch.

These results can be readily interpreted within a predictive coding framework. In these models, speech information at various levels of processing (i.e., semantic, syntactic, phonological) is extracted from the signal and used to activate hypotheses at levels above (feedforward processing) and below (feedback processing). Such an arrangement allows the system to constantly generate probabilistic hypotheses about the upcoming signal to constrain the interpretation of the incoming input on-line.

Unlike the visual context, the auditory context fragment was clearly comprehensible for the observers. Thus, the beneficial effect of the auditory context during Experiment 1 may not be too surprising, as it allows for the possibility of predictions to be formed at higher levels (semantic, syntactic) as well as lower ones (phonological, articulatory). As such, the benefit of context in the auditory version is consistent with previous ERP evidences for auditory-based predictions being used on-line in the comprehension of spoken language [27]. It may be also related to previous demonstrations of on-line predictions being used in the comprehension of written language (e.g., [28], [29]).

However, to our knowledge, this study provides the first demonstration that prior visual speech-reading information can be used to benefit speech processing in a similar way. One important difference, however, is that the visual speech signal provided very little information to our participants, who are not trained lip-readers, at the levels of syntax and semantics [4], [36]. Therefore, we believe that in the audiovisual matching task used in our experiments, the phonological or pre-phonological levels are the most likely used for cross-modal transfer from vision to audition. For instance, phonology is claimed to be amongst the earliest representational levels at which auditory and visual aspects of speech can be encoded in a common format (e.g., [30], [31]). As reviewed in the Introduction section, phonology is likely the level where facial articulatory movements correspond most closely to

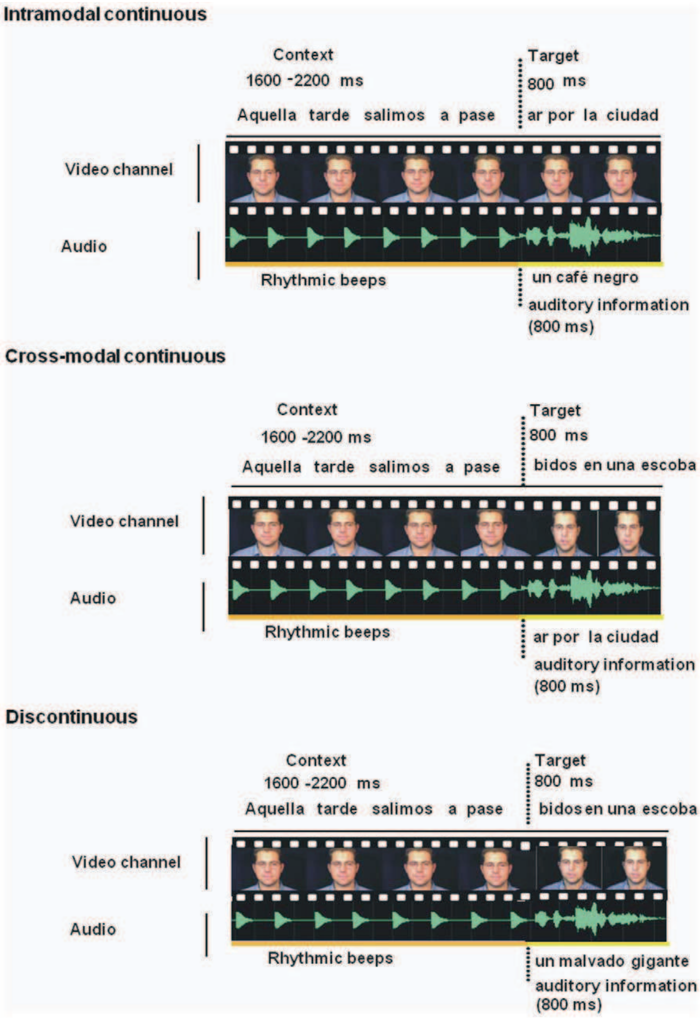


Figure 3. Illustration of mismatch stimulus sequences for the visual version of Experiment 2. In the example shown, for the intra-modal continuous mismatch condition, the lip movements of the context and target fragment were a continuation of the same sentence, but there was no prior information in the auditory channel (rhythmic beeps). In the cross-modal continuous mismatch condition, the lip movements of the context were continuous with the auditory channel of the target fragment. Finally, in the discontinuous mismatch condition, the lip movements of the context and target fragment corresponded to a different sentence. English translation of the sentences: That afternoon we went out to walk… around the town/ a black coffee/ riding a broomstick/ a wicked giant.

doi:10.1371/journal.pone.0025198.g003

acoustic signals, perhaps based on the link they both are supposed to have with the articulatory representations used in speech production [22], [24], [30], [31], [32].

To support this interpretation, we estimated the amount of semantic and syntactic information that could be extracted from the visual leading context in our stimuli. In order to do it, we tested twelve new participants with thirty-nine of the sentences used in Experiments 1 and 2, presented only visually. Participants were asked to report, after watching each sentence, the words that they had been able to recognize. We scored the proportion of content words correctly reported (i.e., nouns, verbs and adjectives but not functional words such as articles or prepositions). The mean percentage of correctly reported words was 3.2%, which supports our claim that information at lexical or higher levels could be hardly extracted from the visual context. It is more likely that the information extracted and used in cross-modal transfer is of a pre-lexical nature (phonological, pre-phonological or perhaps even prosodic) rather than semantic.

The distinction between the possible role of phonological and pre-phonological levels in our results is, at this point, difficult. Some theories of audiovisual fusion claim for the existence of a common format at an early, pre-phonological level of represen-

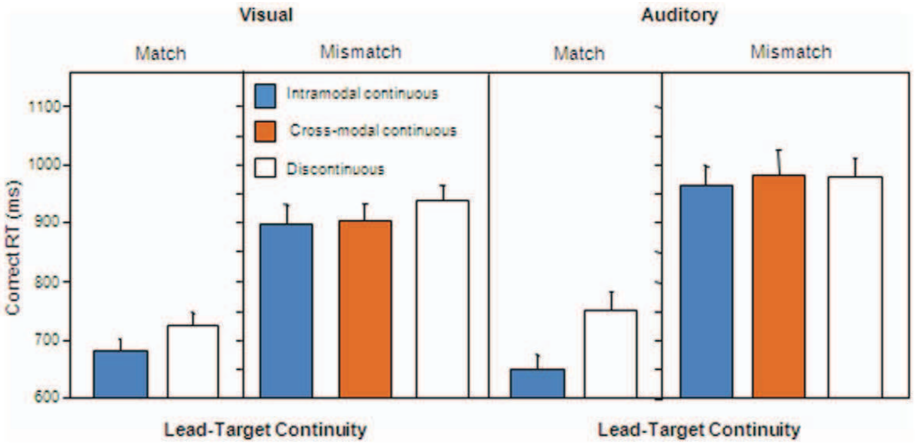


Figure 4. Mean correct RT (in milliseconds) in Experiment 2. Visual (left panel) and auditory (right panel) versions. Error bars represent one standard error of the mean.
doi:10.1371/journal.pone.0025198.g004

tation [37], [38]. We cannot rule out or confirm the possibility that the prediction effects will be based on such levels of representation with our current evidence. A potential way to address the role of phonological vs. pre-phonological representations would be to test for prediction effects in an unknown language. If prediction effects equivalent to those seen here happen at a phonological level rather than in a pre-phonological one, then some minimal degree of phonological knowledge about the language will be necessary for cross-modal transfer to occur.

Our data imply that visual speech information can be used to constrain processing of subsequent auditory information, through a real-time intra-modal transfer as well as a cross-modal transfer of information. This cross-modal benefit is, however, unidirectional from visual to auditory, but not vice-versa. Why the cross-modal transfer was asymmetric, showing benefits of leading visual information on audition, but not the reverse? Our interpretation is that this is consistent with bio-mechanical constraints on language production, whereby the visual information available to an observer precedes in time the corresponding acoustic information [34]. It also fits well with previous ERP findings in which auditory evoked potentials occur earlier when correlated visual information is present [26], [35]. However, an alternative explanation for the present asymmetry in cross-modal effects is that speech comprehension based on the visual channel alone is so much more difficult than when based on the auditory channel alone. As such, the visual leading context may prompt participants to try to actively simulate the sounds based on the facial gestures. In contrast, merely listening to an auditory leading context would not prompt the same degree of active involvement in the task, given that comprehension is easy. To test this hypothesis we conducted a control experiment, identical to the auditory version of Experiment 2, with the exception that a simultaneous noise mask was added to the auditory channel (Signal to Noise Ratio = −5 dB) in order to render it barely intelligible. Despite the increased effort now required to understand the auditory channel, the correct RT data replicated the main result of the auditory version in Experiment 2 (RTs = 1156.76 ms vs. 1143.78 ms; t(19) = 1.22, p = 0.23). This result rules out the *difficulty* hypothesis, although it must be admitted that the asymmetry in our results could be due to strategic differences resulting from extended experience with audio emulation from lip-reading but not visual emulation from audio perception, making the cross-modal transfer more likely from vision to audio than in the opposite direction. Our data does not allow us to resolve this question at present.

Interestingly, in Experiment 1 (auditory and visual version) the benefits of prediction tended to be larger when the task demanded the detection of audiovisual mismatch rather than a match, whereas matching trials showed a benefit of continuity only in Experiment 2. This is in accord with a recent suggestion of an important processing difference on audiovisual match versus mismatch signals [39]. Arnal et al. proposed that when sensory modalities match, they engage preferentially direct connections between visual and auditory areas. In contrast, mismatching information across modalities engages a slower, more indirect network, whereby visual input is integrated and compared with the auditory input via association areas (i.e., the Superior Temporal Sulcus, STS). As such, the process of detecting match in the present study may have occurred too rapidly to be indexed by our response time measure in Experiment 1. The quicker responses to matching trials, as compared to mismatching ones, together with the significant bias to respond 'match' in several of the conditions tested in Experiment 1 (informative context (visual version), $C = -0.04$, t(15) = 0.62, p = 0.54; $C = -0.19$, t(17) = −3.87, $p<0.01$(auditory); no context, $C = -0.37$, t(15) = −5.35, $p<0.01$(visual); $C = -0.38$, t(17) = −5.61, $p<0.01$(auditory)), may reflect a strategy in which participants would default to a matching response a priori. From this perspective, checking for disconfirmation (mismatching responses), would take longer than checking for a confirmation (matching responses). The significant bias toward matching responses in Experiment 1 would support this hypothesis. Note, however, that in Experiment 2, precisely where the on-line cross-modal transfer was shown, there were no significant criterion shifts. Therefore, this particular strategy cannot be the only cause of the RT pattern reported here.

The neural mechanism that mediates this improvement of audiovisual processing following a visual context still remains unknown. We could speculate about the involvement of the mirror neuron system, as suggested by some authors. According to the model proposed by Skipper et al. [24], for example, while perceiving visual information, the motor system is engaged in comparing a hypothesis based on previous experience (forward model) and the perceived information. This makes possible to speed up processing about incoming information that matches expectations.

In conclusion, the present study documents an important case of on-line cross-modal transfer of information in speech perception. Specifically, it demonstrates that visual speech signal in a sentence can facilitate the quick extraction of sufficient information for the detection of a match or mismatch in a subsequent audiovisual portion of the sentence. Our results support that on-line speech perception benefits from a leading visual information, that can be used both to constrain the interpretation of subsequent visual (intra-modal) and auditory (cross-modal) processing. In the case of leading auditory information, the benefit occurs only within the same sensory modality. These results may reflect the well known precedence of visual to acoustic consequences of articulation. We contend that this predictive ability may play a facilitatory role in everyday communication, enabling phonological predictions, based on visual cues, of what we are about to hear.

Methods

Experiment 1: Benefits of prior visual and auditory information

Participants. Data from 34 native Spanish speakers (10 males, mean age 23.4 years) were included in Experiment 1. Data from eight participants who failed to meet a performance criterion of 65% accuracy in the audiovisual matching task were not included, so that their data did not alter our conclusions. All participants reported normal audition and normal or corrected-to-normal vision, and were naive to the purpose of the experiment. The protocol was run under the approval of the University of Barcelona ethics committee, and all participants gave written informed consent. Sixteen participants were assigned to the visual leading context version; 18 to the auditory leading context version.

Materials and procedure. The stimuli consisted of high resolution audiovisual recordings of a male speaking fifty-two complete sentences in Spanish, as indicated in the Appendix S1. Each sentence was edited with Adobe Premiere Pro 1.5, to last 2400, 2600, 2800 and 3000 milliseconds, and included a 560 ms linear fade-in ramp and a 360 ms linear fade-out ramp. Participants viewed the video recordings from a distance of 60 cm on a 17″ CRT computer monitor that showed the full face of the speaker face in the center of the screen. The audio channel was played through two loudspeakers located on each side of the monitor, at a comfortable listening intensity of 65 dB SPL. A program using DMDX software [40] was used to organize the randomization, presentation and timing of the experiments.

Temporal uncertainty was created by sampling randomly and equiprobably among four leading context durations (1600, 1800, 2000 and 2200 ms), prior to the presentation of the 800 ms target fragment. Trials began with a central fixation circle (0.8° visual angle, 500 ms), followed by the presentation of a sentence context (1600–2200 ms) plus target (800 ms). Following each response or time-out (1800 ms deadline) the screen blanked for 800 ms before the next trial began. To confer ecological validity to our design, we left at random the level of discriminability of the particular articulatory gesture in which each of the sentences change form context to target. We just avoided that the transition would occur during a speech (silent) pause in the sentence.

Participants judged, as quickly and accurately as possible, whether the target fragment of the sentences had matching or mismatching audiovisual channels. Responses were made with the index and middle fingers on two neighboring keys, with the assignment of finger to response counterbalanced across subjects. The target fragment consisted of the final 800 ms of each sentence, and it always included both audio and visual channels. To create mismatching targets from these recordings, the audio (or visual, depending on the version) channel of the original fragment was randomly replaced with that of another sentence.

In order to test the effect that both modalities could have over the audiovisual matching task, we ran two different versions of Experiment 1. In one version, we presented an auditory leading context, and in the other version, we presented a visual leading context. In each of the two versions, there were four different types of trials, formed from the orthogonal combination of whether the leading context was a sentence fragment or not (leading context, no context) and whether the audiovisual channels in the target fragment were matching or mismatching. The leading context was always either the original audio or the original visual fragment of the sentence that preceded the target fragment, and thus it continued from the context through the target fragment. The channel that was not informative during this unimodal leading context was replaced. The replacement of the auditory channel was a sequence of rhythmic beats (300 Hz tones, 120 ms duration each, presented at 5 Hz, as shown in Figure 1), that was comparable to the rhythm of speech, and the visual channel was replaced with a still face of the speaker. For the no context conditions, used as the baseline, a still frame of the speaker's face was combined with rhythmic beats. It is important to note that the leading context manipulation (present or absent) was orthogonal with respect to the task and response set, which was whether the audiovisual channels were matching or non-matching. Each participant responded to a total of 208 trials in either the visual or the auditory version, with each of the 52 original sentences edited to create the 2×2 design: leading context vs. no context, and matching vs. mismatching target. Only in two of the four times that each sentence was presented to each participant, it was shown on its complete form, including context, making any possibility of learning very unlikely. These sentences were sampled randomly without replacement for each participant, with context duration varying randomly and equiprobably amongst the four possible durations (1600 to 2200). Participants practiced on a subset of 20 sentences prior to testing. Each experimental session lasted approximately 30 min.

Experiment 2: Cross-modal vs. intra-modal predictions

Participants. A different group of participants, formed by 32 native Spanish speakers (10 male, mean age 23.1 years) participated in Experiment 2. Data from 17 additional participants who failed to meet the 65% performance criterion were not included, so that their data did not alter our conclusions. Sixteen participants were assigned to the visual leading context version; 16 to the auditory leading context version.

Materials and procedure. Forty audiovisual sentences similar to those used in Experiment 1 were selected. As in Experiment 1, we created two versions of the experiment, one to test for visual-to-auditory prediction (called *visual version* for simplicity) and one to test for auditory-to-visual prediction (called *auditory version*). As in Experiment 1, participants judged if the target fragment was audio-visually matching or mismatching.

The critical comparisons in this experiment involved the three audiovisual mismatching target conditions illustrated in Figure 3. The condition called *intra-modal continuous* was identical to the context condition of Experiment 1, in that the context channel was continuous with the same channel in the target fragment. In the new condition called *cross-modal continuous*, the leading context channel was continuous with the alternative modality channel in the target fragment. Finally, the *discontinuous* condition served as a comparison for both of these continuous conditions, in that it required the same response (a mismatch judgment), but the leading context provided no information about the message in the target clip (since it belonged to a different sentence).

Each participant was tested in a total of 200 trials, distributed in 5 equivalent blocks of 40 trials in which each trial type was equiprobable. Only in the two continuous conditions participants were presented with the complete form of the sentences, to avoid any possibility of learning. The experimental session lasted about 30 min.

Supporting Information

Appendix S1 Spanish sentences and their English translation. (DOC)

Author Contributions

Conceived and designed the experiments: CS-G AA JE SS-F. Performed the experiments: CS-G. Analyzed the data: CS-G SS-F. Contributed reagents/materials/analysis tools: CS-G AA JE SS-F. Wrote the paper: CS-G AA JE SS-F.

References

1. Stein BE, Meredith M (1993) The merging of the senses. Cambridge, MA: MIT Press.
2. Calvert GA, Spence C, Stein BE (2004) The Handbook of Multisensory Processing. Cambridge, MA: MIT Press.
3. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264: 746–748.
4. Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26: 212–215.
5. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17: 1147–1153.
6. MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. Br J Audiol 21: 131–141.
7. Arnold P, Hill F (2001) Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. Br J Psychol 92 Part 2: 339–355.
8. Navarra J, Soto-Faraco S (2007) Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol Res 71: 4–12.
9. Angelucci A, Bullier J (2003) Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? J Physiol Paris 97: 141–154.
10. Bubic A, von Cramon DY, Schubotz RI (2010) Prediction, cognition and the brain. Front Hum Neurosci 4: 25.
11. Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360: 815–836.
12. Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2: 79–87.
13. Wolpert DM (1997) Computational approaches to motor control. Trends Cogn Sci 1: 209–216.
14. Bar M (2004) Visual objects in context. Nat Rev Neurosci 5: 617–629.
15. Lamme VA (1995) The neurophysiology of figure-ground segregation in primary visual cortex. J Neurosci 15: 1605–1615.
16. Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL (2002) Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci U S A 99: 15164–15169.
17. Keller PE, Koch I (2008) Action planning in sequential skills: relations to music performance. Q J Exp Psychol (Colchester) 61: 275–291.
18. Di Lollo V, Enns JT, Rensink RA (2000) Competition for consciousness among visual events: the psychophysics of reentrant visual processes. J Exp Psychol Gen 129: 481–507.
19. Enns JT, Lleras A (2008) What's next? New evidence for prediction in human vision. TRENDS Cogn Sci 12: 327–333.
20. Spratling MW (2008) Predictive coding as a model of biased competition in visual attention. Vision Res 48: 1391–1408.
21. Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends Cogn Sci 13: 403–409.
22. Pickering MJ, Garrod S (2007) Do people use language production to make predictions during comprehension? TRENDS Cogn Sci 11: 105–110.
23. Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. Philos Trans R Soc Lond B Biol Sci 363: 1071–1086.
24. Skipper JI, van Wassenhove V, Nusbaum HC, Small SL (2007) Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb Cortex 17: 2387–2399.
25. Stevens KN, Halle M (1967) Remarks on analysis by synthesis and distinctive features. In: Wathen-Dunn W, ed. Models for the perception of speech and visual form: proceedings of a symposium. Cambridge, MA: MIT Press. pp 88–102.
26. Van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102: 1181–1186.
27. Van Berkum JJ, Brown CM, Zwitserlood P, Kooijman V, Hagoort P (2005) Anticipating upcoming words in discourse: evidence from ERPs and reading times. J Exp Psychol Learn Mem Cogn 31: 443–467.
28. Dambacher M, Rolfs M, Gollner K, Kliegl R, Jacobs AM (2009) Event-related potentials reveal rapid verification of predicted visual input. PLoS One 4: e5047.
29. DeLong KA, Urbach TP, Kutas M (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat Neurosci 8: 1117–1121.
30. Fowler CA (2004) Speech as a supramodal or amodal phenomenon. In: Calvert G, Spence C, Stein BE, eds. The Handbook of Multisensory Processing. Cambridge, MA: MIT Press. pp 189–202.
31. Rosenblum LD, Miller RM, Sanchez K (2007) Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. Psychol Sci 18: 392–396.
32. Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. Neuroimage 25: 76–89.
33. Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E (2003) "Putting the face to the voice": matching identity across modality. Curr Biol 13: 1709–1714.
34. Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5: e1000436.
35. Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci 19: 1964–1973.
36. Soto-Faraco S, Navarra J, Weikum WM, Vouloumanos A, Sebastian-Galles N, et al. (2007) Discriminating languages by speech-reading. Percept Psychophys 69: 218–231.
37. Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B, Campbell R, eds. Hearing by Eye: The Psychology of Lipreading. New York (NY): Lawrence Erlbaum Associates. pp 3–51.
38. Schwartz JL, Robert-Ribes J, Escudier P (1998) Ten years after Summerfield … a taxonomy of models for audiovisual fusion in speech perception. In: Campbell R, Dodd B, Burnham D, eds. Hearing by Eye, II. Perspectives and directions in research on audiovisual aspects of language processing. Hove (UK): Psychology Press. pp 85–108.
39. Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29: 13445–13453.
40. Forster KI, Forster JC (2003) DMDX: a windows display program with millisecond accuracy. Behav Res Methods Instrum Comput 35: 116–124.

2.3 Sánchez-García, C., Enns, J. T., and Soto-Faraco, S.

Cross-modal prediction in speech depends on prior linguistic experience. Exp Brain Res. 2013 Feb 06;225(4): 499-511. DOI 10.1007/s00221-012-3390-3

2.4 Angèle Brunellière, Carolina Sánchez-García, Nara Ikumi, Salvador Soto-Faraco

Visual information constrains early and late stages of spoken-word recognition in sentence context. International Journal of Psychophysiology. 2013 Jun 22;89(1): 136-47. DOI 10.1016/j.ijpsycho.2013.06.016

3. CONCLUSIONS

3.1 Summary of results

The present dissertation addressed possible cross-modal predictive mechanisms operating online during speech perception. To this aim, we used sentences consisting of a unimodal context ending in an audiovisual target to explore the possibility of online intra-modal (i.e. visual-to-visual and auditory-to-auditory) as well as cross-modal predictions (i.e. from auditory to visual modality and from visual to auditory modality). As expected, we found beneficial effects when the context was target-consistent within sensory modalities. Importantly, we demonstrate for the first time that prediction in speech also operates at a cross-modal level, more precisely from visual to acoustic continuity (but not vice-versa) (Sánchez-García et al. 2011). Based on these initial findings we hypothesized that cross-modal prediction takes place initially at a pre-lexical level of information processing, possibly phonological or pre-phonological. In a follow-up study we therefore addressed whether these predictive mechanisms operated specifically at a pre-phonological or they occurred at a phonological level by comparing the effects of cross-modal prediction in native and non-native languages. We found predictive effects during the processing of native language only, leading us to conclude that visual speech cues have a predictive influence during early stages of processing, operating in a cross-modal fashion mostly based on phonological representations, rather than pre-phonological information (Sections 2.2 and 2.3; Sánchez-García et al. 2011; 2013). In addition, this thesis explored whether this phonologically-based prediction from visual articulatory information interacts with the effects arising from constraints imposed by the semantic context during sentence processing. To answer this question, we looked at the possibility that visual information might have an influence that carries on beyond phonological stages of processing. We used ERPs to address if visual temporal anticipation, i.e. a more or less visual predictive viseme* at word onset (i.e., available to the perceiver more or less early in time with respect to its acoustic correlates) might interact somehow with the semantic integration of words into the sentence context. We observed that highly informative visual cues at word onset modulate the effects of preceding semantic context at the stage of lexical selection. Below, I present a summary of these findings.

In the first study (Section 2.2; Sánchez-García et al., 2011), we showed that prior speech context can be used to form predictions about the incoming input within the same sensory modality (visual or auditory), speeding up the detection of audiovisual match/mismatch. Furthermore, in some cases, information can also be transferred from one modality to the other. That is, a cross-modal prediction effect was found from visual-to-auditory modality during speech perception. The focus of this dissertation is precisely the process whereby information extracted from the visual articulatory gestures of the speaker is used to make an auditory (cross-modal) prediction during speech perception.

In this first study we observed that visual speech context speeds up audiovisual speech processing. In particular, it allowed participants to detect a mismatch between lips and speech sounds faster than when the visual context was not continuous with any modality during the AV target. We inferred that visual information could be used in order to make predictions about subsequent auditory information, therefore facilitating detection of a mismatch. Our hypothesis, described in the Introduction, was that cross-modal predictive mechanisms might operate in a unidirectional manner, from visual-to-auditory modality, capitalizing on the natural anticipation of visual information with respect to its acoustic correlates occurring in speech production. Confirming our hypothesis, evidence for prediction was only observed from vision to audition, but not in the opposite direction, that is, when the prior context was auditory. As described in Section 2.1, a second issue was to address the nature of this prediction, discriminating between the possible pre-phonological or phonological basis of this online transfer of information.

This was addressed in the second study (Section 2.3; Sánchez-García et al. 2013), where the nature of the visual-to-auditory cross-modal prediction effect was explored. The distinction between the phonological and pre-phonological levels of processing is based on the following logic: 1. If predictions are based on pre-phonological information, then the visual speech cues might be enough for prediction to occur in any language, even if unfamiliar to the perceiver. 2. If, on the other hand, the spatiotemporal relation between visual and auditory information is not enough for prediction to occur and, instead, a more abstract level of representation (i.e. phonology) is necessary, then cross-modal prediction effects will only be observed in a language the perceiver masters. We therefore tested visual-to-auditory prediction effects in the perceivers' native language and in a language not familiar to them. We observed that the benefit from prior visual information on AV mismatch detection was restricted to the native-language situation, where the perceiver has solid knowledge of the language's phonological repertoire. Based on this result, we suggested that cross-modal correlations based on basic spatiotemporal audiovisual correlations present in speech, but that are not language specific, might not be enough to support cross-modal online predictions. Rather, a specific knowledge about the phonological repertoire of the perceived language seems to be necessary to capitalize on cross-modal online prediction during speech perception. Our results thus imply that the anticipatory visual correlates of the auditory phonemes are useful to make online predictions during speech. Interestingly, this mechanism is more effective when perceivers know the perceptual consequences of the articulatory gesture they are seeing (i.e. the relation between visual articulatory features and phonological representations).

In the third study included in this dissertation (Section 2.4.; Brunellière et al. 2013), we explored the possibility that the effect of visual saliency (more or less informative visual cues) at word onset might interact with the effects of semantic context on lexical selection stages during sentence processing. To do so, we took advantage from the fact that highly visually salient phonemes (i.e. articulated at front parts of the articulatory apparatus, such as the lips in the phoneme /p/) provide more constraining information than more ambiguous phonemes (i.e. /k/) with respect to its auditory correlates. Based on this, we observed that visual saliency interacts with the effects imposed by semantic context at a relatively late stage (reflected in a modulation of the late N400 component). We interpreted that this phonological-semantic interaction takes place by the time of selecting the word that best matches with the expectations following a biasing semantic context. Based on the results of this ERP study, we speculated that visually salient phonemes mainly help the early rejection of inappropriate lexical candidates, in keeping with the fact that visual information is available ahead of time in visually salient vs. ambiguous phonemes. This provides evidence that visual cues might therefore support a more efficient / earlier lexical selection.

3.2 Implications of the present findings

3.2.1 Online predictive mechanisms during speech perception operate cross-modally based on anticipatory visual information.

Previous research has explored the extent to which linguistic information conveyed in visual speech contributes to speech perception (Bernstein, 2005; Calvert, Spence, & Stein, 2004; Massaro, 1987, 1998; Massaro & Stork, 1998; Rosenblum, 2005; Summerfield, 1987; Sumby & Pollack, 1954). Visual information not only contributes to improve recognition and identification of auditory speech under some (acoustically degraded) conditions, but it can be by itself enough for identifying words (Auer & Bernstein, 1997; Auer, 2002; Lachs, Weiss, & Pisoni, 2000; Mattys, Bernstein, & Auer, 2002) as well as to discriminate languages from one another (Soto-Faraco et al., 2007; Ronquest, Levi, & Pisoni, 2010) even from early on in life (Weikum et al., 2007). All these studies illustrate that the visual modality is rich in linguistic information, perhaps more than had been initially thought (Bernstein, Demorest, & Tucker, 1998; Samuelsson & Rönnberg, 1993). Furthermore, previous studies have shown that when accompanying auditory information in natural speech, visual information can speed up subsequent speech processing (van Wassenhove et al. 2005; Stekelenburg & Vroomen, 2007; Arnal et al. 2009). The results of this thesis add to this body of evidence.

As seen, visual speech has the potential to convey extensive linguistic information and this information is extracted and used at different levels during speech perception. Evidences from pre-phonological (Schwartz et al. 2004; Stekelenburg & Vroomen, 2007; Arnal et al. 2009), phonological (van Wassenhove et al. 2005; Arnal et al. 2009; all the studies presented in this dissertation (Sánchez-García et al. 2011; 2013; Brunellière et al. 2013)) and lexical (Kim, Davis, & Krins, 2004; Buschwald et al. 2009) levels of information processing have been convincingly shown.

In the studies presented in this thesis, we tap on the role of visual speech information regarding predictive mechanisms operating online in a cross-modal fashion during speech processing. All together, our results show that under some conditions, visual information pre-activates features in an anticipatory fashion as speech unfolds, speeding up the processing of the subsequent speech signal. In sum, visual speech cues seem to have an influence during early stages of processing, operating in a cross-modal fashion at a phonological level (Sections 2.2. and 2.3; Sánchez-García et al. 2011; 2013) and at a later stage, during lexical processing (Section 2.4; Brunellière et al. 2013), interacting with the semantic context of the sentences.

When exposed to natural speech, perceivers use a variety of informational sources to constrain the interpretation of the incoming message (see Kutas, DeLong & Smith, (2011) for a review). During speech comprehension, different linguistic representations (semantics, syntax, phonology...) are available to the perceiver, who uses the information to make linguistic predictions at multiple levels (Pickering & Garrod, 2006; 2013). That is, under some conditions, immediately preceding speech context can be used to form predictions about the upcoming input. Indeed, our results integrate well within this framework. For example, in Sánchez-García et al. 2011 (Section 2.2) we report behavioral evidence for online predictive mechanisms operating within the same modality (i.e. auditory-to-auditory and visual-to-visual), in line with previous ERPs studies on spoken and written language, which have shown intramodal prediction at the level of syntax (Van Berkum et al. 2005), semantics (Dambacher et al. 2009) or phonology (DeLong et al. 2005). One of the most important finding arising from the studies presented in this thesis (Sections 2.2 and 2.3) is that online prediction can occur cross-modally, albeit in a unidirectional manner, from visual to auditory modality. This constitutes, to the best of our knowledge, the first behavioural demonstration that prior visual speech-reading information by itself can be used to benefit subsequent speech processing in an online manner.

As discussed, predictions can operate at different stages during speech processing. But, what is the concrete linguistic level at which information is extracted from visual input to make cross-modal prediction possible? Which is the mechanism underlying this unidirectional perceptual benefit? In the paragraphs above, as well as in the discussions of the experimental section, we have suggested the phonological level as relevant. In this section we discuss in some extent this and several other possible levels of information that could be at the base of the cross-modal visual-to-auditory predictive mechanisms operating in the studies presented in Sections 2.2 and 2.3.

a. Phonological representations are at the basis of cross-modal prediction

In order to account for the level of processing supporting the cross-modal prediction based on visual information suggested by our initial findings (Section 2.2; Sánchez-Garcia et al. 2011), one could think about several possibilities, which might include semantics, syntax, phonology, pre-phonology…. Among all those levels of information, we argued that the predictive mechanisms operating online in a cross-modal fashion during speech perception are firstly based on a phonological level of information. Evidence from the first study allowed us to discard higher levels of information processing (semantic, syntactic, lexical...) due to the scarcity of information provided by visual speech at these levels (Bernstein et al. 1998; Soto-Faraco et al. 2007; Altieri, Pisoni, & Townsend, 2011). Yet, that study left open the possibility that prediction was based on phonological or pre-phonological (more domain-general) levels of processing of the speech signal. The claim for a major role of phonological information is supported by the results of our second study (Section 2.3; Sánchez-Garcia et al. 2013), where we observed predictive effects during speech processing only in the participants' native language, but not when testing for visual-to-auditory prediction in participants' non-native languages. Such a result might indicate that cross-modal correlations based on spatiotemporal dynamics between visual and acoustic signals might not be, on themselves, sufficient to support effective cross-modal prediction during speech perception. Instead, specific knowledge about the phonological repertoire of the language on the perceiver's side is required to capitalize effectively on anticipatory visual information during AV speech processing.

Previous findings suggested that visual anticipatory information is not only informative about the timing of the arrival of auditory information (Schwartz et al. 2004; Stekelenburg & Vroomen, 2007), but it also bears on phonologically structured information. In fact, the natural visual anticipation with respect to its acoustics correlates (Chandrasekaran et al. 2009) differs between phonemes. In some cases, when articulatory gestures are articulated in the front part of the articulatory system (for instance, in bilabials as /p/ or /b/), they are visible to the perceiver early on in time. The perceiver can use this visual information to constraint subsequent auditory information, speeding up speech processing. This effect has been observed during the presentation of syllables (van Wassenhove et al. 2005; Stekelenburg & Vroomen, 2007; Arnal et al. 2009) and here we show it during the presentation of words embedded in sentences (Section 2.4; Brunellière et al. 2013). Moreover, in Brunellière et al. 2013 (Section 2.4), we demonstrate that visual information facilitates lexical selection through the use of anticipatory mechanisms based on phonological information.

Our results from the three studies presented in this dissertation support phonological information as the ground for cross-modal prediction mechanisms operating during speech perception. We claim that cross-modal predictive mechanisms occur primarily at a phonological segmental level, based on associations between visual articulatory features and phonological representations. This possibility links well with the idea that, at the phonological level, auditory and visual modalities share a common format, which might be provided by the articulatory representations used in speech production (Pickering & Garrod, 2006; Skipper et al. 2005, 2007; Fowler et al. 2004; Rosenblum et al. 2007). In the next subsections I will discuss the phonological level as a putative base for visual predictive mechanisms during speech perception in contrast with other linguistic levels of information.

b. Phonological vs. Lexical level

Previous studies have shown that visual articulatory information may have a contribution during lexical recognition processes (Jesse & Massaro, 2010). For instance, a visually spoken word facilitates subsequent processing of the same spoken word during auditory word-recognition (Buchwald et al. 2009) or lexical decision tasks (Kim et al., 2004). Even when only viewing the articulation of a syllable that matches the onset of a word subsequently presented, visual information facilitates its posterior auditory recognition (Fort et al. 2012). Moreover, visual information seems to activate the same lexico-semantic network than auditory information (Dodd, Oerlemens, & Robinson, 1989). These studies illustrate that visual information can facilitate speech processing at a phonological level in a cross-modal fashion, supporting lexical access, even when this visual information is not enough for the recognition of the word. In our last study (Section 2.4; Brunellière et al. 2013) the influence of visual speech cues over the stage of lexical selection was observed. Since only visually salient phonemes helped a more efficient selection of a word from the lexicon*, we assume this effect was based on phonological information.

But, to which extent reliable lexical information can be extracted from visual speech? Some evidences have shown that visual information carries little lexical or semantic information on its own (Bernstein et al. 1998; Soto-Faraco et al. 2007; Altieri et al. 2011). In fact, in Section 2.2 (Sánchez-García et al. 2011), we explicitly tested the amount of lexical information participants were able to extract from our materials by asking a group of (Spanish) participants to lip-read as much words as they could from the (Spanish) sentences. The result was that subjects could not visually recognize more than 4 % of the words from the sentences. This is in accord with the fact that the capacity of overtly identify speech-read words is often very poor in normal hearers (Auer, 2002), at least when dealing with unconstrained speech in sentences, despite large individual differences in this ability (Bernstein et al. 1998, 2000).

According to the results observed in our studies (Sections 2.3 and 2.4; Sánchez-García et al. 2013 and Brunellière et al. 2013), the contribution of lexical cues does not seem to be very prominent a priori, due to the difficulty of extracting information at this level. Thus, by elimination and based on the arguments above, predictive mechanisms seem to hinge mostly on phonological information. Nevertheless we don't discard the possibility that the lexical level might play some role in prediction, even if this does not seem to be the most important source of visual-to-auditory prediction during speech.

In fact, based in the results presented in Sections 2.2 and 2.3 (Sánchez-García et al. 2011; 2013), it might be possible that the faster resolution of phonemes thanks to visual anticipatory information carries over to the lexical level, pre-activating words in the lexicon* that, when matching with the auditory continuation, speed up speech processing. As shown by Kim et al. (2004) visual information can act as a prime at the phonological level, facilitating lexical access. In Kim et al.'s study, participants were not able to identify the words by lip-reading but they did show a visual to auditory priming effect in naming as well as in written and auditory lexical decision tasks with the same materials. In addition, we suggest that lexical representations might be based on abstract, phonological, language specific representations (Pallier, Colomé, & Sebastián-Gallés, 2001; Buschwald et al. 2009), which might explain the lack of cross-modal prediction evidences in an unfamiliar language, in the case that cross-modal prediction might occur at the lexical level too.

The visual phonological-based effects at a late stage during sentence processing (i.e. lexical selection) observed in Section 2.4 (Brunellière et al., 2013), constitute a novelty with respect to previous ERPs studies regarding visual influence in audiovisual processing. The typical temporal modulation of the N100 component depending on the visual saliency of the phonemes, previously found in words (Mengin et al., 2012; Shahin, Kerlin, Bhat, & Miller, 2012), syllables or vowels (van Wassenhove et al. 2005; Besle et al. 2004; Klucharev, Möttönen, & Sams, 2003; Stekelenburg & Vroomen, 2007; Arnal et al. 2009), was replicated in our study (Section 2.4; Brunellière et al. 2013), but here using words embedded in sentential context. In contrast with previous studies, the use of whole sentences as stimuli allowed us to explore the effect of visual speech information in a more natural context. This is important not only because speech is often experienced in sentential context in everyday life, but also because it offers the opportunity to study any possible interaction between different kinds of constraints present in natural speech (in this case, visual saliency and meaning).

The effect of visual speech information during word processing in sentences in Section 2.4 (Brunellière et al, 2013) was also reflected by an increased AV-related negativity in the late part of the N400 component compared with Audio only speech, which only occurred when the word onset contained visually salient phonemes (i.e. /p/) with respect to less salient phonemes (i.e. /k/) (Section 2.4.; Brunellière et al. 2013). This might be related with the fact that visual information in highly salient phonemes is available earlier in time and it supports a stronger basis for rejection of inadequate lexical candidates in word targets beginning with these cues. In agreement with these results, in a behavioural study and using a priming plus lexical decision task, Fort et al. (2012) showed that even when only showing the articulatory gestures corresponding to the initial syllable of a word (which always started with a highly visually salient phoneme, i.e. /b, p, m, v, f, s/), lexical processes were activated, facilitating subsequent auditory recognition of that same word. Taken together, Fort's results and ours (Section 2.4; Brunellière et al. 2013), suggest that as soon as the articulatory gestures are available in the visual signal, they can be used to start lexical processes.

In line with this, the temporal difference between visual and auditory information during natural speech (Chandrasekaran et al. 2009; Cathiard, Lallouache, Mohamadi, & Abry (1995) (see also Smeele, 1994 and Jesse & Massaro, 2010) sets the appropriate time-window for the generation of predictive signals that influence auditory perception, supporting phonological prediction based on visual information. That is, seeing visually salient articulatory gestures (i.e. /p/) might facilitate predictions about the identity of the forthcoming auditory information (van Wassenhove et al. 2005), supporting the rejection of inadequate lexical candidates. During sentence processing, in addition to the phonology-based-predictions, there are expectations created from the semantic context. Therefore, when visual information about the initial phoneme of the target word in a sentence arrives, this (more or less visually salient) information might support rejection of lexical candidates not matching with the semantic expectations from the context, facilitating lexical selection of words during speech processing (Section 2.4; Brunellière et al. 2013). The upcoming auditory information (received later on in time) might refine the selection of the word from the lexicon*, providing complementary features after a first pre-selection based on the phonological visual information.

After discussing the interaction between phonological and lexical levels of processing, in the next subsection we will discuss the possibility of predictive mechanisms operating at (or interfacing with) processing levels earlier than the phonological.

c. Phonological vs. Pre-phonological information

Given the evidences supporting facilitation of audiovisual processing based on spatiotemporal matching across modalities regarding low-level features (Stekelenburg & Vroomen, 2007; Arnal et al. 2009; Green, 1998; Green & Kuhl, 1991; Grant, 2001; Kim & Davis, 2003; Rosenblum, 2005), the contribution from pre-phonological representations during speech processing cannot be completely ruled-out (Summerfield, 1987; Schwartz et al. 1998; 2004). For instance, Schwartz et al. (2004) showed that an identical lip-gesture (and therefore non-informative about the phonetic content of the sound) combined with different auditory French syllables improved identification of the presented syllables in noise in comparison with the auditory only presentation. Interestingly, as in Schwartz's study the visual phonemic information was not congruent in content with the presented auditory information, the visual modality only carried information about the timing of the auditory stimuli. Nevertheless, a benefit in intelligibility was observed when visual information was presented simultaneously with the auditory input. This and other studies support the existence of interactions prior to phonetic categorization (Green & Miller, 1985; Norrix & Green, 1996; Schwartz et al. 1998).

Obviously, this evidence runs against our conclusion about the lack of predictive effects based on representations prior to phonological categorization. We hypothesize that at least for fast, online cross-modal prediction mechanisms during speech perception, knowledge about the phonological bases of the language seems to matter for prediction to be efficient (Section 2.3; Sánchez-Garcia et al. 2013). In addition, we suggest that this phonologically-mediated prediction processes might be supported by the generation of an internal model which anticipates the consequences of the observed articulatory movement, and that this mechanism is only efficient when there is a phonological-code shared from talker and perceiver (Sections 2.2 and 2.3; Sánchez-García et al. 2011; 2013). This proposal will be described in more detail in Section 3.2.2 of this discussion.

d. Segmental phonology vs. Prosody

Regarding the nature of the phonologically-based cross-modal predictive mechanism that we have proposed (Sections 2.2 and 2.3; Sánchez-García et al. 2011; 2013), it is worth noting that phonological information includes several layers, most notably the dissociation between classical segmental phonology* and prosody*. Previous studies have shown that the mechanics of speech production determine not only the sound of the voice but also the movement of the face (Vatikiotis-Bateson & Yehia, 1996; Yehia et al. 1998). For instance, visible natural head movements from a talker correlate with fundamental frequency F0 and amplitude (Yehia et al. 2002). In addition, seeing the head movements of the speaker improves intelligibility of speech in noise (Munhall et al. 2004), even when only the upper part of the head is visible (Davis & Kim, 2006). Accordingly, the observation of the upper part of the head is enough for participants to match cross-modally a given sentence, first presented auditorily and subsequently visually (Davis & Kim, 2006), even when the visual and auditory stimuli are a different token of the same sentence. The same partial view of the head is also enough to distinguish different prosodic patterns even when the semantic context of the sentence is the same (Cvejic, Kim, & Davis, 2010). It has been also proposed that perceivers can be sensitive to the prosodic information and use the rhythmic pattern to match talker identity across modalities (Kamachi et al. 2003; Lander et al. 2007).

These evidences supporting cross-modal transfer of prosodic information suggest that participants might as well have made use of movement patterns arising from the prosodic contour (i.e. rhythm) of the sentences in order to make the matching during our studies (Sánchez-García et al. 2011; 2013).

With respect to the results presented here, the contribution of such visual prosodic cues is still unknown, as we cannot be sure from which part of the speaker's face participants extracted the information they were using to make the prediction (in our studies we presented the whole face of the talker).

Supporting the possibility that information is not only extracted from the lips of a speaker, results from eye-movement studies have shown that during speech perception, listeners not always look at the mouth area from the speaker (Lansing & McConkie, 1999; Vatikiotis-Bateson et al. 1998). This is true from early on in life. A recent study showed that from 12 month-old babies shift their attention from the mouth to the eyes of the speaker when perceiving their native language

(Lewkowicz & Hansen-Tift, 2012). However, this pattern changed when exposed to a non-native language. In that case, and at the same age, babies kept their attention fixed in the mouth of the speaker. Therefore, it could be the case that movements from head/face/eye-brows could support some of the cross-modal transfer of information that facilitates the audiovisual processing, especially when participants are exposed to their native language, as observed in our study in Section 2.3 (Sánchez-García et al. 2013).

Some studies have measured eye movements during audiovisual speech showing that subjects foveated primarily on the stimulus speaker's eyes or mouth (Vatikiotis-Bateson, Eigsti, & Yano, 1994; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). In these studies, the proportion of time subjects fixated on the mouth increased as the level of acoustic masking noise increased too. This is interesting because in our studies in Sections 2.2 and 2.3 (Sánchez- García et al., 2011; 2013), the prior context was purely visual. According with the mentioned studies, in visual only speech, as in speech in noise, the mouth might be probably a primary cue to extract linguistic content without sound, pointing to the segmental phonological information as main source for the predictive information, which may be putatively supported by other prosodic cues.

Thus, with the design used in Section 2.3 (Sánchez-García et al., 2013) we cannot pinpoint whether it is segmental phonology* or prosody* what plays a role during the observed cross-modal prediction (see Soto-Faraco, Calabresi, Navarra, Werker & Lewkowicz, 2012, for related discussion). However, we advocate the idea that probably is a combination of primarily segmental phonology, supported by the prosodic cues. A possible way to address this hypothesis, namely that most of the information is extracted from the mouth of the speaker, might be to use, for example, a paradigm in which only the lips from the speaker are visible to the perceiver, in order to discard most of the prosodic cues. In this way, we could explore with more detail which is the information used to make the prediction.

Another related possibility is that the predictive process could be supported by the familiarity with the rhythmic pattern. Several studies have shown that it is possible to discriminate between languages with different rhythm pattern (i.e., stress-timed (i.e., English and German) and syllable-timed (i.e., Spanish)), using rhythmic information (Ronquest et al. 2010; Ramus, Nespor, & Mehler, 2000). This is of relevance because the non-native languages tested in Section 2.3 (Sánchez-García et al., 2013) belonged to a different rhythmic class than the participant's native language. That is, Spanish is a syllable-timed language, while English and German are stress-timed languages (Abercrombie 1967).

Against the possibility that the particular rhythm of a language might be at the bases of the prediction, Lander et al. (2007) showed that the cross-modal transfer of information which allowed to identify a talker was not dependent on the knowledge about the presented language. This indicates that the information extracted from the prosodic cues and used cross-modally is particular of each speaker, rather than particular of a language.

The differences observed between known and unknown language regarding prediction might be due to the fact that visual cues, from segmental phonology or/and prosody are more useful when one knows the language being presented with. For instance in studies of language discrimination by lip-reading, participants were only able to distinguish between two languages when they were familiar with at least, one of them (Soto-Faraco et al. 2007; Ronquest et al. 2010). Accordingly, in Section 2.3 (Sánchez-García et al. 2013), a clear representation of the perceived language was necessary in order to take advantage from visual information through the use of predictive mechanisms.

In the next section we will discuss the effect of prior linguistic knowledge in different aspects of speech processing, focusing on predictive mechanisms during speech perception in non-native languages.

3.2.2 Online predictive mechanisms based on visual information depend on language-specific experience.

In Section 2.3 (Sánchez-García et al., 2013) we have shown that experience with the language is paramount for making rapid online predictions based on visual cues during audiovisual speech processing.

Many researchers have argued that phonemic categories particular of each language are established during the first year of life, tailored to the specific input from the native language environment (Best & McRoberts 2003; Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995; Werker & Tees, 1984). Thereafter, they will act as a "sieve" for the rest of the languages perceived (Best, 1995; Flege, 1995; Pallier, Bosch, & Sebastián-Galles, 1997). For instance, the native language has an effect during discrimination of non-native phonemic contrasts of a second language (L2), even if L2 was learned as early as at the age of 3-4 years-old (e.g. Sebastián-Gallés & Soto-Faraco, 1999).

Moreover, spoken-word recognition studies have shown that lexical selection in a non-native language is sensitive to phonological similarity between that language and the perceiver's native language (Nas, 1983) and words from both languages are activated even if only words from one language are heard (Spivey & Marian, 1999). Also, lexical competition is greater when listening to non-native languages than to native languages, due to the sum of second-language competitors, activated as result of difficulties in phonetic discrimination, in addition to the native competitors (Weber & Cutler, 2004). These examples show how the native language's representations influence the perception of other languages.

During audiovisual speech perception, prior experience with a language also influences the integration of visual and auditory information (see Sekiyama & Tohkura, 1993; Sekiyama, 1997; Sekiyama & Burnham, 2008), including cross-modal temporal processing aspects (Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010). Recently, an fMRI study revealed that activity in the bilateral occipital lobe was stronger for congruent AV stimuli in a non-native language compared to the native language (Barrós-Loscertales et al., 2013), showing that the multisensory processing of native and non-native languages also differs in terms of brain activation.

All the studies described above illustrate how native and non-native language processing varies regarding different linguistic aspects. Thus, it looks like cross-modal predictive mechanisms operating during speech perception might be one more example of the differences regarding the working of perceptual systems in native and non-native languages. Namely, during AV speech processing, the speech information extracted from visible articulatory movements might be only exploited in full and used to make a prediction when the visemic categories belong to the listener's native repertoire. We argue that, in a non-native language, phonological information is perceived but is more difficult to match to any existing category, being less efficient to ground online predictive mechanisms.

Differences between native and non-native audiovisual speech perception seem to be linked with the direct relationship existing between the amount of experience articulating speech sounds and the influence of visual information during audiovisual speech perception. For instance, Desjardins, Rogers, & Werker (1997) showed that perception of audiovisual speech in preschoolers' is considerably less influenced by visual information than in adults, because infants are less experienced than adults producing speech. A similar effect has been shown in adults with impaired speech production abilities (Siva, Stevens, Kuhl, & Meltzoff, 1995).

Accordingly, some studies have shown that the motor system is involved during the process of speech perception and furthermore, that the activation of the motor system is sensible to native and non-native languages perception (Wilson & Iacoboni, 2006; Swaminathan et al., 2013, among others). These evidences make us wonder whether our findings could be related with the proposal of a strong relationship between speech perception and production systems. That is, phonemic recognition during speech perception is possible because speaker and observer share the same articulatory motor repertoire. We discuss this possibility in the next subsection.

3.2.3 Perception-production links and visual predictive mechanisms: A predictive coding framework.

In the last decade some researchers have proposed that the motor system (involving motor areas of the brain) is involved in speech perception (see Introduction; Section 1.2.2). Even if the present thesis does not focus on the role of the motor system in speech perception, we think that our findings could be related with the speech perception and production link. In this subsection we discuss our results within this framework.

Some studies have brought evidences that motor systems could provide a specific functional contribution to the perception of speech sounds. For instance, TMS studies have shown that the somatotopic organization of the motor cortex is reflected during speech comprehension (Fadiga et al. 2002; D'Ausilio et al. 2009). For instance, Fadiga et al. (2002) showed an increase of motor-evoked potentials (MEPs) recorded from the listeners' tongue muscles when the onset of heard words specifically involved, when pronounced, tongue movements. In the same line, D'Ausilio et al. (2009) found that stimulating the motor representation controlling the articulator producing the perceived speech sound improved the perception of that given sound, while inhibitory effects were seen when stimulating motor representations of articulators related to discordant speech sounds.

Furthermore, the involvement of motor areas in speech perception seems to be sensitive to prior knowledge about the presented language. For example, in an fMRI study, Wilson & Iacoboni (2006) showed a greater activation of motor areas during listening to non-native vs. native phonemes. Recently, Swaminathan et al. (2013) in a TMS study showed a higher motor excitability during observation of a known language compared with an unknown language or non-speech mouth movements, suggesting that motor resonance is enhanced specifically during observation of mouth movements that convey linguistic information that speaker and perceiver share. Together with other studies described in the Introduction (Watkins et al. 2003; Watkins & Paus, 2004; Skipper et al. 2005; 2007), these results support the idea that during speech perception, an online simulation (i.e. internal model) about the movements to articulate the sounds might be in function. That is, while perceiving speech, mechanisms associated with language production are engaged (i.e. the motor system is involved during speech perception), accordingly to the original idea of the Motor Theory of Speech Perception (Liberman & Mattingley, 1985; Wilson & Iacoboni, 2006; but see Venezia, Saberi, Chubb, & Hickok, 2012 for motor activation in TMS studies during speech perception as a response bias effect). Moreover, the motor regions are sensitive to whether or not phonemes are part of the speaker's inventory, which supports the idea that motor areas play an active role in the speech perception process and that the motor system is most readily able to simulate known phonemes (Wilson & Iacoboni, 2006; Swaminathan et al., 2013).

As discussed previously, in Section 2.4 (Brunellière et al., 2013) we showed that the modulation of lexical processing for words beginning with a visually salient phoneme might be mediated by covert imitation of the phoneme by the perceiver. More salient phonemes (such as /p/) are visible earlier in time than less salient phonemes (e.g. /k/) (van Wassenhove et al. 2005) and its internal simulation might allow the system to initiate the process of lexical access earlier in time. This possibility was also suggested by Fort et al. (2012), and also speaks about the production-perception link (Pickering & Garrod, 2006; Skipper et al. 2007; van Wassenhove et al. 2005; Kerzel & Bekkering, 2000; Liberman & Mattingley, 1985).

In this line of thinking, predictive coding models (Pickering & Garrod, 2006; Skipper et al. 2007; van Wassenhove et al. 2005), based on the analysis-by-synthesis approach (Stevens & Halle, 1967), propose that when perceiving speech, information at several levels of processing (i.e. semantic, syntactic, phonological) is extracted from the signal and used continuously to activate/update hypothesis about the consequences of the articulatory movements which might be involved in the production of the incoming utterance*. The hypothesis is mapped into a motor plan that can be followed to reach that goal, which results in a forward prediction of the consequences of executing those motor commands, reflected in an efference copy. The sensory prediction so generated is compared with the current sensory state of the perceiver at each moment. This continous matching process, when successful, results in a speed up of speech processing, by constraining the interpretation of the incoming message.

We suggest that, according with predictive coding models, visual input from seeing the speaker articulating might activate the motor system, which will be engaged in a simulation through a top-down prediction about the consequences of seeing that specific articulatory gestures. The predictions, based on previous experience with the language will be compared with the perceived information, speeding up the processing of subsequent audiovisual information (i.e. bottom-up input) when it matches the expectation. We entertain the hypothesis that this process is much more efficient and reliable if perceiver and producer share the same phonological categories (i.e. from a perception and production point of view). Otherwise, the absence of fine grained knowledge about the articulatory correlates of the observed phonemes may prevent an efficient prediction to be created. This idea is in accord with previous findings within the predictive coding framework, however further research will be necessary to test this hypothesis.

In the next section I will describe the implications that our results may have during communication.

3.2.4 The role of online predictive mechanisms during speech in interactive contexts

In the predictive coding framework, described in the previous section (Pickering & Garrod, 2006; Skipper et al. 2007; van Wassenhove et al. 2005), Pickering & Garrod (2013) recently proposed a model establishing a parallelism between language production and action, and between comprehension and action perception. Following Pickering & Garrod's model, while perceiving speech during a conversation, speakers might use forward production models, by constructing efference copies of the predicted utterance* that will be then compared with the output of the speaker. This means that listeners are able to predict speakers' upcoming utterances because themselves might articulate the utterance that they are listening/observing through covert imitation (i.e. automatic imitation), anticipating the consequences of the articulation and comparing them with the current input received. According with Pickering & Garrod's model, experience with the language seems paramount in order to be able to make a simulation about the utterance. Nevertheless, Pickering & Garrod distinguished between a *prediction-by-simulation* (which might occur, for instance, in a dialogue, because the perceiver has to participate actively and this activates the production system) and *prediction-by-association* (for instance, when a common language is not shared by perceiver and speaker. In this case, the prediction must be done based on the information the perceiver has about the speaker or on what s/he has learned from a similar situation).

According to the model from Pickering & Garrod's, prediction-by-simulation is based on the knowledge about how to articulate the utterance (with the contribution of the motor system), while prediction-by-association relies on auditory information from the speaker. In this framework, our results might indicate that visual information activates prediction-by-simulation mechanisms, which operate when one knows the language and is able to estimate the consequence of the articulatory movement (involving the motor system) but has only available visual information (i.e. not optimal conditions). This is interesting, because consistently with findings from TMS such as the one from D' Ausilio, Bufalari, Salmas, & Fadiga (2012), the contribution of motor systems to speech perception might be restricted to situations in which speech is degraded, such as in noisy conditions, which is a situation a priori comparable to when only visual information is available.

Note that this prediction-by-simulation mechanism acts in situations where the perceiver is actively participating, such as in a conversation (Pickering & Garrod, 2013). Following this idea, the intention from the perceiver to understand the message could be another factor playing a role for the differences we observed between native and non-native language (Section 2.3; Sánchez-García et al. 2013). The motivation to understand the message while the perceiver is involved in a conversation might vary when exposed to a language s/he understands and when s/he does not.

Language is a joint action (Pickering & Garrod, 2006). A joint action is defined as "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment". Joint actions have been defined as depending on the ability to share representations, predict actions and integrate the predicted effects of one's own and others' actions (Sebanz, Bekkering, &

Knoblich, 2006). A dialogue is a successful form of joint action, because both interlocutors want to communicate and they share a common goal. Moreover, both of them share a phonetic code, in which both of them know the relation between acoustic and articulatory representations of the language. Regarding our results, it might be that when participants were presented with a non-native language, in which case perceiver and speaker does not share a common representational code, and it is difficult to predict the consequences of the articulatory movements, participants do not have the sensation of an interaction (join action), and predictive mechanisms are not deployed. If this is true, predictive mechanisms will operate to improve speech processing when perceivers are immersed in an interactive process, and they benefit from the advantages of the mechanism to make communication more fluid.

3.3 Summary of conclusions

The experiments presented in this doctoral dissertation advance several important conclusions in the investigation of the predictive mechanisms underlying audiovisual speech perception. The main conclusions of this thesis are the following:

1. Visual articulatory cues are extracted online during speech perception, and used in a cross-modal fashion to speed up processing of subsequent speech.

2. In order to profit from such predictive anticipatory mechanism, speaker and perceiver must share a common knowledge about the phonological representations of the language.

3. In addition, visual articulatory cues, when highly salient, not only act at an early stage of processing, but they also interact with predictions based on a different level (i.e. semantic*) supporting lexical selection of words during sentence processing.

A way to summarize the implication of our findings in natural situations is to imagine ourselves while having a conversation. The speech input must be processed by the perceiver's part as speech unfolds in time, following a rate set by the speaker. Therefore perceivers have to be able to extract as much cues as possible in real time. In order to create a percept in the more efficient fashion, the perceptual system encodes incoming sensory information, extracting cues from the auditory and the visual modalities and exploiting redundancies and correlations between them. Fortunately, some features are highly correlated across both signals, and visual information is available temporally in advance. This facilitates that during audiovisual speech perception, pre-activation and anticipatory mechanisms operate at different levels, speeding up linguistic processing.

In conclusion, we suggest here that the benefits from visual speech cues during audiovisual speech perception might be part of the linguistic mechanisms working to improve communication strategies. Such mechanisms could mediate a gain in fluidity when one is involved in a linguistic interactive process, such as is, for instance, a conversation.

Glossary

Idiolect: a variety of language that is unique to a person, as manifested by the patterns of vocabulary, grammar, and pronunciation that they use. The idiolect is an amodal articulatory property that can structure both the acoustic and the visual media.

Mental lexicon: A collection of words in long-term memory that mediates access between perception and lexical knowledge.

Pre-lexical processing: In this context, the neural processing of speech sounds before the representation of word identity and meaning.

Phoneme: "The smallest contrastive linguistic unit which may bring about a change of meaning". Chomsky, N.; Halle, M. (1968). The Sound Pattern of English, Harper and Row.

Priming: paradigm to examine changes in responses to a 'target' stimulus when the target is preceded by a 'prime' stimulus. These changes (typically in response time or response accuracy) reflect the relationship between the target and prime stimuli in the cognitive processing required for the task. Priming studies are typically used in psycholinguistics to address issues of whether and when certain representations are active in the course of language processing.

Prosody: the temporal patterns of loudness and pitch associated with stress, intonation, and speaking rhythm. The prosody focuses on aspects of the sound system "above" the level of segments, such as timing, stress and rhythm.

Segmental phonology: Segmental phonology analyses the speech into distinctive units, or phonemes (= 'segmental phonemes'). It focuses on speech sounds (segments), their internal composition and external interactions.

Semantic: Relating to the meaning of things, in this case words and language.

Utterance: An utterance is said to consist of phonetic segments, each consisting of a constellation of articulatory figures.

Viseme: any of several speech sounds which look the same, for example when lip reading (Fisher 1968). The concept of viseme, which classifies visual speech gestures associated to a group of phonemes that are highly confusable upon visual information, such as {/p/, /b/, /m/}.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Aldine, Chicago

Altieri, N., & Townsend, J. T. (2011). *An assessment of behavioral dynamic information processing measures in audiovisual speech perception.* Frontiers in psychology, 2, 238.

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). *Some normative data on lip-reading skills (L).* The Journal of the Acoustical Society of America, 130, 1.

Arnal L. H., Morillon B., Kell C. A., & Giraud A. L. (2009) *Dual neural routing of visual facilitation in speech processing.* Journal of Neuroscience, 29, 13445-13453.

Arnal L. H., Wyart V., & Giraud A. L. (2011). *Transitions in neural oscillations reflect prediction errors generated in audiovisual speech.* Nature Neuroscience, 14 (6), 797-803.

Arnold P., & Hill F. (2001). *Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact.* British Journal of Psychology, 92 (2), 339-355.

Auer, E. T., Jr. (2002). *The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness.* Psychonomic Bulletin and Review, 9: 341-347.

Auer, E. T. Jr., & Bernstein, L. E. (1997). *Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness.* Journal of the Acoustical Society of America, 102, 3704-3710.

Auer, E. T. Jr., & Bernstein, L. E. (2007). *Enhanced visual speech perception in individuals with early-onset hearing impairment.* Journal of Speech, Language and Hearing Research, 50(5), 1157.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., & Halgren, E. (2006). *Top-down facilitation of visual recognition.* Proceedings of the National Academy of Sciences of the United States of America, 103(2), 449-454.

Bar, M. (2007). *The proactive brain: using analogies and associations to generate predictions.* Trends in cognitive sciences, 11(7), 280-289.

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Ávila Rivera, C., & Soto-Faraco, S. (2013). *Neural correlates of audiovisual speech processing in a second language.* Brain and language, 126(3), 253-262.

Benoît, C., & Le Goff, B. (1998). *Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP.* Speech Communication, *26*(1), 117-129.

Bernstein L. E. (2005). Phonetic perception by the speech perceiving brain. Pisoni DB, Remez RE (eds.), The handbook of speech perception (pp 51–78). Blackwell, Malden.

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C., Don, M., & Singh, M. (2002). *Visual speech perception without primary auditory cortex activation.* NeuroReport, 13, 311-315.

Bernstein, L.E., Auer E.T Jr., & Takayanagi, S. (2004). *Auditory speech detection in noise enhanced by lipreading.* Speech Communication, 44, 5-18.

Bernstein L. E., Demorest M. E., & Tucker P. E. (1998). What makes a good speechreader? First you have to find one. Campbell R, Dodd B, Burnham D (eds.), Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech. (pp 211-227). Psychology Press/Erlbaum. Hove, England (UK) Taylor and Francis.

Bernstein L. E., Demorest M. E., & Tucker P. E (2000). *Speech perception without hearing.* Perception and psychophysics, 62(2), 233-252.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). *Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans.* The Journal of Neuroscience, 28(52), 14301-14310.

Besle J., Fort A., Delpuech C., & Giard M. H. (2004). *Bimodal speech: Early suppressive visual effects in the human auditory cortex.* European Journal of Neuroscience, 20, 2225–2234.

Best, C. T. (1995). A direct realist view of cross-language speech, perception. Strange W (ed.), Speech perception and linguistic experience. (p. 171–204). York Press, Timonium.

Best C. C., & McRoberts G. W. (2003). *Infant perception of non-native consonant contrasts that adults assimilate in different ways.* Language and Speech, 46, 183–216.

Best C. T., McRoberts G. W., LaFleur R., &Silver-Isenstadt J. (1995). *Divergent developmental patterns for infants' perception of two nonnative consonant contrasts.* Infant behavior and development, 18, 339–350.

Biau, E., & Soto-Faraco, S. (2013). *Beat gestures modulate auditory integration in speech perception.* Brain and Language, 124 (2), 143-152.

Bishop, D. V. M., Brown, B. B., & Robson, J. (1990). *The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals.* Journal of Speech, Language and Hearing Research, 33(2), 210.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). *Prediction, cognition and the brain.* Frontiers in Human Neuroscience, 4, 25.

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). *Visual speech primes open-set recognition of spoken words.* Language and cognitive processes, 24(4), 580-610.

Burnham, D. K. (1998). Language specificity in the development of auditory-visual speech perception. Campbell, R., Dodd, B. and Burnham, D. (Eds.), Hearing by eye II: advances in the psychology of speechreading and auditory–visual speech (pp. 27–60). Hove, UK: Psychology Press.

Callan, D., Jones, J.A., Munhall, K.G., Kroos, C., Callan, A., & Vatikiotis-Bateson, E. (2003). *Neural processes underlying perceptual enhancement by visual speech gestures.* Neuroreport, 14, 2213-2218.

Callan, D., Jones, J.A., Munhall, K.G., Kroos, C., Callan, A., & Vatikiotis-Bateson, E. (2004). *Multisensory-integration sites identified by perception of spatial wavelet filtered visual speech gesture information.* Journal of Cognitive Neuroscience, 16, 805-816.

Calvert, G. A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., & David, A.S. (1997). *Activation of auditory cortex during silent lipreading.* Science, 276, 593-596.

Calvert, G.A., & Campbell, R. (2003). *Reading speech from still and moving faces: the neural substrates of visible speech.* Journal of Cognitive Neuroscience, 15, 57-70.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). The handbook of multisensory processing. Cambridge: MA:MIT Press.

Calvert, G.A., & Thesen, T. (2004). *Multisensory integration: methodological approaches and emerging principles in the human brain.* Journal of Physiology, 98, 191-205.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., Brammer, M.J., David, A. S., & Williams, S. C. R. (2001). *Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning).* Cognitive Brain Research, 12, 233-243.

Cathiard, M-A., Lallouache, M. T., Mohamadi, T., & Abry, C. (1995). Configurational vs. temporal coherence in audio-visual speech perception. K. Elenius and P. Branderud (Eds.), Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm: ICPhS, Vol. 3, 218–221.

Cathiard, M. A., Tiberghien, G., Tseva, A., Lallouache, M. T., & Escudier, P. (1991). Visual perception of anticipatory rounding during acoustic pauses: A cross-language study. *Proceedingss of the XIIth International Congress of Phonetic Sciences*, 19-24 Août, Aix-en-Provence, France, 4, 50-53.

Chandrasekaran C., Trubanova A., Stillittano S., Caplier A., & Ghazanfar A. A. (2009). *The natural statistics of audiovisual speech.* PLoS Computational Biology, 5(7), e1000436.

Cherry E. C. (1953). *Some experiments on the recognition of speech, with one and two ears.* JOURNAL OF THE ACOUSTICAL SOCIETY OF America, 25, 975-979.

Clark, H. H., & Wilkes-Gibbs, D. (1986) *Referring as a collaborative process.* Cognition 22, 1–39.

Cvejic, E., Kim, J., & Davis, C. (2010). *Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion.* Speech Communication, 52(6), 555-564.

Dambacher, M., Rolfs, M., Gollner, K., Kliegl, R., & Jacobs, A. M. (2009). *Event-related potentials reveal rapid verification of predicted visual input.* PloS One, 4(3), e5047.

D’Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). *The role of the motor system in discriminating normal and degraded speech sounds*. Cortex, 48(7), 882-887.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). *The motor somatotopy of speech perception*. Current Biology, 19(5), 381-385.

Davis, C., & Kim, J. (2006). *Audio-visual speech perception off the top of the head*. Cognition, 100(3), B21-B31.

DeLong K. A., Urbach T.P., & Kutas M. (2005). *Probabilistic word pre-activation during language comprehension inferred from electrical brain activity*. Nature Neuroscience, 8, 1117-1121.

Desjardins R. N., Rogers J., & Werker J.F. (1997). *An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks*. Journal of Experimental Child Psychology, 66, 85–110.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). *Understanding motor events: a neurophysiological study*. Experimental brain research, 91(1), 176-180.

Dodd, B., Oerlemens, M., & Robinson, R. (1989). *Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli*. Visible Language, 22, 59–77.

Driver, J., & Spence, C. (2000). *Multisensory perception: Beyond modularity and convergence*. Current Biology, 10(20), R731-R735.

Enns, J. T., & Lleras, A. (2008). *What's next? new evidence for prediction in human vision.* Trends in Cognitive Sciences, 12(9), 327-333.

Fadiga L., Craighero L., Buccino G., & Rizzolatti G. (2002). *Speech listening specifically modulates the excitability of tongue muscles: A TMS study.* European Journal of Neuroscience, 15, 399-402.

Flege, J. (1995). Second-language Speech Learning: Theory, Findings, and □ Problems. In W. Strange (Ed) Speech Perception and Linguistic Experience: □ Issues in Cross-language research. Timonium, MD: York Press, Pp. 229-273.

Fowler, C.A. (2004). Speech as a supramodal or a modal phenomenon. In Calvert GA, Spence C, Stein BE (eds.), The Handbook of multisensory processing.. The MIT Press, Cambridge. Pp 189–202.

Fowler, C. A. & Rosenblum, L. D. (1989). *The Perception of Phonetic Gestures.* Haskins Laboratories Status Report on Speech Research, l00, 102-117

Fowler, C. A., & Rosenblum, L. D. (1991). *The perception of phonetic gestures.* Modularity and the motor theory of speech perception, 33-59.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2012). *Seeing the initial articulatory gestures of a word triggers lexical access.* Language and Cognitive Processes, 1-17.

Friston, K., Kilner, J., & Harrison, L. (2006). *A free energy principle for the brain.* Journal of Physiology-Paris, 100(1), 70-87.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). *Action recognition in the premotor cortex.* Brain, 119(2), 593-609.

Ghazanfar, A. A., Chandrasekaran, C., & Logothetis, N. K. (2008). *Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys.* The Journal of Neuroscience, 28(17), 4457-4469.

Ghazanfar, A. A., & Schroeder, C. E. (2006). *Is neocortex essentially multisensory?* Trends in cognitive sciences, 10(6), 278-285.

Gilbert, D. T. & Wilson, T. D. (2009). *Why the brain talks to itself: sources of error in emotional prediction.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 364, 1335–1341.

Grant, K. W. (2001). *The effect of speechreading on masked detection thresholds for filtered speech.* The Journal of the Acoustical Society of America, 109(5), 2272-2275.

Grant K. W., & Seitz P. F. (2000). *The use of visible speech cues for improving auditory detection of spoken sentences.* The Journal of the Acoustical Society of America, 108(3), 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), 2677-2690.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: implications for theories of speech perception. Campbell et al. (Eds.), Hearing by eye, II: Advances in the psychology of speechreading and auditory-visual speech. (p. 3-25), Hove (UK): Psychology Press.

Green, K. P., & Miller, J. L. (1985). *On the role of visual rate information in phonetic perception.* Perception and psychophysics, 38(3), 269-276.

Green K. P., & Kuhl P. K. (1991). *Integral processing of visual place and auditory voicing information during phonetic perception*. Journal of experimental psychology: Human perception and performance, 17, 278–288.

Hardison, D. M. (2005). *Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment.* Applied Psycholinguistics, 26(4), 579.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). *Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English.* Speech communication, 47(3), 360-378.

Hickok, G., Costanzo, M., Capasso, R., & Miceli, G. (2011). *The role of Broca's area in speech perception: evidence from aphasia revisited.* Brain and language, *119*(3), 214-220.

Hickok, G., Holt, L. L., & Lotto, A. J. (2009). *Response to Wilson: What Does Motor Cortex Contribute to Speech Perception?* Trends in Cognitive Sciences, 13 (8), 330-331.

Hirata, Y., & Kelly, S. D. (2010). *Effects of lips and hands on auditory learning of second-language speech sounds.* Journal of Speech, Language and Hearing Research, 53(2), 298.

Holt, L. L., & Lotto, A. J. (2008). *Speech perception within an auditory cognitive science framework*. Current Directions in Psychological Science, 17(1), 42-46.

Iacoboni, M. (2008). *The role of premotor cortex in speech perception: evidence from fMRI and rTMS*. Journal of Physiology-Paris, 102(1), 31-34.

Jääskeläinen, I. P. (2010). *The role of speech production system in audiovisual speech perception.* The open neuroimaging journal, 4, 30.

James, W. (1890). The Principles of Psychology. New York: Dover Publications.

Jesse, A., and Massaro, D. W. (2010). *The temporal distribution of information in audiovisual spoken-word identification.* Attention, Perception, and Psychophysics, 72(1), 209-225.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). "*Putting the face to the voice": Matching identity across modality*. Current Biology, 13(19), 1709-1714.

Kauramäki, J., Jääskeläinen, I. P., Hari, R., Möttönen, R., Rauschecker, J. P., & Sams, M. (2010). *Lipreading and covert speech production similarly modulate human auditory-cortex responses to pure tones*. The Journal of Neuroscience, 30(4), 1314-1321.

Kayser, C., Petkov, C. I., Remedios, R., & Logothetis, N. K. (2012). "Multisensory influences on auditory processing: perspectives from fMRI and electrophysiology". M. M. Murray and M. T. Wallace (eds). The Neural Bases of Multisensory Processes (p. 99–113). Boca Raton, FL: CRC.

Keller P. E., & Koch I. (2008). *Action planning in sequential skills: relations to music performance.* Quarterly journal of experimental psychology, 61, 275-291

Kerzel, D., & Bekkering, H. (2000). *Motor activation from visible speech: evidence from stimulus response compatibility.* Journal of Experimental Psychology: Human Perception and Performance, 26(2), 634.

Kim, J. & Davis, C. (2003). *Hearing foreign voices: Does knowing what is said affect visual-masked-speech detection?* Perception, 32(1), 111-120.

Kim, J., Davis, C., & Krins, P. (2004). *Amodal processing of visual speech as revealed by priming.* Cognition, 93(1), B39-B47.

Klatt, D.H. (1980). *Speech perception. A model of acoustic-phonemic analysis and lexical access.* Journal of Phonetics, 8, 279-312.

Klucharev, V., Möttönen, R., & Sams, M., (2003). *Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception.* Brain and Cognitive Research 18, 65–75.

Kutas, M., DeLong, D. L, & Smith, N. J. (2011). A look around at what lies ahead: prediction and predictability in language processing. Bar, M. (Ed.), Predictions in the Brain: Using Our Past to Generate a Future: Using Our Past to Generate a Future. (p.190). Oxford University Press.

Lachs, L., Weiss, J. W., & Pisoni, D. B. (2000). *Use of partial stimulus information by cochlear implant users and listeners with normal hearing in identifying spoken words: Some preliminary analyses.* The Volta Review, 102(4), 303.

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). *It's not what you say but the way you say it: matching faces and voices.* Journal of Experimental Psychology: Human Perception and Performance, 33(4), 905.

Lansing, C. R., & McConkie, G. W. (1999). *Attention to facial regions in segmental and prosodic visual speech perception tasks.* Journal of speech, language and hearing research, 42(3), 526.

Lewkowicz D. J., & Hansen-Tift A. M. (2012). *Infants deploy selective attention to the mouth of a talking face when learning speech.* Proceedings of the National Academy of Sciences of the United States of America, 109, 1431–1436.

Liberman, A. M., & Mattingly, I. G. (1985). *The motor theory of speech perception revised.* Cognition, 21(1), 1-36.

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). *Reflections on mirror neurons and speech perception.* Trends in cognitive sciences, 13(3), 110-114.

MacDonald J., & McGurk H. (1978). *Visual influences on speech perception processes.* Perception and Psychophysics, 24 (3), 253-257.

MacLeod A., & Summerfield Q. (1987). *Quantifying the contribution of vision to speech perception in noise.* British journal of audiology, 21, 131-141.

Massaro, D.W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. (p. 66-74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press: Cambridge, MA.

Massaro, D. W., & Cohen, M. M. (1993). *The paradigm and the fuzzy logical model of perception are alive and well.* Journal of Experimental Psychology: General, 122, 115-124.

Massaro, D. W., & Stork, D. G. (1998). *Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech.* American Scientist, 86(3), 236-244.

Mattys, S.L., Bernstein, L.E., & Auer, E.T., (2002). *Stimulus-based lexical distinctiveness as a general word recognition mechanism.* Perception and Psychophysics*,* 64, 667–679.

McGurk, H., & MacDonald, J. (1976). *Hearing lips and seeing voices.* Nature, 264(5588), 746-748.

McClelland, J. L., & Elman, J. L. (1986). *The TRACE model of speech perception.* Cognitive Psychology, 18, 1-86.

McGettigan, C., Agnew, Z. K., & Scott, S. K. (2010). *Are articulatory commands automatically and involuntarily activated during speech perception?.* Proceedings of the National Academy of Sciences, 107(12), E42-E42.

McNeill, D. (2005). Gesture and Thought. Chicago: University of Chicago Press.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). *The essential role of premotor cortex in speech perception.* Current Biology, 17(19), 1692-1696.

Mengin, O., Flitton, A., Jones, C. R. G., de Haan, M., Baldeweg, T., & Charman, T. (2012). *Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing.* Autism Research 5, 39–48.

Miller, L. M., & D'Esposito, M. (2005). *Perceptual fusion and stimulus coincidence in the cross-modal integration of speech.* Journal of Neuroscience, 25(25), 5884-5893.

Miller, R. M., Sánchez, K., & Rosenblum, L. D. (2010). *Alignment to visual speech information.* Attention, Perception, and Psychophysics, 72(6), 1614-1625.

Mohammed, T., Campbell, R., Macsweeney, M. X., Milne, E., Hansen, P., & Coleman, M. (2005). *Speechreading skill and visual movement sensitivity are related in deaf speechreaders.* Perception-London, 34(2), 205-216.

Möttönen, R., Järveläinen, J., Sams, M., & Hari, R. (2005). *Viewing speech modulates activity in the left SI mouth cortex.* Neuroimage, 24(3), 731-737.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). *Processing of changes in visual speech in the human auditory cortex.* Cognitive Brain Research, 13, 417-25.

Möttönen, R., & Watkins, K. E. (2009). *Motor representations of articulators contribute to categorical perception of speech sounds.* The Journal of Neuroscience, 29(31), 9819-9825.

Munhall, K. G., & Buchan, J. N. (2004). *Something in the way she moves.* Trends in Cognitive sciences, 8(2), 51-53.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). *Visual prosody and speech intelligibility head movement improves auditory speech perception.* Psychological Science, 15, 133–137.

Murakami, T., Restle, J., & Ziemann, U. (2011). Observation-execution matching and action inhibition in human primary motor cortex during viewing of speech-related lip movements or listening to speech. *Neuropsychologia*, *49*(7), 2045-2054.

Nas, G. (1983). *Visual word recognition in bilinguals: Evidence for a cooperation between visual and sound based codes during access to a common lexical store.* Journal of Verbal Learning and Verbal Behavior, 22, 526–534.

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.

Navarra, J., & Soto-Faraco, S. (2007). *Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds.* Psychological Research, 71(1), 4-12.

Norrix, L. W., & Green, K. P. (1996). *Auditory-visual context effects on the perception of /r/ and /l/ in a stop cluster.* Journal of the Acoustical Society of America, 99, 2591.

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), 333-338.

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). *An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex.* PLOS One, 8(6), e68959.

Pallier, C., Bosch, L., & Sebastián-Gallés N. (1997). *A limit on behavioral plasticity in speech perception.* Cognition 64, B9–17.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). *The influence of native-language phonology on lexical access: exemplar-based vs. Abstract lexical entries.* Psychological Science, 12(6), 445–450.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I.P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). *Primary auditory cortex* *activation by visual speech: an fMRI study at 3 T.* Neuroreport, 16, 125-128.

Pickering, M. J., & Garrod, S. (2004). *Toward a mechanistic psychology of dialogue.* Behavioral and brain sciences, 27(2), 169-189.

Pickering, M. J., & Garrod, S. (2006). *Do people use language production to make predictions during comprehension?* Trends in cognitive sciences, 11, 105-110.

Pickering, M. J., & Garrod, S. (2013). *Forward models and their implications for production, comprehension, and dialogue.* Behavioral and Brain Sciences, 49-64.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). *Motor cortex maps articulatory features of speech sounds.* Proceedings of the National Academy of Sciences, 103(20), 7865-7870.

Ramus, F., Nespor, M., & Mehler, J. (2000). *Correlates of linguistic rhythm in the speech signal.* Cognition, 73(3):265-92

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. Dodd, B. and Campbell, R. (Eds.), Hearing by eye: The psychology of lip-reading. (p. 97-113). Hillsdale, NJ, England: Lawrence Erlbaum Associates Hillsdale: LEA.

Rizzolatti, G., & Craighero, L. (2004). *The mirror-neuron system.* Annual review of neuroscience, 27, 169-192.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). *Premotor cortex and the recognition of motor actions.* Cognitive Brain Research, 3(2), 131-141.

Robert-Ribes, J., Piquemal, M., Schwartz J-L. & Escudier, P. (1996). Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. Stork, D. G., and Hennecke, M. E. (Eds.), Speechreading by Humans and Machines. Berlin: Springer-Verlag.

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., & Hickok, G. (2011). *Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system.* Neurocase, 17(2), 178-187.

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. (2010). *Language identification from visual-only speech signals.* Attention, Perception, and Psychophysics, 72(6), 1601-1613.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. Pisoni, D. B., Remez, R. E. (eds.), The handbook of speech perception. (p. 51–78). Blackwell Publishing, Malden.

Rosenblum, L. D., Miller, R. M., & Sánchez, K. (2007). *Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects.* Psychological Science, 18(5), 392-396.

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, 17(5), 1147-1153.

Rouger, J., Fraysse, B., Deguine, O., & Barone, P (2007). *McGurk effects in cochlear implanted deaf subjects.* Brain Research, 1188, 87-99.

Sams, M., Aulanko, R., Hamalainen, M., Hari., R., Lounasmaa, O.V., Lu, S.T., & Simola, J. (1991). *Seeing speech: visual information from lip movements modifies activity in the human auditory cortex.* Neuroscience Letters, 127, 141-145.

Sams, M., Möttönen, R., & Sihvonen, T. (2005). *Seeing and hearing others and oneself talk.* Brain Research. Cognitive Brain Research, 23(2-3), 429-435.

Samuelsson, S., & Rönnberg, J. (1993). *Implicit and explicit use of scripted constraints in lip-reading.* European Journal of Cognitive Psychology, 5, 201-233.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). *Neuronal oscillations and visual amplification of speech.* Trends in cognitive sciences, 12(3), 106-113.

Schwartz, J., Berthommier, F., & Savariaux, C. (2004). *Seeing to hear better: Evidence for early audio-visual interactions in speech identification.* Cognition, 93(2), 69-78.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech. Campell, R., Dodd, B., and Burnham, D. (eds.), (p.85-108). Hove (UK): Psychology Press.

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, *10*(4), 295-302.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). *Identification of a pathway for intelligible speech in the left temporal lobe*. Brain, 123(12), 2400–2406.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). *Joint action: bodies and minds moving together*. Trends in cognitive sciences, 10(2), 70-76.

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012*). A bilingual advantage in visual language discrimination in infancy*. Psychological Science, 23(9), 994-999.

Sebastián-Galles, N., & Soto-Faraco, S. (1999). *Online processing of native and non-native phonemic contrasts in early bilinguals.* Cognition 72, 111–123.

Sekiyama, K., (1997). *Cultural and Linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects*. Perception and psychophysics, 59, 73–80.

Sekiyama, K., & Burnham, D. (2008). *Impact of language on development of auditory-visual speech perception.* Developmental science, 11, 306–320.

Sekiyama, K., & Tohkura, Y. (1993). *Inter-language differences in the influence of visual cues in speech perception.* Journal of Phonetics, 21, 427–444.

Shahin, A. J., Kerlin, J. R., Bhat, J., & Miller, L. M. (2012). *Neural restoration of degraded audiovisual speech.* NeuroImage, 60, 530–538.

Siva, N., Stevens, E. B., Kuhl, P. K., & Meltzoff, A. N. (1995). *A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions.* The Journal of the Acoustical Society of America, 98, 2983.

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.

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, 17(10), 2387-2399.

Smeele, P. M. T. (1994). Perceiving speech: Integrating auditory and visual speech. Unpublished doctoral dissertation, Delft University of Technology, The Netherlands.

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J. F., & Lewkowicz, D. J. (2012). The development of audiovisual speech perception. Bremner AJ, Lewkowicz DJ, Spencer C (eds.), Multisensory development. (p. 207–228). Oxford University Press, Oxford.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Galles, N., & Werker, J. F. (2007). *Discriminating languages by speech-reading*. Perception and psychophysics, 69, 218-231.

Spence, C. & Driver, J. (Eds.) (2004). Crossmodal space and crossmodal attention. Oxford: Oxford University Press.

Spivey, M., & Marian, V. (1999). *Crosstalk between native and second languages: Partial activation of an irrelevant lexicon.* Psychological Science, 10, 281–284.

Stein, B. E., & Meredith, M. (1993). The merging of the senses. Cambridge MA: MIT Press.

Stekelenburg, J. J., & Vroomen, J. (2007). *Neural correlates of multisensory integration of ecologically valid audiovisual events.* Journal of cognitive neuroscience, 19, 1964-1973.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. Wathen-Dunn W, (ed.), Models for the perception of speech and visual form: proceedings of a symposium. (p. 88–102). Cambridge, MA: MIT Press.

Sumby, W. H., & Pollack, I. (1954). *Visual contribution to speech intelligibility in noise.* Journal of Acoustic Society of America, 26(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. Dodd, B., and

Campbell, R. (eds.), Hearing by eye: the psychology of lip reading. (p. 3–51). Lawrence Erlbaum Associates, New York.

Swaminathan, S., MacSweeney, M., Boyles, R., Waters, D., Watkins, K. E., & Möttönen, R. (2013). *Motor excitability during visual perception of known and unknown spoken languages*. Brain and language, 126(1), 1-7.

ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & Van Atteveldt, N. (2013). *Audiovisual onset differences are used to determine syllable identity for ambiguous audiovisual stimulus pairs*. Frontiers in Psychology, 4, 331.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). *Anticipating upcoming words in discourse: evidence from ERPs and reading times*. Journal of experimental psychology. Learning, memory, and cognition, 31, 443-467.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). *Visual speech speeds up the neural processing of auditory speech*. Proceedings of the National Academy of Sciences of the United States of America, 102(4), 1181-1186.

Vatikiotis-Bateson, E., Eigsti, I. M., & Yano, S. (1994). *Listener eye movement behavior during audiovisual perception*. Journal of the Acoustical Society of Japan, 94-3, 679-680.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). *Eye movement of perceivers during audiovisualspeech perception*. Perception and Psychophysics, 60(6), 926-940.

Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.C., & Terzopoulos, D., (1996). The dynamics of audiovisual behavior in speech. Stork, D., Hennecke, M. (Eds.), Speech Reading by Humans and Machines. Vol. 150, (p. 221–232). NATO-ASI Series, Series F, Computers and Systems Sciences. Springer, Berlin.

Vatikiotis-Bateson, E., & Yehia, H., (1996). *Physiological modeling of facial motion during speech*. The Acoustical Society of Japan, H-96 65.

Venezia, J. H., Saberi, K., Chubb, C., & Hickok, G. (2012). *Response bias modulates the speech motor system during syllable discrimination.* Frontiers in psychology, 28, 3-157.

Watkins, K., & Paus, T. (2004). *Modulation of motor excitability during speech perception: the role of Broca's area.* Journal of Cognitive Neuroscience, 16(6), 978-987.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). *Seeing and hearing speech excites the motor system involved in speech production.* Neuropsychologia, 41(8), 989-994.

Weber, A., & Cutler, A. (2004). *Lexical competition in non-native spoken-word recognition.* Journal of Memory and Language, 50(1), 1-25.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Galles, N., & Werker, J.F. (2007). *Visual language discrimination in infancy*. Science, 316, 1159.

Werker, J.F., & Tees, R.C. (1984). *Phonemic and phonetic factors in adult cross-language speech perception.* The Journal of the Acoustical Society of America, 75(6), 1866–1878.

Werker, J. F., & Tees, R. C. (2002). *Cross-language speech perception: Evidence for perceptual reorganization during the first year of life.* Infant Behavior and Development, 25(1), 121-133.

Wilson, S. M., & Iacoboni, M. (2006). *Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception.* Neuroimage, 33(1), 316-325.

Wilson, S.M., Pinar Saygin, A., Sereno, M.I., & Iacoboni, M. (2004). *Listening to speech activates motor areas involved in speech production.* Nature, 7(7), 701-702.

Wolpert, D. M. (1997). *Computational approaches to motor control.* Trends in Cognitive Sciences, 1(6), 209-216.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). *Linking facial animation, head motion and speech acoustics.* Journal of Phonetics, 30(3), 555-568.

Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. (1998). *Quantitative association of vocal-tract and facial behaviour.* Speech Communication, 16, 23-43.

Yeung, H. H., & Werker, J. F. (2013). *Lip Movements Affect Infants' Audiovisual Speech Perception.* Psychological science, 24(5), 603-612.

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). *Temporal context in speech processing and attentional stream selection: a* *behavioral and neural perspective*. Brain and language, *122*(3), 151-161.

Annex. The time course of audio-visual phoneme identification: A high temporal resolution study.

Carolina Sánchez-García [1], Sonia Kandel[2,3], Christophe Savariaux[4] and Salvador Soto-Faraco[1,5]

1. Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain
2. Univ. Grenoble Alpes, CNRS LPNC UMR 5105, BP 47, 38040 Grenoble Cedex 9, France
3. Institut Universitaire de France, 103, bd Saint-Michel, 75005 Paris, France
4. Université Stendhal, GIPSA-lab, Dpt. Parole et Cognition (CNRS UMR 5216), BP 25, 38040 Grenoble Cedex 9, France
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Abstract

We addressed the temporal course of the contribution of visual and auditory speech information in phoneme identification in Spanish. The participants identified phonemes in disyllabic speech stimuli that were presented in Audiovisual, Audio-only and Visual-only condition. Speech tokens to be identified differed in one particular consonantal phoneme (/pa*C*a/), whose degree of visual and auditory saliency varied (i.e. /f/, /s/, /θ /, /r/ and /g/). We used a gating paradigm, in which the quantity of information presented was increasing in steps that could be as short as 10 ms thanks to high speed camera recordings. We estimated the amount of information necessary to correctly identify the target phoneme in each type of presentation. The results revealed that the timing of the identification process depends on the relative strength of each modality in terms of information (saliency). For phonemes where one modality was very informative by itself, the amount of information necessary to identify the phoneme in the bimodal condition equaled the most highly informative unimodal. In contrast, when the phoneme's visual and auditory information was not salient but complementary, audiovisual identification occurred earlier than in both unimodal conditions (i.e. /θ/). Therefore, the integration of vision and audition not always leads to a benefit, as it has been often argued. Our study suggests that the relative visual and auditory saliency of each

phoneme should be considered by classical audiovisual speech integration models.

Introduction

In face to face spoken communication, speakers provide both auditory and visual information to listeners. Many studies illustrate that, when available, the sound and the sight of the speaker contribute to speech perception and, in the end, to the understanding of the message (see Calvert, Spence, and Stein, 2004; Campbell, 2004). This is partly due to the fact that visual information plays an important role in phoneme identification and this accelerates the word recognition process (Fort, Spinelli, Savariaux, and Kandel, 2010; Fort et al., 2012). It has been extensively demonstrated that visual information about the speaker's oro-facial movements enhances speech comprehension in adverse conditions. For instance, when auditory information is degraded, such as in noisy conditions (Sumby and Pollack, 1954; Benoit, Mohamadi, and Kandel, 1994; Ross, Saint-Amour, Leavitt, Javitt, and Foxe, 2007), while interacting in a non-native language (e.g., Burnham, 1998; Navarra, and Soto-Faraco, 2007; Barrós-Loscertales et al., 2013), or in hearing-impaired listeners (Grant, Walden, and Seitz, 1998; Rouger, Fraysse, Deguine, and Barone, 2007). Even when visual and auditory information are not congruent in phonetic content, visual information has been shown to have a strong influence on the final percept, as happens during the McGurk effect (McGurk and MacDonald, 1976), where listening to the spoken syllable /ba/ while simultaneously watching the lip movements corresponding to the syllable /ga/ often results in the illusory perception of /da/. Finally, an illustrative example of the relevance of visual information is that it allows for discrimination between two languages when the perceiver is familiar with at least one of them (Soto-Faraco et al., 2007) and also when he/she is not (Sebastián-Gallés, Albareda-Castellot, Weikum, and Werker, 2012), being this ability already present very early in infancy (Weikum et al., 2007).

Although it is widely admitted that visual information plays an important role in speech perception, few studies have focused on its time course. This is surprising given the rich temporal structure that characterizes the speech signal. In addition, the few existing experiments that examined the temporal evolution of audiovisual integration during speech focused on the amount of presented information from each modality, using stimuli in AV phonological conflict (Munhall and Tohkura, 1998) or during tasks of word recognition (Jesse and Massaro, 2010). To our knowledge, only Jesse and Massaro considered the possibility that the audiovisual benefit could be modulated by the saliency of visual and auditory information. The present study investigated the time course of audiovisual phoneme identification by comparing the distribution of unimodal and bimodal information as speech unfolds in time. We hypothesized that the relative saliency of each modality regarding phoneme identity will regulate the evolution of the audiovisual benefit. We understand by saliency the perceptual weight from each modality in terms of informativeness. This means that highly visible articulatory movements or acoustic information such as the burst of a stop consonant may be so informative as to annul the audiovisual benefit. Visual information increases speech intelligibility because, in some cases, visual and auditory information are complementary (Miller and Nicely, 1955; Smeele , Sitting, and Heuven, 1992; Robert-Ribes, Schwartz, Lallouache, and Escudier, 1998). Visual information is useful to disambiguate phonologically close speech sounds, for instance, those that only differ in place of articulation. However, in other cases auditory information is more robust and the contribution of vision is limited. Yet, in other cases, both modalities provide redundant information, such as similarities in the dynamic pattern of the temporal properties of the speech stream (Campbell, 2008). Audiovisual benefit over auditory-alone listening conditions has been reported to be larger in the first case, when both modalities are complementary (Grant and Walden, 1996; Massaro, 1998; Summerfield, 1987), but the contribution of each sensory modality could be modulated by the visual and/or auditory saliency of each specific phoneme.

Some studies have investigated the contribution of visual information on phoneme discrimination. They proposed to group visual speech segments based on the discriminability from one another. This led to the concept of *viseme**, which classifies visual speech gestures associated to a group of phonemes that are highly confusable upon visual information (Fischer, 1968; Gentil, 1981), such as {/p/, /b/, /m/}. However, finer levels of detail can be extracted from visual information within a viseme class (Bernstein, Iverson, and Auer, 1997). For instance, Mayer, Abel, Barbosa, Black, and Vatikiotis□Bateson (2011) recently showed that some differences in articulation can be perceived, influencing the process of phoneme identification. In addition, while watching someone speaking, it is possible to extract visual cues related to acoustic features (Vatikiotis-Bateson, Munhall, Hirayama, Lee, and Terzopoulos, 1996) such as prosody, which allow for syllable identification (Munhall, Jones, Callan, Kuratate, and Vatikiotis-Bateson, 2004). Other studies capitalized on the possible role of the detailed phonological information carried in each sensory modality during audiovisual speech perception (e.g., Green and Miller 1985; Green and Kuhl, 1991; Benoît et al. 1994; Smeele, 1994; van Wassenhove, Grant, and Poeppel, 2005). For example, Benoît et al. (1994) examined phonemic identification of French consonants, /b, v, z, 3, r, l/, in different vocalic contexts /a, i, y/, presenting CVCVCV non-words in audiovisual and audio only conditions with various signal/noise levels of masking white noise. As in previous studies, their results revealed that phoneme identification was enhanced in the audiovisual condition with respect to the audio-only condition, and that the contribution of visual information increased with noise (see Altieri and Townsend, 2011 for a discussion). More interesting is that their data showed that audiovisual intelligibility for consonants depends on vocalic context. Due to co-articulation, consonant identification in the audiovisual presentation was better in an /a/ context (i.e., /CaCaCa/) than in an /i/ context, being the /y/ context the less intelligible. It is therefore difficult to speak about visual phoneme saliency without considering the context in which the phoneme is produced.

Although the studies discussed above are indeed informative about the possible differential contribution of visual information as a function of its phonological content, they do not provide a clear idea of the temporal profile of these influences. One must take into account that visual and auditory information are available at different moments in time, because speech unfolds in time (Escudier, Benoıˆt, Lallouache, 1990; Cathiard, Tiberghien, Tseva, Lallouache, Escudier, 1991; Smeele, 1994; Abry, Lallouache, and Cathiard, 1996; Chandrasekaran, Trubanova, Stillittano, Caplier, Ghazanfar, 2009). In this respect, the process of audiovisual integration must also be considered in terms of its temporal course. In natural speech, the auditory signal is the direct consequence of the movements of the speaker's articulators (visual signal), and therefore, both signals are in close relationship (Chandrasekaran et al. 2009). Many articulatory movements are directly visible to the speaker and often precede in time their acoustic correlates, sometimes over 100 ms (Escudier et al, 1990; Cathiard et al., 1991; Smeele, 1994; Abry et al. 1996). Some authors have recently argued that the information is exploited by the speech perception system as soon as it is available, so that phoneme identification based on the earlier arriving of the visual input can start before the auditory input is available. For instance, Escudier et al. (1990) showed that the rounding gesture for French /y/ in a /i/ →/y/ transition was detected visually well before the acoustic information about the identity of the phoneme was available (Escudier et al., 1990; Cathiard et al., 1991). For consonants, Smeele (1994) showed that plosive bilabials and labiodentals, both having a highly visually salient place of articulation, were identified earlier in time when the speaker's image was available, in addition to the sound. Also, seeing the lips close in preparation to articulate a /b/ sound can be enough visual information to narrow down a labial place of articulation even before the auditory information arrives (see also, Chandrasekaran et al., 2009). Indeed, some findings suggest that the extent to which the visual information can affect the time course of phoneme identification depends on the visual saliency of the articulatory features involved (e.g., Van Wassenhove et al., 2005; Arnal L. H., Morillon B., Kell C. A., Giraud A. L., 2009). In the present study we examined the temporal processing of auditory and visual information as they unfold in the speech signal. The stimuli had consonants of varying visual / auditory saliency. We investigated how saliency affected the audiovisual integration processes.

Previous studies addressing the question of how audiovisual integration evolves as speech unfolds in time and how the perceptual system copes with the different rates of information flow in each modality (Munhall and Tohkura, 1998; Jesse and Massaro, 2010) used the gating paradigm (Grosjean, 1980; 1996; Smeele, 1994; Munhall and Tohkura, 1998; Jesse and Massaro, 2010; Troille, Cathiard, and Abry, 2010). The gating paradigm (Grosjean (1980; 1996)) has been widely used in psycholinguistics, mostly in auditory word recognition (e.g. Warren and Marslen-Wilson, 1987, 1988). In this task, participants have to identify a speech token from a limited amount of information (i.e., phoneme, syllable …). The stimulus is segmented into fragments (i.e., the gates) and is presented gradually in steps of increasing length. This procedure aims at estimating the amount of information listeners need to recognize a word. The experimenter can therefore "track" the process of identification over time in a fine-grained manner. Smeele (1994) examined which phonetic features were extracted from visual and auditory modalities and at which moment in time. The stimuli were Dutch CV syllables increasing in fragments of 40 ms. Several consonants, varying in place of articulation, manner and voicing were presented (i.e. /b, d, p, t, k, v, f, z, s, m, n/), followed by the vowel /a/. Smeele showed that information about place of articulation was extracted from the visual signal, very early in time, while the manner of articulation and voicing were provided by the auditory modality, after the presentation of a longer fragment of stimulus. The audiovisual integration resulted by the combination of those features.

The gating paradigm was also used by Munhall and Tohkura (1998) to study the time course of audiovisual integration in the McGurk illusion. They presented the auditory stimulus /aba/ in its whole duration, dubbed onto fragments of the visual stimulus /aga/ presented in gates of increasing length (in steps of 33 ms). The results revealed that the proportion of fusion responses (i.e., the McGurk effect, /ada/) increased linearly with the accumulation of visual information. When testing the reverse case, that is dubbing auditory fragments of /aba/ onto a complete visual /aga/ stimulus, /ada/ responses did not grow steadily as information was accumulated, but increased abruptly at the gate corresponding to the burst of the acoustic /b/. Munhall and Tohkura suggested that visual and auditory information unfold at different rates, with vision accumulating steadily and audition providing information more transiently. In a recent study, Jesse and Massaro (2010) investigated the time course of unimodal and audiovisual information during a word recognition task. In their study, increasing fragments of a given CVC word were presented audiovisually, visually only or auditorily only. Participants had to identify the word that best matched with the word onset presented, choosing among 66 options. The stimulus set they chose represented all the possible initial English consonants. Globally, the authors found an improvement in performance during the presentation of the audiovisual condition, compared to the auditory only at each gate and across all tested phonemic categories. The amount of information that visual and auditory stimuli carry is distributed differently over time. In particular, visual information mostly concerns place of articulation, and it is available early on in time, whilst auditory information is accumulated more gradually across the presentation of the phoneme, and mostly transmits information about manner of articulation and voicing. This is in line with Smeele's (1994) results. These last conclusions contrast with Munhall and Tohkura's (1998) interpretation, where informativeness of the auditory signal was shown to vary rapidly and in a nonlinear manner. However, as already mentioned by Jesse and Massaro, the pattern found by Munhall and Tohkura might not be extensible to all phonemes (Smits, 2000; Smits et al. 2003), since their stimulus set was limited to three plosive consonants.

The study from Jesse and Massaro (2010) is extremely informative because it covered a wide range of phonemes, unlike many of the previous audiovisual speech perception studies using the Gating technique, which have often focused on one phoneme class (vowels /i/ and /y/ in Cathiard et al., 1991; plosive consonants in Munhall and Tohkura, 1998). This gives a complete picture of what is the contribution of audiovisual processing in everyday life speech perception, where the perceptual system must deal with the whole phonemic / visemic spectrum in a language's repertoire. However, it is noteworthy that Jesse and Massaro's study was carried out with synthetic auditory and visual speech. This limits the interpretation of the results because the articulatory information conveyed by the stimuli (auditory and visual) can go only as far as the authors managed to reproduce the features to create their stimuli. Synthetic speech might be sufficient to show that audiovisual information unfolds and affects speech perception at different rates in time, but we cannot generalize to natural speech, based on evidences such as the one mentioned earlier from Mayer et al. (2011), showing that very fine visual cues, maybe not captured at first glance, provide information that allow for discrimination between phonemes belonging to the same viseme*.

One of the problems that one faces when using naturally produced speech instead of synthetic materials in a gating paradigm is the standard temporal resolution of video-recording and playback equipment, of 25Hz. This obviously imposes a 40 ms limit on the size of Gates in audio-visual stimuli, which is often too slow to capture some informative events due to the quicker temporal variation of speech. For example, changes of 50 % during audiovisual identification of the /y/ vowel have been shown to occur in only 20 ms (Troille et al. 2010). Therefore, the temporal resolution of regular video recording might not be enough to explore speech perception with precision, because it would imply the loss of information about the course of the process. Indeed, auditory-only Gating studies use gates as short as 20 -30 ms (Grosjean, 1980) (note that auditory recording and playback can be done at a much higher temporal resolution with conventional equipment). Here, we were able to examine the temporal course of audiovisual, visual and audio phoneme identification with a gating paradigm in a "fine-grained perspective", by recording the speech stimuli with a high speed camera with a 100Hz sampling rate. We could therefore generate 10 ms gates, which significantly contrasts with the 40 ms of conventional cameras (25 images/ sec) used in previous studies. We believe it is important to use high temporal resolution to be able to capture as many changes as possible during audiovisual perception, given the quick temporal variation of speech.

In this experiment, we examined a relatively wide range of Spanish phonemes with varying auditory and visual saliency. We gated the stimuli visually, auditorily and audio-visually, so we could compare the temporal course of the perception for both modalities independently, as well as in combination. We used

/paCa/ natural speech sequences, where *C* corresponded to the target phoneme, which varied on the degree of visual and auditory saliency. We controlled for co-articulation effects by keeping the target phoneme always in a constant phonetic context, namely /a/. We chose this vowel because it is the context that less affects phoneme intelligibility (Benoît et al., 1994). We will therefore quantify the general benefit of audiovisual integration taking into account the specific properties that determine the saliency of each phoneme. We hypothesized that integration of visual and auditory information might lead to a benefit when both modalities are complementary. However, if one modality (either visual or auditory) carries enough information to allow for phoneme identification, audiovisual presentation will not constitute a gain when compared to the unimodal condition.

Method

Participants

Twenty-six native Spanish speakers (eight males, mean age 24.9 years) were tested. All participants reported normal audition and normal or corrected-to-normal vision, and were naive to the purpose of the experiment. Before the beginning of the experiment, participants gave their written informed consent.

Material

The stimuli consisted of natural productions of /pa*C*a/ video recordings. The stimuli differed in the second consonant slot only (i.e., *C*), which could be one of the five target phonemes. The target phonemes were three voiceless fricatives of different place of articulation and therefore, of varying visual saliency, according with the classification from confusion matrices (Summerfield, 1987): from high to low, /f/ labiodental, /θ/ inter-dental and /s/ alveolar; a flap consonant / r/, which shares an alveolar place of articulation with /s/, but unlike /s/, /r/ is voiced and finally, a voiced velar plosive consonant /g/. All the sequences started by the phoneme /p/ to make all the stimuli start with the mouth closed. Furthermore, the target phoneme was surrounded by the vowel /a/, because it is the vocalic context in which consonants are better identified (Benoît et al., 1994).

A frontal full-face view of a female speaker articulating the stimuli was recorded in a sound-proof room with a high speed camera (S-PRI color (1.3 Go); Vannier-Photelec) which allowed capturing images at 100Hz, thus allowing a playback resolution of 10 milliseconds. The audio signal was recorded simultaneously on the left channel of a Digital Audio Recorder (Tascam PMD-670) at 44,1 kHz. The synchronization signal issued from the camera was recorded on the right channel and used to post-synchronize together the audio and video signal. After recording, the videos were edited using VirtualDub program, in order to choose the number of images corresponding to each token of the stimuli and extract them from the visual stream. A total of four tokens were recorded for each phoneme, choosing the best three tokens for the experimental session, in base of the avoidance of blinking and other artifacts. The one left was used during the training session. The stimuli were presented in three conditions: audio-visual (AV), audio-only (A) and visual-only (V). For each stimulus, the point corresponding to the auditory onset of the target consonant segment (alignment point) was determined by visual inspection of the auditory wave, through a speech editor program (Praat software; Boersma and Weenink, 1996). For the audio-visual condition, the audio and visual streams belonging to each token were always synchronized according to the original sound track. We aligned the acoustic onset of the reference target consonant from the soundtrack of the /paCa/ stimuli with the corresponding image of the onset of the same central consonant in the visual stimulus. The same process was used for each of the 15 stimuli (i.e., 3 tokens x 5 phonemes). E-Prime software (2004) allowed for online loading and presentation of the images corresponding to each auditory stimulus in good synchrony (~1ms accuracy, checked with the Blackbox (Black Box Toolkit, 2004)). During the visual only condition, the sound track of the video-clip was silenced, and during the presentation of the auditory only condition, the video-track was replaced by a black screen with a white fixation cross in the center. The participants were presented with three otherwise equivalent tokens containing each target consonant (i.e., three versions of /pafa/, three of /pasa/ …). Three different tokens of each phoneme were used to discourage participants from using possible subtle acoustic / visual particularities of the stimuli in order to solve the task and focus, instead, on phonemic / visemic aspects.

Since the goal of the study was to examine the evolution of the identification of the target phoneme, the experimental procedure was an adaptation of the Gating technique (Grosjean, 1996). We thus segmented the video recordings in 18 fragments or gates. The *alignment point* corresponded to the auditory release of the target consonant and was used as reference fragment (gate 12; see **Figure 1**). The gates were created by adding or subtracting signal in 10 ms gates around the alignment point and larger ones in the extremities (20 and 30 ms). In total the 18 gates were created with the following cut points: -150 ms, -120 ms, -100ms, -80ms, -70ms, -60ms, -50ms, -40ms, -30ms, -20ms, -10 ms, 0, +10ms, +20ms, +30ms, +50ms, +70ms and full stimulus (variable length), where 0 corresponds to the *alignment point.* The idea underlying this segmentation was to get fine-grained information (10 ms) around the critical information concerning the identity of the target consonant. The presentation could be speeded up before and after this point.

Procedure

The participants were instructed to watch the lips of the speaker and to listen to the speech fragment. The gates of a given stimulus token were presented in an incremental fashion until the end of the sequence (see **Figure 1**).We will refer to each presentation, incrementing length in one gate with the term *trial.* After the presentation of each trial (in its auditory, visual or audiovisual version) the participants had to indicate what they perceived by choosing among five possibilities that appeared written in the screen of the computer (e.g. 1. Pafa; 2. Paga; 3. Pasa; 4. Para; 5. Paza).

Participants gave their answer about the identity of the percept by pressing the corresponding number in the keyboard. They were encouraged to respond after every trial, even if for certain stimuli they had to guess, and did not have limitation of time to respond. Then, they were asked to evaluate on a 9-point scale their confidence in the response (1=totally uncertain; 9=totally confident). After the confidence rating, the procedure was repeated for the next trial of the same stimulus token. Once a token had been presented completely (i.e., until the last trial), the participants pressed the Space bar to continue onto the next token. Participants were presented with a total of 45 tokens (15 per modality). The three conditions, audio-visual, visual only and auditory only were presented in different blocks, with block order counterbalanced across participants. There were a total of 270 trials, corresponding to each of the 18 trials by each of the 15 tokens/modality. The order of tokens presentation within a block was randomized, but as mentioned earlier, the presentation of each token consisted of the serial presentation of the trials corresponding to it, incrementing in one gate duration. The order of the numbered list of answers displayed in the screen was randomized for each token. It did not change from trial to trial of the same token, but the order was renewed every time a new sequence of a token started.

Participants watched the stimuli on a 100 Hz video monitor (Acer GD245HQ) and the acoustic signal was presented through headphones (Sennheiser HD 435 Manhattan) at a comfortable, constant, sound pressure level. E-Prime Software was used to control stimulus presentation and record the participant's responses from the keyboard. Participants were tested individually in a sound attenuated experimental booth, seated approximately 50 cm from the video monitor with a keyboard placed on the table in front of them.

After instructing the participants about the task (written and verbally) they were shown a list with the five different stimuli they would be presented with during the experiment (in order to decrease the possibility of lexical influence at the beginning of the session), and then they ran a short training session (with different tokens from the ones used for the experimental session) to familiarize themselves with the paradigm. The duration of the training plus the experimental session was approximately 45 minutes.

Data analysis

For each stimulus, we measured the **Isolation Point (IP)** and the **Recognition point (RP)**. These are the usual variables used in the gating paradigm (Grosjean, 1996; Sebastián-Gallés and Salvador-Soto, 1999). Following the definitions proposed by Grosjean (1996), the IP is defined as the amount of stimulus information needed to identify the stimulus (without any change in response thereafter), while the RP corresponds to the amount of stimulus information needed to reach and maintain a particular confidence rating when the correct response is given during the identification (isolation) of the phoneme (in our case, a rate of 8 or higher).

Results

Results from Isolation Point and Recognition Point scores are presented in **Table 1 and Table 2**. To test whether there was an advantage of the audiovisual presentation on phoneme identification we conducted Student's t-test comparisons between the Isolation Point of the three experimental conditions (A, V and AV). In the case of /f/, the scores for the audiovisual condition were equivalent to the ones observed in the visual-only presentation. It was identified 30 ms earlier visually than in the auditory modality. For /g/ and /r/ the results for the audiovisual modality were equivalent to the audio-only condition. They both presented an advantage respect to the visual modality of 60 ms for /g/ and 50 ms for /r/. For /θ/, the results revealed a significant audiovisual benefit of 20-30 ms over the unimodal presentations. A significant difference between the visual unimodal and the audiovisual modalities was also observed for /s/, but instead of leading to a benefit, the combination of both modalities resulted in a significant delay of 10 ms during the identification.

Comparing across phonemes, the Isolation Point was reached significantly earlier for the "auditory" phonemes (/g/ and /r/) than for the "visual" phonemes (/f/ and /s/) or when the integration of auditory and visual information was the best option (θ), indicating that when the auditory modality is very informative it is available even before than the visual information.

Recognition point scores showed that the audiovisual modality was the more reliable for the phonemes in which visual modality was the predominant modality. In the case of /g/ and /r/, however, the auditory modality was trusted to the same degree as the audiovisual. **See Figure 2.**

We explored in more detail the temporal evolution of the identification process by calculating the **percentage of correct responses** for each phoneme, at each gate, across participants (**Figure 3)** so we could capture more transient variations over time.

To see at which moment in time a modality played a significant role over the others, we conducted Student's t-tests, comparing the percentage of correct responses in each of the three conditions at each gate (**Table 3**).

For /f/ we observed that visual and audiovisual identification curves overlapped. The supplementary auditory information in the AV condition did not lead to a benefit in the identification process. Despite the visual appearance of differences between visual and audiovisual performance, they did not reach statistical significance at any point in time (all $t<1$), according with the Isolation point analyses.

The distribution of visual and auditory information during the identification of /g/ and /r/ followed a similar pattern. That is, the auditory modality provides enough information for phoneme identification. The auditory curve overlapped with the audiovisual curve, so visual information does not make a significant contribution to the identification processes. Any significant difference was found when comparing visual vs. auditory and visual vs. audiovisual (all t<0.01, from gate 5 for /g/ and 8 for /r/, temporally corresponding to -70 ms and – 40 ms from the alignment point). This profile is congruent with what we found using the Isolation Point scores.

For /θ/ we observed a systematic AV advantage with respect to the unimodal conditions. At early and late points in time the audiovisual scores overlapped with the best unimodal (auditory and visual, respectively). At the middle gates, when visual modality started to be informative (around gate 10), visual and auditory modalities contributed equally and the curves crossed at gate 12 (coinciding with the alignment point), leading to a significant audiovisual enhancement of 20 %.

For /s/ the scores in the visual modality were globally higher than in the auditory modality. However, in contrast with the other phonemes with clear unimodal dominance, the audiovisual scores were significantly lower than the best unimodal presentation (visual). The audiovisual curve is higher than the auditory curve but lower than the visual curve, according with the Isolation Point scores.

Visual, auditory and audiovisual phonetic similarity between phonemes was established by calculating the number of responses of each type for each trial. With these data, confusion matrices were constructed for the three modalities of presentation. On the basis of the confusion matrices we can observe the frequency with which one phoneme was confused with another one. See **Table 4**.

Remarkable from the pattern of phonemic confusion is that auditory identification of /f/ by itself is poor (the 67.3 % participants were not able to identify the phoneme only auditorily), mostly because based on auditory information alone, /f/ is frequently confused with /θ/, even when presented in full (this is in fact a common confusion in everyday life spoken communication, when lexical context is not biasing). Based on visual information alone, participants failed in the discrimination of /g/ on the 56 % of the cases. It is maybe related to the fact that during the production of this phoneme, vibration of the vocal cords reflects mostly in the acoustic signal, not in the optical signal (Yehia, Rubin, and Vatikiotis-Bateson, 1998).

Discussion

We addressed the temporal course of the contribution of visual and auditory speech information in phoneme identification in Spanish. The objective was to reveal possible patterns of audiovisual integration as information from each sensory source accumulates in time and varies in degree of saliency. The participants identified phonemes embedded in disyllabic natural speech stimuli that were presented in Audiovisual, Audio-only and Visual-only condition. The results revealed that when visual and auditory information provided relevant cues about phoneme identity (visual for /f/; auditory for /g/ and /r/), the audiovisual identification curves overlapped the dominant unimodal source of information. So audiovisual integration -if present- did not lead to any benefit over the best unimodal condition. At variance with this pattern, for the phoneme /s/ we observed that the visual modality alone allowed for correct identification well before the audiovisual presentation. This indicates that the presence of auditory information slowed the identification process, suggesting that audiovisual integration led to a less efficient performance. Finally, for /θ/ the results yielded an audiovisual benefit with respect to the unimodal conditions. That is, audiovisual integration led to earlier identification as compared to the best unimodal condition. This might suggest that the features provided by both modalities complemented each other.

Taken together, our results did not show a systematic audiovisual benefit, as it has been reported in previous studies (Smeele, 1994; Steven and Massaro, 2004; Jesse and Massaro, 2010). They revealed instead that the interactions between visual and auditory information are variable. Audiovisual information may sometimes enhance identification (e.g., /θ/) but also disrupt phoneme processing (e.g., /s/) depending on the information provided by each sensory modality. In a study using a similar gating paradigm, Jesse and Massaro (2010) showed that performance during word recognition was characterized by a size-constant audiovisual benefit over the auditory only condition at each gate during the presentation of the stimuli. These differences between their results and ours could be explained by differences in the experimental design. The most important difference is that, in Jesse and Massaro's study the way used to construct the gates led to variations across the temporal placement of the corresponding gates across different phonemes. That is, the gates were made based on the duration of the initial consonant and vowel of their stimuli in order to have the same number of gates for all the presented words (in such a way that at the third gate the first consonant had been completely presented and at the sixth gate both, initial consonant plus vowel had been completely presented). This leads to gates of different lengths across stimuli. Perhaps more importantly, in Jesse and Massaro's study, the first gate of each stimulus carried a high amount of information. Identification scores of phonemes belonging to the same viseme class at the first gate were very close to the ones they obtained when presenting the full stimulus (mean 29% vs. 36%, respectively across stimuli). It is therefore likely that Jesse and Massaro's study reveals only a small window of the whole temporal course of information accumulation, leaving out some highly relevant instants of phoneme processing. This could be due to the fact that the duration of their gates was much longer than in our study. As said, thanks to our high speed camera, we could generate 10 ms gates and collect very fine grained data on the phoneme identification process. For example, in the present study the identification of /g/ and /r/ in the auditory condition occurred very early in time (around gate 5, -70ms), while in the audiovisual modality it occurred later on (gates 11-13, -10 to 10 ms). This early processing could not be examined by Jesse and Massaro and indeed, they did not observe any unimodal benefit over the audiovisual at any time. It is noteworthy that Smeele

(1994) also supported the existence of an AV benefit for bimodal vs. unimodal perception, arguing that visual information about place of articulation and auditory information about manner and voicing, improves intelligibility when combined during the perception of plosive and fricative phonemes. Again, temporal resolution differences -their gates were 40 ms long- could explain divergences between the two data collections. Smeele's gates could have been too long to observe the processes occurring at the earlier stages of phoneme identification.

Our results also indicated that phoneme identification curves in the A, V and AV conditions were modulated by the perceptual saliency of visual and auditory information. Saliency has been shown to be directly related to the extent to which visual information has an influence in the time course of phoneme identification (Van Wassenhove et al. 2005).

Highly visible phonemes have been shown to present less acoustic differences, and vice-versa (Summerfield, 1987). The visual signal conveys information about place of articulation (Smeele, 1994; Jesse and Massaro, 2010), frication and tongue-tip movement (i.e. the tip of the tongue moves visibly during production). In our study, visual information played an important role during the identification or /f/, /s/ and /θ/. This is in line with Jesse and Massaro's (2010) study as well as Smeele's (1994). Their results revealed that visual information plays an important role in the identification of this kind of fricative phonemes.

The differences in the perception of / θ / and /s/, even when both share most of the features, might be due to differences in the weight/quantity of the features transmitted by the visual modality, and its combination in each case. For / θ / the arrival of auditory information about tongue-tip movement and dental adduction might be combined with visual information about place of articulation, leading to an audiovisual benefit when both of these features are detected. Although a minimal amount of information from both modalities is required to observe an audiovisual enhancement during speech perception (Grant and Walden, 1996; Massaro, 1998), the benefit of audiovisual integration seems to be larger when unimodal performance is low (Massaro, 1998). That is, visual information contributes to an audio-visual benefit substantially bigger when auditory input is degraded (Altieri and Townsend, 2011) or it is ambiguous (Rouger et al., 2007). In fact, the benefit is particularly important in intermediate noise levels (i.e., signal-to-noise levels around -10 dB to -12dB), which are the most common in everyday life (Ross et al., 2007). In our study, it might be that each unimodal modality by itself is not sufficiently clear to arrive to identification, and therefore, a benefit is observed as consequence of the combination of both. In the case of /s/, the audiovisual identification scores were lower than the visual identification scores. Indeed, /s/ has been shown to be special in some acoustic aspects. For example, during its production, female frequency spectra tend to be higher than the male spectra (Schwartz, 1968; Stelmachowicz, P. G., Pittman, A. L., Hoover, B. M., Lewis, D. E., and Moeller, M.

P., 2004) and it is the most likely phoneme to be impaired in any speech pathology in any language (Luchsinger, Godfrey, Arnold, and Baar, 1965).

The patterns of the curves presented in Figure 3 revealed that the visual identification in the phonemes /s/ and /θ / is continuously increasing; the results are less clear for /f/. Interestingly, these are the cases in which visual information plays an important role during the audiovisual integration, meaning that visual information is received in an increasingly smooth manner. Instead, the curves observed for the phonemes /g/ and /r/, which are very salient acoustically, are defined by more discontinuous informative steps. In the case of the phoneme /s/, while the visual only curve is accumulative, the graph is characterized by the attraction of the audiovisual curve towards the dynamics of the auditory information (characterized by peaks), delaying the identification process in comparison with the visual information alone. This is interesting, because it suggests that in this case, audio-visual integration would be an average of the two modalities, instead of overlapping the best unimodal, thus effectively leading to interference with respect to the best unimodal (i.e., visual alone). The patterns found indicate that the information that the acoustic signal provides is not homogenous, but there are specific instants highly informative, which give the cues for phoneme identification. This is in agreement with the pattern of results reported by Munhall and Tohkura's (1998) gating study with the McGurk effect, but here extending it to a bigger range of phonemes. Munhall and Tohkura showed that the visual speech signal provided continuous and increasingly available information to the perceiver. For the auditory contribution, instead, they observed discrete peaks that had different degrees of impact during perception. For example, the consonant burst provided highly salient acoustic information and a silence during articulation of bilabial plosives gives information about the duration of the consonant.

However, as mentioned already in the introduction section, when looking to Jesse and Massaro's (2010) results, visual information was available early on in time, from the first moments of the presentation, while the auditory one was accumulated over time. These results are in principle contradictory to Munhall and Tokura's and our own results, showing, one more time, that the course of information is a complex process, highly variable depending on the presented stimuli. One must bear in mind, that the comparison between these studies is compromised by the large methodological differences, including the use of real vs. synthesized stimuli, the study of audiovisual illusions (McGurk) vs. audiovisually congruent events, the different temporal resolution and the different temporal window explored.

In sum, our results show that perceptual saliency from the different sensory modalities affects the course of phoneme discrimination and it constitutes a determining factor for an audiovisual benefit. They support the view that visual and auditory information are processed in a parallel interactive manner as information is available. Some of visual and auditory features would be processed early on, while integration of higher order features will occur later on during the process of phoneme identification (van Wassenhove et al. 2005, Jesse and Massaro, 2010, Altieri and Towsend, 2011). Altieri and Townsend (2011) proposed that the processing of auditory and visual information occurs in an interactive parallel manner, where a decision is taken when the system has enough information from one of the two modalities to reach identification. Moreover, Altieri and Townsend found that the greatest audiovisual benefit occurs at -18 dB. However when auditory information was perfectly clear, as in the present study), visual information disrupted more than benefited audiovisual performance. Our study is in agreement with this observation. In the audiovisual presentation of highly salient phonemes in one modality (auditory for /g/ and /r/; visual for /f/ and /s/), the addition of information from the other (less informative) modality did not yield to any benefit. Then, the most informative cue has the greatest impact on the judgments. The identification process relies on the information carried on by the strongest unisensory modality. It is likely that a decision would be taken before the processing of the information from both sensory modalities is completed, leading to phoneme categorization based on unimodal information, supporting the existence of a decision stopping rule (Altieri and Townsend 2011). However, when unimodal information is not strong enough to allow for a decision, participants wait to have complementary information (about different features) from the other modality, leading in this case the audiovisual integration to a benefit (i.e. /θ/).

In conclusion, our results suggest that audiovisual integration may not always be the best option during the process of phoneme identification in natural speech. To a large extent, it depends on the information about the features that are transmitted by visual and auditory modalities, as well as on its complementary, in terms of saliency. The present study suggests that theories of speech perception should take into account the specificities of auditory and visual saliency of the percept. Concretely, we have shown the importance of the saliency properties regarding the audiovisual benefit during speech perception.

References

Abry, C., Lallouache, M. T., & Cathiard, M. A. (1996). *How can coarticulation models account for speech sensitivity to audio-visual desynchronization?.* NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES, 150, 247-256.

Altieri, N., & Townsend, J. T. (2011). *An assessment of behavioral dynamic information processing measures in audiovisual speech perception.* Frontiers in psychology, 2.

Arnal L. H., Morillon B., Kell C. A., & Giraud A. L. (2009). *Dual neural routing of visual facilitation in speech processing.* Journal of Neuroscience, 29: 13445-13453.

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Ávila Rivera, C., & Soto-Faraco, S. (2013). *Neural correlates of audiovisual speech processing in a second language.* Brain and language, 126(3), 253-262.

Benoit, C., Mohamadi, T., & Kandel, S. (1994). *Effects of phonetic context on audio-visual intelligibility of French.* Journal of Speech, Language and Hearing Research, 37(5), 1195.

Bernstein, Z. E., Iverson, P., & Auer Jr, E. T. (1997). *Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception.* In Audio-Visual Speech Processing: Computational and Cognitive Science Approaches.

Boersma, P. & Weenink, D. (2007) Praat (Version 4.5.25) [Software]. Latest version available for download from www.praat.org.

Burnham, D. K. (1998). Language specificity in the development of auditory-visual speech perception. Campbell, R., Dodd, B. and Burnham, D. (Eds.), Hearing by eye II: advances in the psychology of speechreading and auditory–visual speech (pp. 27–60). Hove, UK: Psychology Press.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). The handbook of multisensory processing. Cambridge: MA:MIT Press.

Campbell, R. (2004). Audiovisual speech processing. In K. Brown (Eds.), The Encyclopedia of Language and Linguistics (2nd edition). Oxford, UK: Elsevier.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1493), 1001-1010.

Chandrasekaran C., Trubanova A., Stillittano S., Caplier A., & Ghazanfar A. A. (2009). *The natural statistics of audiovisual speech.* PLoS Computational Biology, 5(7) e1000436.

Cathiard, M. A., Tiberghien, G., Tseva, A., Lallouache, M. T., & Escudier, P. (1991). Visual perception of anticipatory rounding during acoustic pauses: A cross-language study. Procedures of the XIIth International Congress of Phonetic Sciences, 19-24 Août, Aix-en-Provence, France, 4, 50-53.

Escudier, P., Benoıˆt, C., & Lallouache, M.-T. (1990). Identification visuelle de stimuli associe´s a` l'opposition /i/-/y/: e´tude statique. Colloque de physique, supple´ment au n_ 2, tome 51, 1er Congre`s Franc_ais d'Acoustique, C2-541-544.

Fischer, C. G. (1968). *Confusions among visually perceived consonants.* Journal of speech and hearing research, pp. 796-804.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). *The word superiority effect in audiovisual speech perception.* Speech Communication, 52(6), 525-532.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2012). *Seeing the initial articulatory gestures of a word triggers lexical access.* Language and Cognitive Processes, 1-17.

Gentil, M. (1981). *Etude de la perception de la parole: lecture labial et sosies labiaux*. IBM France.

Grant, K. W., & Walden, B. E. (1996). *Evaluating the articulation index for auditory–visual consonant recognition*. The Journal of the Acoustical Society of America, 100, 2415.

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 Pt 1), 2677-2690.

Green, K. P., & Miller, J. L. (1985). *On the role of visual rate information in phonetic perception.* Perception and psychophysics, 38(3), 269-276.

Green K. P., & Kuhl P. K. (1991). *Integral processing of visual place and auditory voicing information during phonetic perception.* Journal of experimental psychology. Human perception and performance, 17:278–288.

Grosjean, F. (1980). *Spoken word recognition processes and the gating paradigm.* Perception and Psychophysics, 28(4), 267-283.

Grosjean, F. (1996). *Gating.* Language and cognitive processes, 11(6), 597-604.

Jesse, A., & Massaro, D. W. (2010). *The temporal distribution of information in audiovisual spoken-word identification.* Attention, Perception, and Psychophysics, 72(1), 209-225.

Luchsinger, R., Godfrey E.. Arnold, & Baar, E. (1965). Voice, speech, language. Wadsworth.

Mayer, C., Abel, J., Barbosa, A., Black, A., & Vatikiotis Bateson, E. (2011). *The labial viseme reconsidered: Evidence from production and* *perception.* The Journal of the Acoustical Society of America, 129, 2456.

Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press: Cambridge, MA.

McGurk, H., & MacDonald, J. (1976). *Hearing lips and seeing voices.* Nature, 264(5588), 746-748.

Miller, G. A., & Nicely, P. E. (1955). *An analysis of perceptual confusions among some English consonants.* The Journal of the Acoustical Society of America, 27, 338.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). *Visual prosody and speech intelligibility head movement improves auditory speech perception.* Psychological Science, 15, 133–137.

Munhall, K. G., & Tohkura, Y. (1998). *Audiovisual gating and the time course of speech perception.* The Journal of the Acoustical Society of America, 104, 530.

Navarra, J., & Soto-Faraco, S. (2007). *Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds.* Psychological Research, 71(1), 4-12.

Navarra, J., Sebastián-Gallés, N., & Soto-Faraco, S. (2005*). The perception of second language sounds in early bilinguals: new evidence from an implicit measure*. Journal of Experimental Psychology: Human Perception and Performance, 31(5), 912.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). *Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise.* The Journal of the Acoustical Society of America, 103, 3677.

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), 1147-1153.

Rouger, J., Fraysse, B., Deguine, O., & Barone, P (2007). *McGurk effects in cochlear implanted deaf subjects.* Brain Research, 1188, 87-99.

Schwartz, M. F. (1968*). Identification of speaker sex from isolated, voiceless fricatives.* The Journal of the Acoustical Society of America, 43, 1178.

Sebastián-Galles, N., & Soto-Faraco, S. (1999). *Online processing of native and non-native phonemic contrasts in early bilinguals.* Cognition 72:111–123.

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). *A bilingual advantage in visual language discrimination in infancy*. Psychological Science, 23(9), 994-999.

Smeele, P. M. T. (1994). Perceiving speech: Integrating auditory and visual speech. Unpublished doctoral dissertation, Delft University of Technology, The Netherlands.

Smeele, P. M., Sittig, A. C., & Heuven, V. J. V. (1992). Intelligibility of audio-visually desynchronised speech: Asymmetrical effect of phoneme position. In Second International Conference on Spoken Language Processing.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Galles, N., & Werker, J. F. (2007). *Discriminating languages by speech-reading.* Perception and psychophysics, 69:218-231.

Steven, K., & Massaro, D. W. (2004). *Audiovisual speech gating: Examining information and information processing.* Cognitive Processing, 5(2), 106-112.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. Dodd, B., and Campbell, R. (eds.), Hearing by eye: the psychology of lip reading. (p. 3–51). Lawrence Erlbaum Associates, New York.

Sumby, W. H., & Pollack, I. (1954). *Visual contribution to speech intelligibility in noise.* Journal of Acoustic Society of America, 26(2), 212-215.

Stelmachowicz, P. G., Pittman, A. L., Hoover, B. M., Lewis, D. E., & Moeller, M. P. (2004). *The importance of high-frequency audibility in the speech and language development of children with hearing loss.* Archives of Otolaryngology—Head and Neck Surgery, 130(5), 556.

Troille, E., Cathiard, M. A., & Abry, C. (2010). *Speech face perception is locked to anticipation in speech production.* Speech Communication, 52(6), 513-524.

Warren, P., & Marslen-Wilson, W. (1987). *Continuous uptake of acoustic cues in spoken word recognition.* Perception and Psychophysics, 41(3), 262-275.

Warren, P., & Marslen-Wilson, W. (1988). *Cues to lexical choice: Discriminating place and voice.* Perception and Psychophysics, 43(1), 21-30.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Galles, N., & Werker, J.F. (2007). *Visual language discrimination in infancy.* Science, 316:1159.

Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.C., & Terzopoulos, D., (1996). The dynamics of audiovisual behavior in speech. Stork, D., Hennecke, M. (Eds.), Speech Reading by Humans and Machines. Vol. 150, (p. 221–232).

NATO-ASI Series, Series F, Computers and Systems Sciences. Springer, Berlin.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences of the United States of America, 102(4), 1181-1186.

Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. (1998). *Quantitative association of vocal-tract and facial behaviour*. Speech Communication, 16, 23-43.

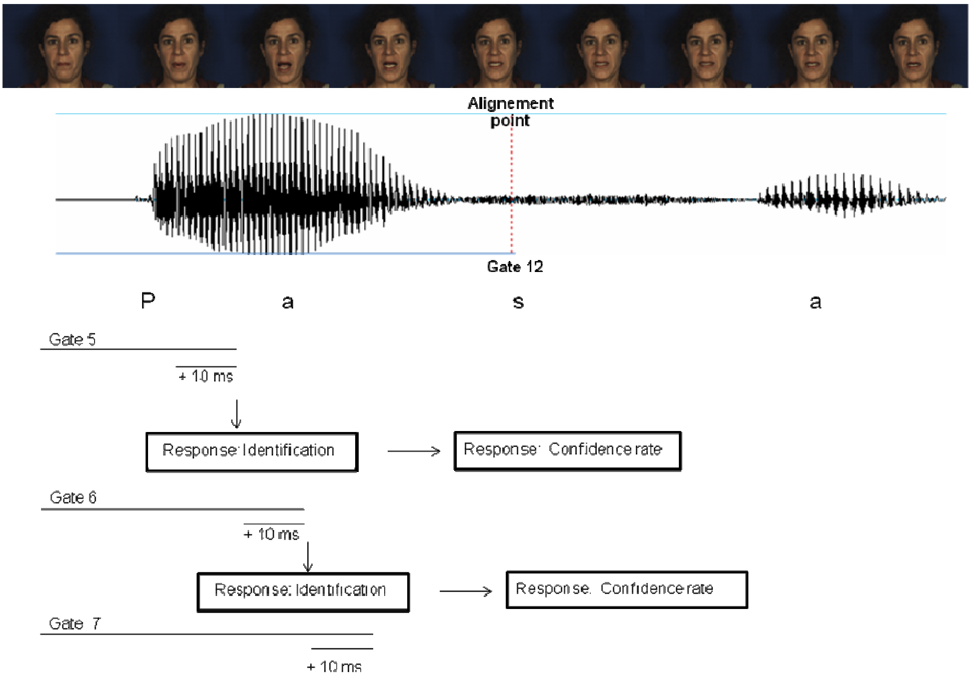


Figure 1. Experimental procedure.

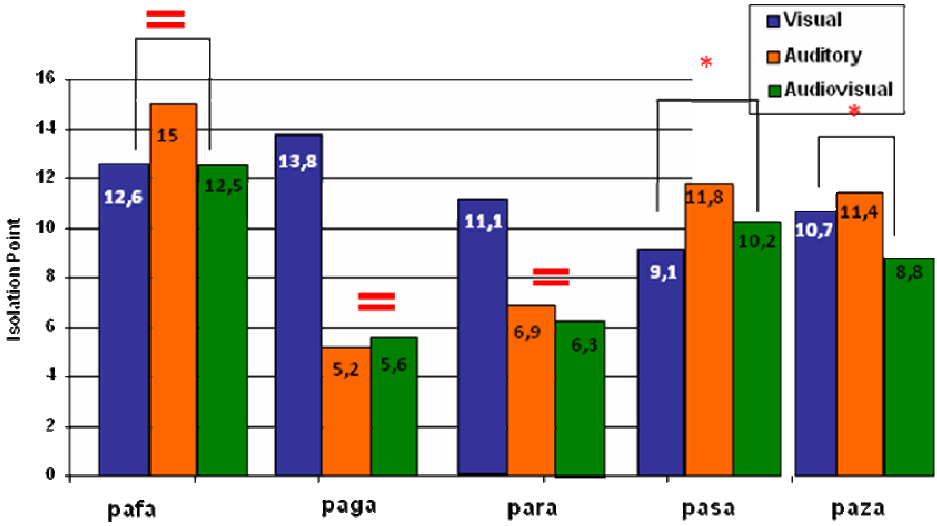


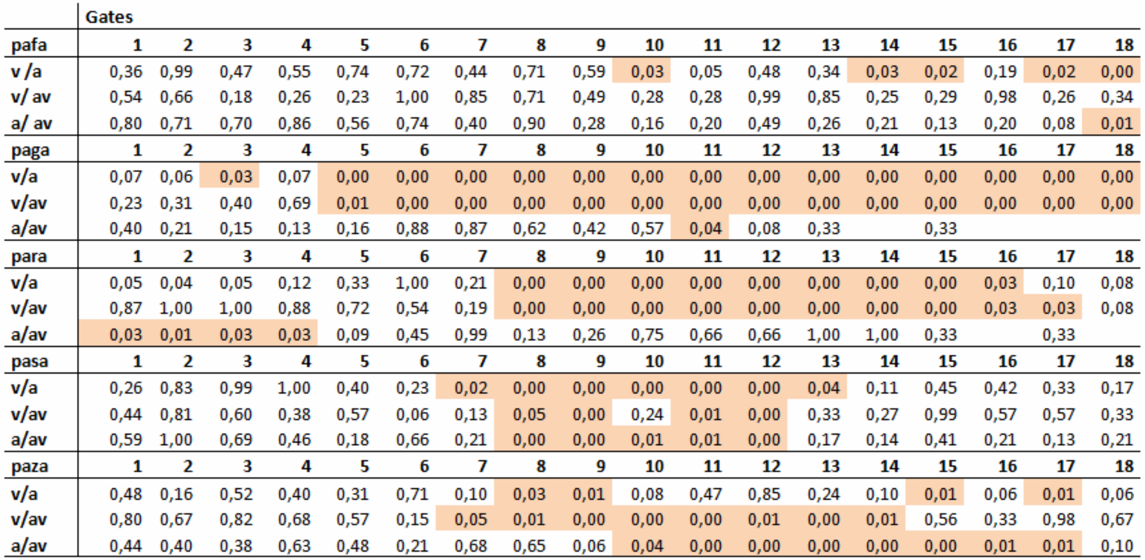
Figure 2. Mean Isolation Point for each phoneme and modality.

	Visual	Auditory	AV	Visual vs. AV		Auditory vs. AV		Visual vs. Auditory	
				t-value	p-value	t-value	p-value	t-value	p-value
Pafa	12.6	15	12.5	0.139	n/s	4.36	<0.01	-0.36	<0.01
Paga	13.8	5.2	5.6	11.8	<0.01	-0.73	n/s	10.9	<0.01
Para	11.1	6.9	6.3	7.36	<0.01	1.36	n/s	6.62	<0.01
Pasa	9.1	11.8	10.2	-2.89	<0.01	3.25	<0.01	-5.17	<0.01
Paza	10.7	11.4	8.8	3.80	<0.01	4.34	<0.01	-1.09	n/s

Table 1. Isolation Point scores

	Visual	Auditory	AV	Visual vs. AV		Auditory vs. AV		Visual vs. Auditory	
				t-value	p-value	t-value	p-value	t-value	p-value
Pafa	16.4	17	15.6	2.69	<0.01	4.41	<0.01	-1.67	n/s
Paga	17.3	12.09	11.9	14.3	<0.01	0.45	n/s	11.7	<0.01
Para	16	11.8	12.4	9.48	<0.01	-1.71	n/s	9.94	<0.01
Pasa	14.8	14.6	13.7	3.79	<0.01	3.37	<0.01	0.43	n/s
Paza	15.3	15.3	13.3	6.26	<0.01	5.44	<0.01	0.08	n/s

Table 2. Recognition Point scores



	Gates																	
pafa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
v /a	0,36	0,99	0,47	0,55	0,74	0,72	0,44	0,71	0,59	0,03	0,05	0,48	0,34	0,03	0,02	0,19	0,02	0,00
v/ av	0,54	0,66	0,18	0,26	0,23	1,00	0,85	0,71	0,49	0,28	0,28	0,99	0,85	0,25	0,29	0,98	0,26	0,34
a/ av	0,80	0,71	0,70	0,86	0,56	0,74	0,40	0,90	0,28	0,16	0,20	0,49	0,26	0,21	0,13	0,20	0,08	0,01
paga	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
v/a	0,07	0,06	0,03	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
v/av	0,23	0,31	0,40	0,69	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
a/av	0,40	0,21	0,15	0,13	0,16	0,88	0,87	0,62	0,42	0,57	0,04	0,08	0,33		0,33			
para	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
v/a	0,05	0,04	0,05	0,12	0,33	1,00	0,21	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,10	0,08
v/av	0,87	1,00	1,00	0,88	0,72	0,54	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,03	0,08
a/av	0,03	0,01	0,03	0,03	0,09	0,45	0,99	0,13	0,26	0,75	0,66	0,66	1,00	1,00	0,33		0,33	
pasa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
v/a	0,26	0,83	0,99	1,00	0,40	0,23	0,02	0,00	0,00	0,00	0,00	0,00	0,04	0,11	0,45	0,42	0,33	0,17
v/av	0,44	0,81	0,60	0,38	0,57	0,06	0,13	0,05	0,00	0,24	0,01	0,00	0,33	0,27	0,99	0,57	0,57	0,33
a/av	0,59	1,00	0,69	0,46	0,18	0,66	0,21	0,00	0,00	0,01	0,01	0,00	0,17	0,14	0,41	0,21	0,13	0,21
paza	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
v/a	0,48	0,16	0,52	0,40	0,31	0,71	0,10	0,03	0,01	0,08	0,47	0,85	0,24	0,10	0,01	0,06	0,01	0,06
v/av	0,80	0,67	0,82	0,68	0,57	0,15	0,05	0,01	0,00	0,00	0,00	0,01	0,00	0,01	0,56	0,33	0,98	0,67
a/av	0,44	0,40	0,38	0,63	0,48	0,21	0,68	0,65	0,06	0,04	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,10

Table 3. Student's t-test comparisons of the percentage of correct responses between visual (V), auditory (A) and AV (audiovisual) modalities of presentation at each gate, for each phoneme.

	Visual only					Audio only					AV				
	f	**z**	**g**	**r**	**s**	**f**	**z**	**g**	**r**	**s**	**f**	**z**	**g**	**r**	**s**
pafa 1	210	114	62	55	27	119	151	66	72	60	213	110	73	56	16
pafa 2	172	111	92	55	38	171	139	57	71	30	159	127	80	60	42
pafa 3	246	61	62	44	55	168	100	49	87	64	210	117	63	30	48
	209,3	95,3	72,0	51,3	40,0	152,7	130,0	57,3	76,7	51,3	194	118	72	49	35
%	44,73	20	15	11	8,5	32,6	27,78	12,3	16	11	41	25	15	10	7,5
paza 1	64	250	74	35	45	113	217	67	43	28	51	272	30	49	66
paza 3	73	205	80	57	53	89	232	33	52	62	56	274	37	64	37
paza 4	77	280	41	35	35	69	265	15	37	82	42	320	25	35	46
	71,33	245	65	42	44	90,3	238	38,3	44	57,3	50	289	31	49	50
%	15,24	52	14	9	9,5	19,3	50,85	8,19	9,4	12,3	11	62	6,6	11	11
paga 2	54	57	222	94	41	22	25	370	38	13	8	35	356	38	31
paga 3	47	44	242	45	90	21	11	405	22	9	21	22	373	38	14
paga 4	50	84	148	138	48	47	30	328	31	32	33	14	351	42	28
	50,33	62	204	92	60	30	22	368	30	18	21	24	360	39	24
%	10,75	13	44	20	13	6,41	4,701	78,6	6,5	3,85	4,4	5,1	77	8,4	5,2

	f	z	g	r	s	f	z	g	r	s	f	z	g	r	s
para 1	13	33	134	224	64	46	27	45	312	38	14	55	57	315	27
para 3	52	54	84	250	28	39	16	58	329	26	13	11	56	341	47
para 4	22	71	91	268	16	39	22	50	330	27	16	32	17	369	34
	29	53	103	247	36	41,3	21,67	51	324	30,3	14	33	43	342	36
%	6,197	11	22	53	7,7	8,83	4,63	10,9	69	6,48	3,1	7	9,3	73	7,7
pasa 1	23	37	45	63	300	39	74	55	74	226	35	46	63	53	271
pasa 2	38	18	31	72	309	80	63	41	56	228	62	41	25	64	276
pasa 3	43	24	46	90	265	56	112	30	46	224	35	91	51	66	225
	34,67	26	41	75	291	58,3	83	42	59	226	44	59	46	61	257
%	7,407	5,6	8,7	16	62	12,5	17,74	8,97	13	48,3	9,4	13	9,9	13	55

Table 4. Confusion matrices between phonemes during the discrimination task in visual (V), auditory (A) and AV (audiovisual) modalities of presentation.

brainglot
bilingualism & cognitive neuroscience