ABSTRACT

Title of Dissertation: CORTICAL DYNAMICS OF AUDITORY-VISUAL SPEECH: A FORWARD MODEL OF MULTISENSORY INTEGRATION.

Virginie van Wassenhove, Ph.D., 2004

Dissertation Directed By: David Poeppel, Ph.D., Department of Linguistics, Department of Biology, Neuroscience and Cognitive Science Program

In noisy settings, seeing the interlocutor's face helps to disambiguate what is being said. For this to happen, the brain must integrate auditory and visual information. Three major problems are (1) *bringing together* separate sensory streams of information, (2) *extracting* auditory and visual speech information, and (3) *identifying* this information as a unified auditory-visual percept. In this dissertation, a new representational framework for auditory visual (AV) speech integration is offered. The experimental work (psychophysics and electrophysiology (EEG)) suggests specific neural mechanisms for solving problems (1), (2), and (3) that are consistent with a (forward) 'analysis-by-synthesis' view of AV speech integration.

In Chapter I, *multisensory perception* and *integration* are reviewed. A unified conceptual framework serves as background for the study of AV speech integration.

In Chapter II, psychophysics testing the perception of desynchronized AV speech inputs show the existence of a ~250ms temporal window of integration in AV speech integration.

In Chapter III, an EEG study shows that visual speech modulates early on the neural processing of auditory speech. Two functionally independent modulations are (i) a ~250ms amplitude reduction of auditory evoked potentials (AEPs) and (ii) a systematic temporal facilitation of the same AEPs as a function of the saliency of visual speech.

In Chapter IV, an EEG study of desynchronized AV speech inputs shows that (i) fine-grained (gamma, ~25ms) and (ii) coarse-grained (theta, ~250ms) neural mechanisms simultaneously mediate the processing of AV speech.

In Chapter V, a new illusory effect is proposed, where non-speech visual signals modify the *perceptual quality* of auditory objects. EEG results show very different patterns of activation as compared to those observed in AV speech integration. An MEG experiment is subsequently proposed to test hypotheses on the origins of these differences.

In Chapter VI, the 'analysis-by-synthesis' model of AV speech integration is contrasted with major speech theories. From a Cognitive

Neuroscience perspective, the ‘analysis-by synthesis’ model is argued to offer the most sensible representational system for AV speech integration.

This thesis shows that AV speech integration results from both the statistical nature of stimulation and the inherent predictive capabilities of the nervous system.

CORTICAL DYNAMICS OF AUDITORY-VISUAL SPEECH:

A FORWARD MODEL OF MULTISENSORY INTEGRATION.

By

Virginie van Wassenhove

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:

Professor David Poeppel, Chair
Dr. Ken W. Grant
Professor Richard Payne
Professor Jonathan Z. Simon

ACKNOWLEDGMENTS

It has been a privilege to have David Poeppel as a mentor, and I thank him for his trusting and encouraging supervision. I thank him for having made graduate school an intellectual challenge and for having provided me with such an exceptional work environment. I thank him also for his ever-raising expectations concealed in an enthusiastic and positive attitude.

I thank Ken W. Grant for his enthusiasm and support throughout my graduate school. I greatly benefited from our discussions and many points raised in this thesis would not have been without his critical insights.

I thank both of you, David and Ken, for forgetting to mention that this 'cool McGurk illusion' would only keep me thinking for the next few years. It has been an honor to work with you both.

I thank Jonathan Z. Simon for his help on signal processing and his serving as a member of my thesis committee. I thank Richard Payne for having been a wonderful professor who secured my interests in the neurosciences during my senior year of undergraduate school. I thank him for having supported my application to the graduate program and for accepting to serve as a member of my thesis committee. I thank John Jeka for kindly accepting to serve as dean's representative and Norbert Hornstein for having tried real hard. I thank David Yager for having served as a member of my cursus committee for two years.

My sincere thank you to Sandy Davis, who got me and kept me in the graduate program. Thank you Sandy for your precious administrative help but especially for your emotional support and kindness. I also thank Robert Magee for his help in financial matters and in traveling arrangements, and for the interlude jokes.

All the members of the NACS program and of the CNL lab have provided an excellent research environment. I am grateful to have met such a diversity of talented minds; all of them have influenced my thinking in many ways.

I thank my office mates for bearing with me for so many hours. Anna Salajegheh, Luisa Meroni and Ana Gouvea, you have been wonderful friends and colleagues and will remain such. Your patience and care have been a tremendous help for me, both personally and academically. I thank Masaya for helping me making our office a 24/7 station. I thank Anthony Boemio for the impatient discussions of our first years of grad school and for his technical help all throughout. I thank Diogo Almeida, without his help some EEG work would not have been completed on time. I thank Jeff Walker for his limitless patience and our well deserved nicotinic breaks. Thank you to Anita Bowles for her editing help on this dissertation, for her commitment to the lab in the past months and for her friendship. Thank you to David Boothe for kicking me out of work.

Thank you to all the graduate students, postdocs and professors I have shared classes, journal clubs, committees, outings and graduate life with... It was a great journey!

I especially thank dear friends and colleagues for their trust and care, and their sharing thoughts on so many diverse and fascinating topics.

I thank Robert Nolan for putting so much optimism in my work and in me; for his ever inspiring reflections, where art and science may meet.

I thank Kuei Yuan Tseng for his kindness, his coaching throughout my dissertation writing and for our intuitive debates on (neuro) (scientific) (bottom-up) (top-down) issues.

I thank Michael Romalis for his genuine curiosity and perspective on neurosciences; may the neurons spike this cortex up!

I thank Nicolas Leroy for a friendship that sustained long distances through time.

I warmly thank David King and Taylor King, so strongly associated with my life here, despite it all.

I especially thank my family for their patience and their moral support. I hope their bearing the distance will at last bring them a little pride and much reassurance. *Je remercie surtout ma famille pour sa patience et son soutien moral. En échange des années de longue distance, j'espère vous apporter un peu de fierté et beaucoup de soulagement.*

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A: auditory

AES: anterior ectosylvian sulcus

AS: Analysis-by Synthesis

AV: auditory-visual

EEG: electroencephalography

FBI: frequency band of interest

FLMP: Fuzzy Logical Model of Perception

fMRI: functional magnetic resonance imaging

HG: Heschl's gyrus

ID: identification

ITG: inferior temporal gyrus

MN: multisensory neuron(s)

MEG: magnetoencephalography

MST: Motor Theory of Speech Perception

MTG: middle temporal gyrus

PET: positron emission tomography

POA: place-of-articulation

PFC: pre-frontal cortex

ROI: region of interest

SC: superior colliculus

STG: superior temporal gyrus

STP: superior temporal polysensory

STPa: anterior superior temporal polysensory

STS: superior temporal sulcus

TOJ: temporal order judgment

V: visual

VOT: voice-onset time

Thesis Outline

'Turn on the light, so I can hear you better'. If that sounds a little odd, here are the reasons why turning on the light is useful.

The face conveys speech information that helps detect and disambiguate the auditory speech signals. In noisy settings, we often look at our interlocutor 'instinctively' as if watching the face increased our understanding of the auditory speech. It actually does. The normal-hearing population performs better in detecting an auditory event when they can see the face articulating the speech sounds and, in fact, speech intelligibility is improved.

What does '*perform better'* mean? '*Perform better'* means (i) faster and (ii) more accurately. These two notions are ecologically significant for the survival of an individual, in any species. Put another way, seeing and hearing a lion in the wild benefits our escape behavior, thus our chance of survival. The behavioral benefit of two sources of sensory information (as opposed to a single source) has been shown in various contexts. The problem is that in Neurosciences and in Cognitive Neurosciences textbooks, we learn that sensory pathways are distinct perceptual systems. Yet, this *'better performance'* must arise from the synergistic interactions of auditory and visual systems and these interactions result in a more efficient (and coherent) representation of the world.

This thesis focuses on auditory-visual speech in the context of multisensory integration. The questions that will be addressed in the following chapters are of three kinds:

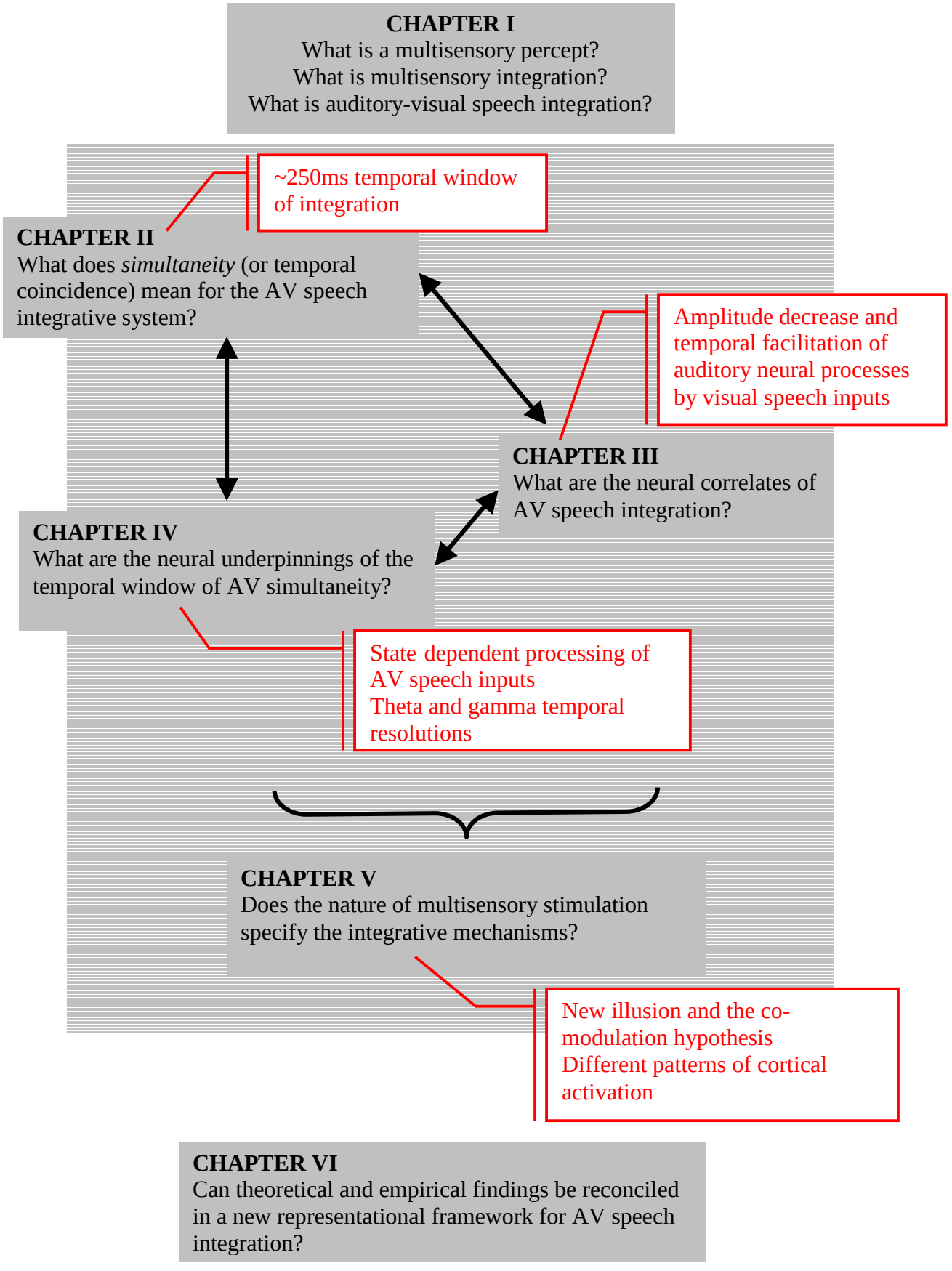
1. In the formation of a multisensory percept, information must combine from separate sensory modalities. The integration process necessitates that information be transmitted via a connected network of neurons. For instance, neural sites that receive inputs from different sensory modalities or 'multisensory convergent sites' are hypothetically well suited to integrate and output a multisensory percept. However in this thesis, I will argue that the *multisensory-by-convergence* hypothesis only provides an anatomical gateway that is ultimately insufficient for AV speech integration.
   Hence, my first question is: given that the cortex is an *anatomically* and *dynamically* defined ensemble of neural subsystems, (i) what are the cortical dynamics of auditory-visual speech integration? and (ii) what do they tell us as far as multisensory integrative processes?

2. Second, not all sensory information can be integrated. If such were the case, perception would be reducible to a 'multisensory instantaneous whole' rather than a collection of 'perceptual events'. It is thus necessary to assume that (i) only particular types of information are meaningful for the system as 'multisensory events' and (ii) that both the physical (external world) and perceptual (internal world) attributes of sensory events constrain the integration process.
   Hence, my second question is: what are the integrative constraints on multisensory perception? Specifically, what are the roles of *sensory invariance* and/or *perceptual redundancy* in the formation of a multisensory object?

3. Finally, a major step forward in the understanding of how a system works is realized when one is able to predict the response of the system given the inputs. My third question is: can we start predicting the cortical dynamics of multisensory integration given the statistical description of (multisensory) inputs?

Each question is of a very general nature and will be treated throughout each chapter. Most of the thesis (except Chapters I and V) is focused on AV speech integration, thus each of these questions are here treated in the context of AV speech integration. Specific sub-questions will be described in each chapter.

The following map provides a dynamic outline of the thesis.


CHAPTER I
What is a multisensory percept?
What is multisensory integration?
What is auditory-visual speech integration?
~250ms temporal window of integration
CHAPTER II
What does simultaneity (or temporal coincidence) mean for the AV speech integrative system?
Amplitude decrease and temporal facilitation of auditory neural processes by visual speech inputs
CHAPTER III
What are the neural correlates of AV speech integration?
CHAPTER IV
What are the neural underpinnings of the temporal window of AV simultaneity?
State dependent processing of AV speech inputs
Theta and gamma temporal resolutions
CHAPTER V
Does the nature of multisensory stimulation specify the integrative mechanisms?
New illusion and the co-modulation hypothesis
Different patterns of cortical activation
CHAPTER VI
Can theoretical and empirical findings be reconciled in a new representational framework for AV speech integration?

Chapter 1: Multisensory perception

Everything should be made as simple as possible, but not simpler.

Albert Einstein

The sum of the parts is not the whole!

Lao Tse *(B.C. 600), 39th saying in Tao-te-King*

The rationale of this introductory chapter is to provide a general overview of both theoretical and empirical issues in the field of multisensory perception. In particular, the notions of 'multisensory perception' and 'multisensory integration' will be reviewed and explicitly defined.

In part 1, I will begin with the notion of percept formation with regard to classic Gestalt principles. In multisensory perception, 'sensory invariance' relates to a kind of information available across sensory systems that may provide the basis for a multisensory feature. Developmental results will be discussed to clarify the nature of sensory invariance and the emergent classification of multisensory percepts. The

discussion will then be extended to memory systems in order to highlight the potential contribution of sensory-specific and multisensory areas in the formation of multisensory percept.

In part 2, I will review the neural substrates of multisensory percept formation. Specifically, I will discuss the current 'multisensory convergence' hypothesis and raise the issue of how multiple levels of multisensory interaction -readily observable psychophysically- can be implemented in neural systems. A broader definition of multisensory integration will be provided.

Part 3 will focus on current issues raised in auditory-visual speech research, and provide a broad background to the issues that will be developed in the following chapters.

1.1 What is a multisensory percept?

From continuous sensory inputs to discrete representations

The array of energy in the external world provides the sensory systems with a continuous flow of information. The sampling of information by the nervous system is, however, naturally constrained early on by the temporal resolution of the sensory receptors and later on by the underlying dynamics of neural populations. For example, temporal discontinuities as short as ~1 or 2 ms can be detected in the auditory domain (Plomp, 1964; Penner 1977), but the temporal threshold of the visual system is on the ~25 ms scale (e.g. Hirsh & Sherrick, 1961). In the formation of more complex perceptual units -e.g. a speech sound or a musical tune-, larger integration

times are observed on a ~25ms to ~250ms time-scale in both the auditory (e.g. Näätänen, 1992; Yabe *et al.*, 1998; Winkler *et al*, 1998; Poeppel, 2003; Shinozaki*et al.*, 2003) and visual systems (e.g. Heller *et al*, 1995; Mackick & Livingston, 1998).

The evidence for neural computations over multiple time scales -i.e. with different temporal resolutions- suggests that (i) various levels of information extraction exist that are associated with different degrees of resolution and (ii) that each of these levels of information extraction may correspond to a particular perceptual representation.

In Gestalt psychology, a percept results from the synthesis of sensory 'primitives' (Hartmann, 1974). The grouping of sensory primitives is driven by principled spatial and temporal configurations of features -e.g. continuity or proximity. Additionally, arrays of energy in the environment are classically defined according to the sensory system dedicated to their processing. For instance, the luminance contours of a visual event refer to the kinds of information extracted in the visual system, while the spectral characteristics or the amplitude of an acoustic pattern refer to information extracted in the auditory system. From an information-theoretic standpoint, the Gestaltist principles should be verified in the nervous system as a set of computational principles used in the extraction and recognition of sensory inputs as perceptual events.

The extraction of features is fundamental for the recognition (identification) of complex continuous inputs and may be intrinsically associated with perceptual categorization (e.g. Massaro, 1998). Information extraction in the sensory systems is specified by both the architecture and the computational capability of the nervous

system (e.g. multiple-temporal resolutions). However, the extraction of information alone is not sufficient to account for the classic observation that a group of features is rarely, if ever, reducible to their sum. Indeed, a percept is *qualitatively*different from the perceptual features that compose it -i.e. the resulting percept is neither reducible to one feature nor to their linear sum- as captured by the classic Gestaltist saying, "The whole is qualitatively *different* from the sum of its parts" (Ehrenfels, 1890).

Spatio-temporal discretization in multisensory perception

In both auditory and visual perception, the characterization of features and the relations among them has been extensively studied. The sensory-specificity of features results from a straightforward description of their interactions within each physical domain (e.g. continuity in the pattern of luminance). Leaving the unisensory side of perception to address the issue of how our senses interact to build a coherent representation of the world is yet another challenge. In effect, sensory events are analyzed through their acoustic, visual or somatosensory attributes, i.e. processed and represented through separate sensory streams. Yet ultimately, our perception is that of an object or whole, not that of a separate group of sensory attributes (Treisman, 1996). For instance, in a noisy room, on which basis does one match a voice with a particular talker? Or, why is the strike of a tennis ball naturally associated with particular sound qualities? To date, no clear Gestalt principles of multisensory perception have been described, thus preventing the systematic description of multisensory interactions and that of their products, the multisensory percepts. In fact,

the very idea -or lack thereof -, that Gestalt principles apply to multisensory perception leads to various theoretical assumptions not always acknowledged in the experimental literature.

Auditory and visual information originate from two independent sources of energy. More precisely, visual information originates from the transduction of photons arriving at the retina while auditory information originates from the sound pressure waves mechanically transmitted to the basilar membrane. Yet numerous references to 'multisensory redundancy' of information have been used in the literature (e.g. Giard & Peronnet, 1999). Positing 'cross-modal redundancy' of information is not an obvious statement given that audible and visible spectra do not overlap. In fact, this assumption presupposes that a receiver –here, the nervous system- is adapted to detect and process redundancy across sensory channels in spite of no *a priori* physical redundancy. Consequently, in any theoretical and empirical account of multisensory perception, one must ultimately specify (i) the properties of the 'processor' and (ii) define what type of information constitutes a source of incoming sensory 'redundancy'. In particular, one needs to determine if the processor possesses an adequate temporal resolution for segmenting the hypothesized flow of information or whether that provided -and in fact used- by individual sensory systems is sufficient to account for the formation of a multisensory percept. More importantly, *sensory non-invariance* calls upon the notion of *perceptual redundancy*, a 'high-level' (or abstract) property, where features from some proximal objects must have already been extracted.

The amodal spatio-temporal structuring –or sensory invariance- of multisensory events

Because there exists no 'multisensory receptor' to date, and assuming that at some stage in the nervous system segregated sensory streams interact, two *non-specific* types of information are possible candidates for cross-modal redundancy, namely spatial location and timing (and yet to be specified 'multisensory features' see above). Natural multisensory occurrences originate from the same spatial location and share a common temporal structure. These two types of information are readily available in all sensory systems. The non-sensory specific nature of spatial and temporal information provides a convenient source of redundancy again assuming that the *reference frame* – or the coordinate system- in which this information is encoded is similar across sensory domains.

For example, in auditory-visual speech, the articulatory movements of the face and audio speech originate from the same location, and recent findings have suggested that the amplitude envelope of the auditory speech signal correlates with the mouth movements (Grant, 2001). This spatiɵ temporal 'co-modulation' of sensory inputs may provide a basis for multisensory unit formation regardless of the nature, meaningfulness or complexity of stimulation. The spatio-temporal structuring of ambient arrays of energy has classically been considered an 'amodal property' (Lewkowicz, 2000) or a 'sensory invariance' (Gibson, 1966) of the sensory world i.e. independent of the input modality (e.g. Marks, 1978). Is this amodal spatiɵ temporal structure a basic unit of multisensory perception? And if so, does it constitute a feature or primitive -in the Gestalt sense- of multisensory percept formation?

Developmental studies have started to address this question, and a diverse pattern of spatio-temporal cross-modal matching has emerged. In infant studies, two general methodologies are commonly used (Lewkowicz, 2000). In the first method, the perception of intersensory relations is inferred by the infant's gaze duration. For example, a mismatch presentation (e.g. asynchronous intersensory stimulation) would lead to smaller gaze duration than a matched one. A second method consists of familiarizing the infant with a pair of stimuli and then introducing parametric changes in the stimuli. The perception of changes is accompanied with variations of gaze duration that are compared with reference to the initial gaze duration observed in the habituation period.

Simultaneity, duration and rate matching across auditory and visual modalities have been characterized at different stages of development, suggesting the existence of separate computational stages for auditory-visual (AV) interactions. Contrary to the predictions of the 'nativist' view[1], sensory invariance -or the spatio-temporal relationship of an AV event- is acquired gradually yet early in life (Lewkowicz, 2000, 2002). The detection of AV synchrony emerges first and is followed by duration, rate and rhythm matching in the first 10 months of an infant's life. Similarly in the spatial domain, cross-modal associations are established slowly during the first 2 years of life. This gradual acquisition of spatio-temporal relationships suggest that the more complex the pattern, the later the acquisition, in agreement with the 'increasing specificity hypothesis' (Gibson, 1969; Spelke, 1981). For instance, auditory-visual intensity matching has been shown in neonates (Lewkowicz & Turkewitz, 1980) and

[1] The nativist view considers that knowledge is 'innate' i.e. part of our endowment. To the contrary, the empiricist view considers that knowledge is acquired.

three and a half months old infants are sensitive to natural temporal structures, but only later (7 months) can they detect 'arbitrary' cross-modal associations such as pitch and shape (Bahrick, 1992) or even emotion matching in strangers (Walker-Andrews, 1986). Surprisingly however, an early sensitivity to complex AV speech events has been reported in 5 months old infants, for instance in detecting the congruency of an auditory speech inputs with the articulatory movements of the face (Rosenblum *et al*, 1997).

While there is no clear evidence for an innate ability to process multisensory spectro-temporal invariants, the spatio-temporal structuring of arbitrary patterns as well as the nature and ecological relevance -and ultimately the representational value- of incoming information need to be considered as important factors in the development and the spatio-temporal tuning of the multisensory system. The acquisition of intersensory equivalences seems to undergo a perceptual restructuring - i.e. *an internalization of perceptual grouping rules*- that is overall non-linear. Importantly, these results illustrate the need for various degrees of cross-modal, spatio-temporal resolution and also suggest that sensory modalities may interact at various stages of their sensory-specific processing. Ultimately, various levels of multisensory representation should be distinguishable.

Effects of incongruent spatio-temporal structuring on perception

The early representations of cross-modal 'invariance' are of course preserved in adulthood. Cross-modal simultaneity thresholds for simple stimuli (e.g. tone/flash) approximate ~ 25ms (Hirsh & Sherrick, 1961; Lewald *et al*, 2001 ; Zampini *et al*., 2002) but multisensory interactions are much less sensitive for complex stimuli (Dixon & Spitz, 1980; Massaro, *et al*, 1996; Munhall *et al*, 1996). A variety of multisensory effects has been reported. These studies employed different types of stimuli and paradigms based upon the spatio-temporal congruency of stimuli inputs (see Appendix A). An emerging principle is that auditory information tends to bias visual perception in the temporal domain (e.g. Shams *et al*., 2002; Welch *et al*, 1986; Recanzone, 2003), while visual information biases the spatial localization of auditory events (e.g. Howard & Templeton, 1966). For instance, in the 'visual bouncing' illusion, a sound is presented at the point of crossing of two continuously moving visual circle. The presentation of the sound results in the 'bouncing' of the circles otherwise perceived as streaming through each other in a visual alone presentation. In the classic ventriloquism effect, the spatial source of a sound is 'captured' by the presence of a moving visual stimulus (e.g. Howard & Templeton, 1966). The direction of the biasing effect has been suggested to follow that of the 'modality appropriateness' rule, where vision serves as the 'spatial' dominant modality and audition as the 'temporal' dominant modality (Welch & Warren, 1980). However, this dichotomy has recently been questioned in light of recent findings showing for instance a temporal ventriloquism effect (Bertelson & Aschersleben, 2003).

The environment is naturally endowed with rich and diverse sources of sensory information. Developmental results suggest that multisensory perception engages different types of interactions associated with different levels of spatial and temporal resolution and that the ecological relevance of multisensory patterns appears to be a critical feature of the system. Hence, a characterization of multisensory integration solely based on global spatio-temporal coincidence does not appear sufficient to determine the nature of these interactions and more than one mechanism of multisensory interactions needs to be considered.

Multisensory memory systems

If early intersensory interactions - i.e. featural level of representation - exist, can we then talk about multisensory objects as we talk about auditory or visual objects (Kubovy & Van Valkenburg, 2001)? Specifically, results of developmental studies suggest an early sensitivity to cross-modal associations that are in the natural register of the system, regardless of the spatio-temporal complexity of the stimuli at hand (e.g. Rosenblum *et al*, 1997). Perceptual object formation is tightly associated with storage mechanisms, which facilitate the extraction of information most relevant for the biology of the system. Sensory storage is particularly advantageous in preserving extracted information on a short-time scale –i.e. on the order of 200 ms – ensuring that transient inputs and short processing jitters across processing modules are integrated on line. Prior experiences establish long-term prototypical storage, which structures perceptual experience and render the parsing of the sensory world

more efficient. For instance, face perception is readily accessible in its synthetic form due to the natural and regular configuration of eyes, nose, and mouth in a restricted spatial area (Farah, 1990), allowing for rapid processing despite the overall complexity of visual configuration.

Many studies on the perception of visual objects have naturally focused on object recognition, yet most visual objects can also be categorized on the basis of several modalities (e.g. you can see, hear and touch a piano). It is uncertain whether such representations are stored in explicit memory in a sensory-specific manner or as a multisensory (or *abstract*) whole accessible by each sensory system. To date, the dominant view is that storage in explicit memory is "*post-hoc*" to sensory-specific representation and becomes multisensory or rather*amodal* after cross -modal associations have been consolidated by experience (e.g. Mesulam, 1998).

A major question being raised in multisensory representation is whether object recognition re-activates the entire sensory-specific and multisensory network used during encoding. Available studies on object recognition via different modalities assume an online multisensory integration process and the question being addressed is *how* integration takes place, not how information is being retrieved from different modalities. If sensory-specific attributes are re-activated through multisensory object recognition, one predicts interactions of scattered unisensory memories, and this hypothesis may be supported by a global spread of neural activation.

The Superior Temporal Sulcus (STS) and the Middle Temporal Gyrus (MTG) have recently been proposed to mediate multisensory integration in object perception (Beauchamp *et al*, 2003). These regions typically show enhanced activation to the

AV presentation of objects when compared to unisensory presentation of the same objects. In a different paradigm, the learning of novel or arbitrary AV associations -such as the pairing of ideograms and musical tones- elicits pre-frontal activation classically associated with retrieval in explicit memory (e.g. Gonzalo *et al.*, 2000). Frontal enhancements to the presentation of arbitrary AV pairings -tones paired with static or morphing circles- have also been reported in electroencephalographic (EEG) studies (Giard & Peronnet, 1999; Molholm *et al.*, 2001) and the sources of activation were further located with fMRI in the fronto-temporal area (Fort *et al., 2001*). More recently, a similar spread of activation over frontal areas as well as the temporo-parietal junction was observed for the detection of salient or novel multisensory pairings (Downar *et al*, 2002). Although very few studies have yet addressed the question of multisensory object retrieval, a global network of activation is emerging that supports the reactivation hypothesis – i.e. cortices involved in the encoding process are recruited during recall (e.g. Nyberg *et al.*, 2000). The notion of 'salient' (or biologically relevant) stimulus evaluation was further proposed to characterize this global network, to emphasize its possible non-modality specific nature (Downar *et al.*, 2002). The notion of saliency again raises the importance of ecological validity for if artificial pairing induces the need for explicit memory retrieval, salient or ecologically natural stimulation may be more "automatic"[2].

[2] In fact, much evidence in language studies shows that priming effects -i.e. testing implicit memory (for review, see Schacter & Buckner, 1998)- are robust regardless of the input modality (e.g., Graf et al., 1985, Badgaiyan et al., 1999, Curran et al., 1999, Carlesimo et al., 2003), suggesting that (i) either pre-attentive intersensory interactions are effective in a sensory-specific form or (ii) that supra-modal feedback strongly modulates incoming sensory inputs. In language studies, similar frontal patterns of activation are observable and are associated with semantic processing –i.e. 'assignment of meaning' to linguistic stimuli (e.g. Gabrieli *et al.*, 1998; Price, 2000). Because language readily engages auditory (speech) and visual (lipreading and written alphabet) modalities, it stands as a useful

In AV speech, which constitutes a natural form of language-specific stimulation, -in the ecological sense, face-to-face conversation- of language-specific stimulation, the core fMRI results show that auditory cortices and STS are enhanced as compared to the presentation of auditory or visual speech alone (Calvert, 1997, Calvert *et al*, 1999, 2000). (These results will be discussed in more depth throughout the thesis)

Multisensory perception is at an early stage of investigation. Although pioneering studies have pointed out the complexity of the problem, it is only recently that experimental designs have included the potential gradation of multisensory interactions - sensory, perceptual and cognitive. Overall, available studies suggest that the complexity of multisensory perception does not solely reside in the integration of multisensory information but also in the storage and retrieval of amodal and sensory-specific information –see for instance an interesting contribution to the hypothesized amodal nature of visual storage in the inferotemporal (IT) cortex in monkeys (Gibson & Maunsell, 1997). The intricacy of neural processing that underlies various stages of multisensory processing calls upon a reevaluation of the role of (i) implicit memory in the formation of multisensory percepts and (ii) explicit memory as an interface of amodal storage and multisensory percept representation. It is yet to be determined whether stored information in its amodal form also has to be

---

means to study the processing of multisensory inputs and amodal representations. It is generally accepted that visual information in its written form (a *learned* and *arbitrary* visual pattern across languages) is eventually recoded into a phonological form. Interestingly, behavioral facilitation (as measured by faster reaction time) is observed with grapheme and phoneme pairing (e.g. Dijkstra *et al.*, 1993) and a different pattern of activation emerges in particular marked by a decreased activation over the STS (Raij *et al.*, 2000). More recently, a combined magnetoencephalography (MEG) and fMRI study comparing priming of written and spoken word shows activation of a temporo-frontal network interpreted as evidence for a supramodal semantic network (Marinkovic *et al.*, 2003).

computed on-line cross-modally and importantly, at which representational stage supramodal feedback can be triggered. It is clear that initially, computations depend on the basic spatio-temporal coincidence of multisensory stimulation.

Definition of a multisensory percept

In summary, a multisensory percept is herein defined as the unique product resulting from an on-line multisensory integration process. The presentation of co-occurent multisensory events (i.e. within the temporal resolution of the integrative process) serves as a necessary condition for multisensory percept formation. The immediacy of multisensory synthesis is to be distinguished from the retrieval of 'amodal' percepts, whose representation may be accessed independently by each modality. A perceptual output will thus be considered "multisensory" if it differs from any one of the perceptual outputs obtained under unimodal stimulation (i.e. the AV resulting output must differ from auditory alone or visual alone stimulation, providing that the same unimodal stimulation is presented in unimodal or bimodal conditions). If this condition is not applicable the percept will be amodal[3].

Multisensory percepts may exist at various stages of perceptual complexity - e.g. from 'meaningless' temporal pattern to perceptual unit- contingent upon (i) the hierarchical level at which sensory-specific interactions occur and (ii) the accessibility of the resulting output to perceptual awareness.

[3] By 'perceptual output', I here consider any quantified output parameter such as performance (e.g. detection, identification) and/or reaction times.

Accordingly, multisensory interactions may engage diverse neural mechanisms, which may both impact and involve sensory-specific computations.

Available studies of multisensory perception often fail to indicate which level of perceptual organization is being tested, hence oversimplifying the type of interactions and the nature of the perceptual domain and neural systems being involved.

1.2 What is multisensory integration?

From discrete features to whole percept

The flow of sensory information follows a classic time-course from the sensory receptor, where the incoming information is converted into neural energy (transduction stage), to cortical areas where information 'filtering' -as constrained by the receptive fields or tuning properties of individual neurons and the dynamics of neural populations - leads to the featural decomposition.

One of the most thoroughly studied sensory systems is vision, where more than 30 functionally differentiated processing modules have been described in the macaque. These processing modules create a functional hierarchy (i.e. serial processing from V1 to V2 to V3 etc.) where each processing stage extracts a specific kind of information (e.g. orientation) (Hubel & Wiesel, 1977). This hierarchy is associated with a second level of functional differentiation instantiated in parallel -or concurrent- processing streams. In the visual system, the dorsal stream or 'where

pathway' is fast but low-resolution ('global' processing) and underlies a spatial-type of processing (e.g. motion), whereas the ventral 'what pathway' is slower and high-resolution ('local' processing) and underlies general pattern recognition (Ungerleider & Mishkin, 1982). The combination of hierarchical (i.e. serial) and parallel streams of information processing provides a natural substrate for sensory-specific analytical mechanisms (Felleman & van Essen, 1991). Evidence for a similar architecture in the auditory system has been put forward (Ehret, 1997; Rauschecker, 1998; Rauschecker *et al.*, 1998; Kaas *et al*, 1999; Alain *et al*, 2001).

The functional specialization in neurophysiological modules provides a means to analyze sensory information in a featural mode that resembles the Gestaltist principles, where smaller units of perception are ultimately synthesized as a whole. This modular organization has however raised the classic 'binding problem' (Treisman, 1996) -i.e. how extracted features are combined or integrated into a perceptual whole.

First, neurophysiological findings have shown that in a given processing stream, neural receptive fields tend to enlarge as one goes up the hierarchy (Barlow, 1972, 1991). One cell -the 'cardinal' or 'grandmother' cell- could ultimately synthesize the preceding computations of the processing stream into a whole, suggesting that featural binding lies in the natural progression of local computations in the hierarchy. This view has raised numerous questions, among them the problem of resource limitation, that is, where a limited number of neurons must represent a countless number of occurrences. The local coding hypothesis eventually limits the

functionality of a cardinal cell to that of single representation, thus also rendering difficult the representation of novel events.

An alternative to the local coding hypothesis is the 'distributed coding' hypothesis, where several neural binding mechanisms are proposed. For instance, featural integration could be mediated by neuronal ensembles (Singer & Gray, 1995; Engel *et al.*, 1997). The evidence for cortical and subcortical oscillatory mechanisms at different frequency ranges is abundant (e.g. Steriade & Amzica, 1995; Ritz & Sejnowski, 1997) and transient synchronizations of large neural populations are readily recorded via non-invasive methods such as electroencephalography (EEG) and magnetoencephalography (MEG) in humans. A given neural ensemble -or processing module- can take part in simultaneous computations providing that it operates in different frequencies bands (e.g. Başar, 1998, 1999; von Stein *et al*, 2000). [4]

The frequency ranges of neural oscillations have characterized various stages of perceptual processing. For instance, alpha band activity (8-12Hz) is predominant in early sensory-specific stages of information extraction (Dinse *et al*, 1997; Isoglu - Alkaç *et al*, 1999) while the so -called 'gamma-band' (25-75Hz) has been hypothesized to underlie perceptual binding (e.g. Tallon-Baudry & Bertrand, 1999). Furthermore, attentional modulation of oscillatory mechanisms has been described (e.g., Fries *et al*, 2001) and, interestingly, patterns of synchronization in

[4] The intrinsic dynamics of neural systems and their specificity at various scales of the neural groupings must naturally provide a functional substrate for the processing of the environment spectro-temporal complexity and the no less intricate structuring of cognitive events. From a perceptual standpoint, the bidirectionality -upstream and downstream- of information processing in the cortex further necessitates cross talks over distant areas and measures of coherence (and coherency) are a possible means by which to quantify global interactions (e.g. Blake & Yang, 1997; Tononi *et al.*, 1998; Rodriguez *et* al., *1999*; Lachaux *et al.*, 2001).

schizophrenic patients have been reported that differ from those seen in normals (e.g. Lee *et al*, 2003; Spencer *et al.*, 2003).

Oscillatory mechanisms in a specific frequency range do not stand independent of each other. Whereas it is, of course convenient to quantify the perceptual correlates (such as grouping) in one frequency range, it may neglect fundamental dynamic interactions across sensory systems. For instance, low-frequency bands (such as theta band, 4-7Hz) are tightly coupled with the gamma band from the very specifications of their neural generators (da Silva, 1991; Buzsáki, 2002). Recent evidence suggests, for example, that the theta band is modality-independent (Yordonova *et al.*, 2002) while theta and the coupling of theta-gamma bands may underlie working memory functions (e.g. Schack *et al.*, 2002; Howard *et al.*, 2003). The 'burst of gamma' observed in perceptual testing originates from a transient synchrony of neural populations involved in the processing of a sensory event and has been interpreted as the 'emergent realization' or recognition of this event by the nervous system -i.e. the stimulus enters the realm of conscious perception (Sauvé, 1999; Bertrand & Tallon-Baudry, 2000; Hopfield and Brody, 2000; Izhikevich *et al*, 2003).

These correlational observations do not directly pertain to the nature of the representations but rather to the computational processes that may underlie their formation (Shadlen & Movshon, 1999). The understanding of neural systems is at an early stage of scrutiny, but results continuously re-define the realm of possible neural coding with an increasing specificity (e.g. Oram *et al.*, 2002; Panzeri *et al*, 1999; Lestienne, 2001). Principles of information theory and signal processing are being

promisingly applied and provide a systematic means to interface neural signals and experimental stimulation paradigms (e.g. Rieke *et al*, 1999; Thomson, 2001).

Limitations of the multisensory neural convergence view

In multisensory perception research, the nature (i.e. the representational stage) of sensory-specific inputs at which information is being combined has often been oversimplified. To date, the convergence of sensory streams onto multisensory sites of integration is often hypothesized to account for multisensory perceptual binding. Yet, the diversity of multisensory perceptual phenomena suggests that multisensory interactions occur at various stages of perceptual representation.

Here, the 'multisensory convergence' hypothesis is proposed to be insufficient to account for the range of multisensory interactions reported in the literature. By analogy with the cardinal cell or local coding hypothesis, I will argue that multisensory convergence fulfills a *necessary* but not a *sufficient* condition for the formation of a multisensory percept. In particular, constraints must be posited that involve at least two levels of analysis for the completion of a multisensory percept.

First, the sensory invariance of spatio-temporal information provides the substrates for a coarse intersensory binding, whose computational relevance will be described below. While it may suffice for the *detection* of intersensory co-occurrence, it remains insufficient for its *identification*.

It is in a second stage that the representational value (meaningfulness or 'semantics' (Marks, 1978)) will be determined according to the nature of the detected event - an AV speech event vs. a pianist pressing a keyboard. Specifically, it is predicted that the *nature* of a multisensory event will influence which neural pathways are engaged in its computation. Although the latter may appear a reasonable point of departure, no such consensus has yet been reached or clearly stated.

Origins of the multisensory neural convergence view

Multisensory neurons (MN), neurons that receive inputs from more than one sensory modality, have been studied in the deep layers of the superior colliculus (SC). The neurophysiology of this subcortical structure has been extensively described using the cat model (e.g. Stein & Meredith, 1993). More recently, various cortical sites receiving converging inputs from different modalities have started to be characterized in monkeys and humans -e.g. the Superior Temporal Sulcus (Benevento *et al.*, 1977; Desimone & Gross, 1979; Bruce *et al.*, 1986; Hikosaka *et al*, 1988), the orbito- and pre-frontal cortices (e.g. Benevento *et al*, 1977; Fuster *et al*., 2000), and the posterior parietal cortex (Hyvärinen & Shelepin, 1979; Leinonen *et al*, 1980; Cohen & Andersen, 2002) (cf. also Pandya, 1982). The distinctive connectivity and neurophysiology of multisensory neurons –and, in particular, their enhanced response to the presentation of multisensory inputs- have led to the overwhelming assumption that any type of multisensory integration must derive from multisensory convergence. While the enhancement effect or 'supra-additive' response of MN is frequently

referred to in the literature as the indication for ‘multisensory integration’ and indeed became the characteristic ‘verification’ in functional brain imaging techniques for its existence (e.g. Giard & Perronet, 1999; Calvert *et al.*, 2000; Molholm *et al*, 2002), MN display complex response properties. These properties not only depend on the spatio-temporal relationship of multisensory inputs, but also on the intrinsic diversity of their tuning properties (for review, see Stein and Meredith, 1993). Here, a critical summary of the major neurophysiological properties is provided (essentially drawn from studies in the SC of cats, unless otherwise indicated).

First, MN display similar spatial tuning properties for each sensory modality to which they are responsive, hence sensory-specific inputs are said to be in ‘spatial register’ with one another (Stein & Meredith, 1993). It follows that the sensory-specific spatial location of a multisensory event is intrinsically specified by the activated neuron. In macaques, visually dominated MN of the Superior Temporal Polysensory cortex (STP) display a coarser spatial tuning than unisensory cells, and receptive fields as large a 105 x 150 degrees were reported (Desimone & Gross, 1979).

Second, this spatial register is supplemented by complex temporal tuning properties. Temporal resolutions of MN vary from ~20 ms to as much as 1.5 seconds of stimuli desynchronization, within which integrative properties are preserved (Meredith *et al*, 1987; Stein & Meredith, 1993). Three major types of response profiles were described in the SC of cats when stimuli were desynchronized, namely ‘enhancement’, ‘enhancement-depression’ and ’depression’ responses (Meredith *et*

al., 1987). Importantly, the duration of a cell's discharge rate to the first incoming input determines the size of the simultaneity window. In the SC, first spike latencies show great variability for each modality but the trend was shorter for auditory inputs (~20 ms) followed by somatosensory (~30 ms) and visual inputs (~80 ms). Similarly in the orbital cortex of the macaque, various response profiles were obtained to the presentation of desynchronized stimuli to auditory-visual MN. Multisensory responses were observed during both short (~20 ms) and long desynchronization (~400 ms) (Benevento *et al.*, 1977). While the temporal tuning of MN has not been systematically described, it has been suggested that similar tuning properties should be found in STS, frontal or parietal polysensory cortices (Stein & Meredith, 1993; Stein *et al*, 2000). This hypothesized multisensory global network, given its coarse resolution, was proposed early on to be specialized in 'global' computations rather than pattern recognition (or 'local') (Bruce *et al*, 1986 *)*.

In summary, the integrative properties of MN manifest when stimuli are within the defined spatial and temporal resolution of the cell. The resolution of the temporal window defines the 'spatio-temporal coincidence' principle (Stein & Meredith, 1993). Extracellular single unit recordings have shown a supra-additive enhancement of MN discharge rate when presented with spatio-temporal coincident multisensory stimuli. For instance, the simultaneous presentation of a tone (A) and a flash (V) at the same spatial location elicits a discharge rate higher than the summation of the neural outputs when identical stimuli are presented alone, and for a longer duration. It is this particular response type that often serves as an account for various multisensory perceptual phenomena despite the ostensibly different system

levels involved. While a regular pattern of response profile is being shown (i.e. enhancement when stimulation within the neuron receptive field and depression when outside the neuron receptive field), strict spatio-temporal boundaries are not always clear, and supra-additive enhancement (as well as suppression) is observed even when multisensory stimulation is not within the receptive fields (RF) of the neurons (e.g. Kadunce *et al*, 1997). In fact, some MN also show a depressed response to the presentation of multisensory events (Stein and Meredith, 1993). *The reduction of multisensory integrative mechanisms to supra-additivity is thus highly simplistic and neglects important aspects of MN response properties.*

For instance, multisensory neurons are also responsive to unimodal stimulation; they do not respond exclusively in a multisensory context (Stein & Meredith, 1993). In fact, a multisensory cell is generally 'modality-dominant' in that it will respond maximally (i.e. higher spike rate and for a longer duration) to one modality when presented alone. This dominance is not only important for the type of response integration, but also for the temporal tuning of the cells in multisensory stimulation. For instance, multisensory enhancement is maximal for minimally effective unisensory stimulation as captured in the 'inverse effectiveness principle' (Stein and Meredith, 1993). In other words, the 'benefit' of enhanced responses is lowered for a highly – unisensory - dominated cell as compared to a non-dominated one, and response properties can demonstrate sub- or supra-additive effects depending on the absence or presence of the preferred modality, respectively.

If multisensory convergence provides a possible route for the integration of multisensory inputs, the type of information reaching the multisensory convergence sites remain unclear. While the diversity of MN responses depends on the coarse spatio-temporal relationships of multisensory inputs, the lack of stimulus specificity observed even in multisensory cortices remains puzzling (Bruce *et al*, 1986; Hikosaka *et al*, 1988). A growing body of evidence in functional brain imaging studies highlights the involvement of diverse and global cortical networks in multisensory processing, including classic multisensory convergence sites. Furthermore, anatomical and neurophysiological studies have suggested that primary sensory-specific cortices are interconnected directly (e.g. Zhou & Fuster, 2000; Falchier *et al*., 2002; Rockland & Ojima, 2003). Because the superior colliculus is accessed early on in the hierarchical processing stream, one may assume that supra-additive integration at this early processing stage drives subsequent processing, in particular into the cortex. However, it has now been shown that integrative properties (and in particular enhancement) of MN in the SC are gated by multisensory cortical sites such as the Anterior Ectosylvian Field (AEF) in cats (Jiang *et al*, 2001) but also sensory-specific cortices (Wickelgreen & Sterling, 1969; Stein & Arigbede, 1972; Stein & Gallagher, 1981; Meredith & Clemo, 1989). Conversely, as recently pointed out by Schroeder and colleagues (2003), cortical multisensory areas are rather late in the sensory processing stream, and while the integrity of the cortico-subcortical network is necessary for classic multisensory integrative properties observed downstream in subcortical sites, the processing stage at which multisensory sites of

integration operate remains unclear. In the next section, a complementary view to the multisensory convergence hypothesis is suggested.

Multisensory-integration-by-convergence

Having summarized general properties of MN and some of their limitations, new challenges are now being raised. First, do multisensory convergence sites provide enough constraints on incoming inputs? Second, what is the representational status of the multisensory integrated output? Specifically, is integration by multisensory convergence sufficient for the completion of a 'multisensory percept', or is more processing needed?

Consider first a general key issue of (multisensory) perception. In a natural environment rich with (multisensory) stimuli, the perceptual system must extract sufficient (multisensory) information from the background noise not only to detect potentially significant sources of information, but also to identify and categorize them. Specifically, in multisensory perception, the detection stage is further complicated by the need for separate sensory systems to determine that two arrays of energy provide information about a single perceptual source-event. Despite their lack of specificity, it is clear that MN have a crucial role to play in coarse spatio-temporal intersensory binding. In fact, as a result of multisensory convergence, sensory-specific inputs are clearly transformed into a single output –i.e. two presumably different encoding schemes become one. This coding step maximizes intersensory information, i.e. increases the saliency of 'amodal' redundant information - as shown

by the ‘supra-additive’ responses- and may ultimately improve the detection of multisensory events behaviorally (Stein & Meredith, 1993). As such, multisensory convergence unquestionably provides a distinctive stage of multisensory representation.

However, the environment provides a large number of coincident signals which *do not* relate to any perceptual category. For instance, at a red light one may have observed the windshield wipers seemingly in tune with the beat of a musical piece playing on the radio. While experiencing a sense of ‘coherency’ between the auditory and visual inputs, and despite the strong spatio-temporal relatedness of the two events - i.e. invariant physical redundancy -, no unique perceptual category emerges in this case. In fact, there exists no ‘windshield wipers music beat’ category naturally available in the nervous system - i.e. perceptual redundancy - and we are readily aware of the cause-to-effect of this occurrence. Finally, there is no confusion as to which information is provided to each sense, unlike cases of multisensory fusion detailed below.

From this anecdotal example, the point I would like to argue for is that two types of neural processing may occur: (i) multisensory convergence results in a ‘pseudo-unified percept’ eventually suppressed or unbound because it is has no representational value as a unique percept or, (ii) other mechanisms come into play in the formation a natural percept. The former solution is not viable because in most, if not all, multisensory situations, be it perceptually categorizable (a speech token) or not (tone paired with a flash), reaction times are faster in multimodal conditions than in unimodal conditions (e.g. Hershenson, 1962 and Appendix A). Thus, a re-

evaluation of multisensory output is unlikely, for it would at least predict reaction times, if not slower, at best similar to unisensory evaluations. This is overwhelmingly not the case. If it is necessary to posit other mechanisms besides multisensory convergence in the formation of natural percepts, we need to re-focus on the nature of the computations that follow or *parallel* the multisensory-integration-by-convergence stage.

Long-range synchronization and multisensory convergence

i. Multisensory convergence: detecting sensory invariance

The multisensory convergence hypothesis is appealing because it provides a straightforward solution to the initial problem of how two or more separate streams of information are combined. However, the lack of specificity and the low spatio-temporal resolution of MN does not account for the different perceptual outcomes observed in the literature. For instance, the intersensory simultaneity threshold approximates 25ms in auditory-visual conditions (Hirsh & Sherrick, 1961; Stone *et al.*, 2001; Zampini *et al*, 2002). Some MN demonstrate a temporal tuning of ~25ms (Meredith *et al*, 1987), leading to the possible hypothesis: responses of the 'fine grained temporal resolution population' of MN (i.e. the 'enhanced-depression' neurons (Meredith *et al*., 1987)) are best adapted for detecting the temporal coincidence of multisensory events, and whenever the lagging modality input overlaps with the discharge rate within 25ms (initiated by the leading modality input), the MN output will mark 'simultaneity'. Following this hypothesis, why would the

temporal threshold of AV speech be as large as ~200ms (Massaro *et al*., 1996; Munhall *et al*., 1996)? If MN demonstrated a specificity to the complexity of sensory attributes, one could argue that MN populations respond selectively to the presence of particular patterns of inputs, and the populations of MN that tolerate larger stimulus disparities would then be recruited. To date, however there is no compelling evidence suggesting the validity of this hypothesis.

What, then, is the benefit of multisensory convergence from an information-theoretic perspective? *It is not merely that detecting redundant cross-modal information is insufficient to insure a robust perceptual representation but rather that the system needs further constraints to establish whether what has been detected is a perceptually valid event or not.* There is no question that multisensory enhancements may be efficient in boosting detectability, yet the very synergistic response pattern observed at multisensory convergence sites suggests that sensory-specific information is not entirely redundant and thus likely to be *functionally* different. For instance, irrelevant light stimuli enhance the detection of auditory stimuli embedded in noise (e.g. Lovelace *et al*, 2003), just as the detection of auditory speech is improved when supplemented with an articulating face (Grant & Seitz, 2000). A central role of MN may be to improve the detectability of ambiguous stimulation (i.e. noisy input), in which case MN would essentially validate or recalibrate incoming unisensory information. MN neurons are often intermixed with unisensory neurons; this is the case, for instance, in the SC where less than 50% of the neural population is bimodal or trimodal -i.e. multisensory (e.g. Wallace *et al*, 1998). The general principles characterizing the response patterns of MN are in fact consistent with an information-

theoretic approach, in which the gain of information provided by MN corresponds to the reduction of uncertainty in sensory-specific inputs (Patton *et al*, 2002).

Should we then consider that the gain of intelligibility and identification in auditory speech by the addition of facial visual information (e.g. Grant & Walden, 1996) and the displacement of auditory localization by a visual target (e.g. Bertelson & Radeau, 1981) originate from the same unique underlying neural mechanism? Probably not, because as observed in the visual system - and strongly suggested in the auditory - the nature of perceptual representations is fundamentally determined by the spatial and temporal resolutions of the underlying processing modules. In fact, detectability - i.e. the sufficient amount of information indicating a change in the environment - and perceptibility (or identification) - i.e. the sufficient amount of information to qualitatively categorize the perceptual world - are fundamentally different. In the example of intelligibility vs. ventriloquism effects, the former requires a fine temporal resolution and ultimately relates to the meaningfulness of the events, whereas the latter is based on a more global resolution.

What mechanism(s) need to supplement multisensory convergence for perceptual categorization? Assuming that multisensory-integration-by convergence permits accurate detection, the resulting unified output also loses its sensory-specificity. Specifically, featural information that contributes to the formation of a percept in the classic sensory-specific pathways is lost. Yet, we have seen that the contribution of each sensory modality must *functionally* differ. This argument is in line with prior mention of (i) various *levels of resolution* in multisensory interactions and (ii) various *types* of intersensory perceptual effects. For instance, if sensory-

specific information is lost, we cannot explain why visual information does not exert as strong a temporal modulation on auditory events as auditory does for visual processing. Hence, the extraction of sensory-specific information remains unresolved from a multisensory standpoint. The initial problem of how do sensory-specific attributes contribute to the formation a multisensory percept?

ii. Multisensory convergence: weighing unisensory specificity

Another strategy to maximize information is to suppress redundant information - i.e. another mechanism for reducing stimulus uncertainty (e.g. Rieke *et al.*, 1996). The lack of overlapping energies in intersensory inputs prevents that specific or fine resolution information - other than sensory invariance - be redundant. Sensory systems quantize the continuous flow of information into discrete or symbolic units of information transmission (e.g. Van Rullen & Koch, 2003). From a perceptual viewpoint, sensory-specific inputs can be seen as providing redundant information that differs by *nature* from sensory invariance in that they ultimately specify an abstract and discretized internal representation. For instance, the sight or sound of a piano ultimately evokes the perceptual representation of a piano, regardless of the input modality. Note however that the tagging of modality-input remains; one still 'knows' that a tune is *heard* and the white and black arrangement of the keys is *seen*. Assuming now that the specialization of sensory systems renders one system more efficient than the other at identifying a perceptual object – e.g. it is easier to figure out the melody of a piano from the sound rather than seeing the player's movements, even for a musician - there is no benefit in extracting similar 'meaning' from both modalities since one is noisy while the other is sufficient for

perceptual completion. As discussed earlier, the auditory system is well adapted to process time-dependent stimuli (transient stimulations), while the visual system is more efficient at processing spatial information. If perceptual redundancy can be weighted by the multisensory integrative system, it would also provide a means to maximizing incoming sensory-specific information.

MN are found throughout the cortex and provide a distributed multisensory network across major systems (Wallace *et al*, 2004). Not all multisensory areas are interdependent, however. For instance, MN found in the AES of cats projecting to the frontal cortices do not connect to those gating the SC (Jiang *et al*, 2001), suggesting that MN neural populations can form functionally distinct discrete networks. In contrast with the SC, the first spike latency in STS does not differ across sensory modalities (Benevento *et al.*, 1977), yet first spike latency and speed of processing differ in each sensory modality. For instance, auditory inputs reach the primary auditory cortices as early as 10ms post-stimulation (Celesia, 1976; Liégeois Chauvel *et al.*, 1991; Howard *et al.*, 1996) and visual inputs at about 50ms (e.g. Ffytche *et al.*, 1995, Maunsell *et al*, 1999). Provided that information reaches primary cortical areas at different latencies, sensory-specific information has seemingly undergone prior processing before converging onto the MN of STS - in agreement with a recent review (Schroeder *et al*, 2003). From a perceptual standpoint, this result is problematic as it pertains to the representational status of sensory -specific inputs at their arrival on multisensory convergence sites. If multisensory integration occurs after a certain amount of information has been extracted in the sensory-specific

stream, MN should also demonstrate a more specific tuning than what has been observed so far (Desimone & Gross, 1979), they should be tuned more specifically to this encoded information.

A second possibility is that MN need not be more specific to the type of incoming information because their function may reside in synchronizing inputs from various cortical areas - i.e. fundamentally, a coincidence detector role. In this view, MN are a necessary (because it reduces sensory coding scheme into one output) but non-specific hub mediating the dynamics of a more global inter-sensory network. MN would not then be a closing stage of conversion from sensory-specific to multisensory representations. If MN maintain or regulate the synchronization of neural sensory-specific ensembles, one would also predict that they do so at various stages of multisensory integration. Recent findings have shown that MN display oscillatory behavior and could maintain cortico-subcortical state-dependent synchronization at low (alpha range, ~10Hz) and high (gamma range, 30-70 Hz) frequency ranges (Brecht & Singer, 1998, 2001; Saito & Isa, 2003).

Maintaining a unified or *amodal* neural coding in parallel with sensory-specific encoding of information suggests that MN may have a role in regulating the attentional drive of sensory-specific streams. In particular, the 'reduction of uncertainty' hypothesis together with the neurophysiology of MN described earlier, would predict that *unisensory information can be weighed according to the gain of information* it provides for the maintenance of a unified (spatio-temporal coincident neural) event. For instance, recent fMRI studies of arbitrary multisensory pairings have shown a deactivation of sensory-specific cortices (Laurienti *et al*, 2002,

Bushara *et al.,* 2003). A similar modulation of activation has been reported for AV speech (Wright *et al*, 2003) and in various priming paradigms (e.g. Bagdayan, 1999, 2000). In divided attention paradigms, a modulation of early sensory-specific components has also been observed (Luo & Wei, 1990; Hohnsbein *et al*, 1991; Woods *et al*, 1993; Eimer & Schröger, 1998; Oten *et al*, 2000; Eimer *et al*, 2001; Oray *et al*, 2002). Together, these results suggest that the maintenance of multisensory flow of information may be crucial in the regulation of basic neurophysiology of sensory systems, such as state-dependent baseline activity in primary and secondary sensory cortices (e.g. Schulman *et al*, 1997; Pessoa*et al.* , 2003).

Another benefit of the multisensory cells as *inter-sensory relays* is that sensory-specific streams are concurrently preserved. As previously mentioned, each sensory modality is specialized in extracting types of potentially more efficient information than in the other modality, each one ultimately contributing to the hypothesized amodal representations. For instance, in AV speech, information pertaining to place of articulation is accessible via the auditory modality through upward or downward rapid frequency shifts or 'formant transitions' and in the visual modality through surface articulatory movements of the face or 'place-of-articulation'. If sensory-specific processing modules involved in the extraction of place-of-articulation can be brought together, *perceptually* redundant information could be computed. This working hypothesis will be posited in section 1.3 and throughout the manuscript.

Definition of multisensory integration

It is argued here that multisensory integration is a highly complex mode of information processing which is not reducible to one computational stage (i.e. that of multisensory convergence). Perceptual object formation necessitates analytic (featural) and synthetic (global) processes. Concurrent processes of multisensory information extraction are proposed here that involve highly dynamic interactions of multisensory and sensory-specific neural streams.

In particular, the existence of discrete multisensory convergence sites is proposed to insure a regular maintenance of crosstalk between sensory-specific streams, in what would now constitute a multisensory mode of information processing. Although the extraction of featural information remains specific to unisensory streams, MN now demonstrate a specificity of their own that is crucial in two ways. First, this view preserves the classic intersensory coincidence detector role of MN, that their neurophysiology readily illustrates. MN importantly maintain an 'amodal' mode of information processing that relies on sensory invariance - i.e. the very spatial and temporal mapping of the physical world. Second, the *coincidence detection in higher stages of information processing is not passive, but rather regulates or weighs the degree of sensory-specific information to be integrated* in the formation of a multisensory percept. Specifically, it is assumed here that the very *nature* of inputs provided by the sensory streams changes from a one-to-one relationship with the spectro-temporal characteristics of the stimulation to that of the

'inner (mental) world' or neural representations where spectro-temporal relationships are now intrisically defined by the neural inputs and outputs -i.e. the neural code.[5]

Multisensory integration is herein defined as the collection of neural dynamics that are necessary to insure the completion of a multisensory percept. Consequently, any type of classic local interactions may be hypothesized that include (mutual) facilitation, (mutual) inhibition or funnneling effects (Eijkman & Vendrik, 1965). Examples of various types of interaction in multisensory perception have been provided previously where, for instance, early sensory perception shows biasing effects and classic reports of faster reaction times to multisensory events may be related to mutual facilitation effects. Perceptual phenomena such as the McGurk illusion and AV speech in general will be extensively discussed throughout the following chapters.

1.3. Auditory-visual speech, an ecological case of multisensory integration

Because in natural conversational settings visual speech is readily available to the listener, it should come as no surprise that speech information in the auditory and visual channels may have evolved synergistically. Normal hearing and hearing-impaired populations benefit from looking at the interlocutor's facial gestures in auditory speech detection paradigms (Grant & Seitz, 2000) but also in disambiguating auditory utterances (Sumby & Pollack, 1954; Erber, 1975; MacLeod & Summerfield,

[5] Additionally, it is the architecture, the anatomical connectivity of MN, that fundamentally defines the *nature* of 'spatio-temporal' information. Hence, MN do not assign any specific 'meaning' but, rather, assign a representational value in that it regulates the dynamics of unisensory streams of information processing.

1987; Grant & Walden, 1998). One also benefits from the presence of visual speech inputs for extracting emotional content (e.g. deGelder *et al,* 1999). Speech theories do not always integrate visual speech as a source of natural inputs nor is AV speech integration explicitly accounted for in classic speech models (Green, 1996).

Classic findings in auditory-visual speech

A large body of evidence has shown that auditory and visual speech inputs interact in various perceptual contexts. For instance, a classic example of auditory speech mislocation is the ventriloquism effect, where the presentation of auditory speech whose source is located away from that of a moving face is mis-localized to or 'captured by' that of the visual source (e.g. Bertelson & Radeau, 1981; Driver, 1996). In the classic McGurk paradigm (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978), an audio [pa] dubbed onto a visual place-of-articulation [ka] is perceived as [ta]. This effect is commonly known as the McGurk effect or, more precisely, McGurk fusion and can be generalized across places-of-articulation in stop-consonants such that any bilabial dubbed onto a velar will result in misperceiving a alveolar (see Appendix A 1.3). Interestingly, dubbing an audio velar onto a bilabial place-of-articulation (McGurk combination) does not result in fusion but, rather, in 'combination', where the correct identification of audio and video are being combined in a more complex co-articulatory pattern, such as 'paka', 'kapa', etc… (McGurk & McDonald, 1976; MacDonald & McGurk, 1978). Within the speech domain alone, these two types of illusory outputs illustrate the complexity of AV interactions, and results from McGurk interactions suggest that the informational

content carried by each modality is not *equivalent* as far as their auditory-visual fusion / combination potentiality.

Auditory speech is typically sufficient to provide a high level of intelligibility (over the phone for instance) while performance by visual speech alone is much more difficult (Campbell, 1989; Massaro, 1996). The segmentation of information in the visual domain is further constrained by the rate of change in surface articulatory movements thus naturally limited to supra-segmental information (~80 to 200ms). While the discrimination between a bilabial (e.g. [ba] or [pa]) and a velar (e.g. [da] or [ta]) is easily achieved, visual information alone cannot disambiguate within place-of-articulation categories (e.g. bilabial class, [ba] vs. [pa]). Visually-based categories of contrast are called *visemes*, by analogy with phonemes in the auditory modality.

If the formation of AV speech percept can benefit from each modality, that is, if each modality provides non-redundant information, one would predict that the nature of modality-specific information will have an impact on the degree of AV speech integration. One approach to quantify AV interactions and the potential benefit of two sources of information over a single source is to degrade the information in one modality and observe what type of compensatory effects are obtained from the non-impoverished modality. For instance, the benefit of visual inputs in AV speech integration could be enhanced in better lip-readers. Measures of speechreading ability are indeed a good predictor of AV speech integration performance (Grant *et al.*, 1998), and recent findings also suggest that the efficiency of integration does not only depend upon the amount of information extracted in each sensory modality, but also in the variability of this information (Grant *et al.*, 1998).

In spite of their limited saliency, visual speech inputs robustly influence auditory speech even when degraded. Numerous filtering types do not totally attenuate the integration process (e.g. Rosenblum & Saldaña, 1996; Campbell & Massaro, 1997; Jordan *et al*, 2000; MacDonald *at al.*, 2000). These results suggest that, like in the auditory channels, multiple cues may be available from a facial display, including luminance patterns (Jordan *et al*, 2000) and kinematics (Rosenblum & Saldaña, 1996). However, it is remarkable that neither the gender (Walker *et al*, 1995; and see example appendix C) nor the familiar ity (Rosenblum & Yakel, 2001) of the face impacts the robustness of AV speech integration. The processing of visual speech appears functionally dissociated from the processing of faces, although it shares some of its sub-processes (Campbell, 1986, 1992). Case studies (in particular, prosopagnosic and akinetopsic patients) also suggest that both form and motion are necessary for the processing of visual and auditory-visual speech (Campbell, 1992; Campbell *et al.*, 1990, 1997). Visual speech was further proposed to access early on the phonetic module (Campbell, 1992), in which case visual inputs essentially provide place-of-articulation information. Given that visual speech representation is visemic and in light of the robust maintenance of AV integration with degraded visual inputs, it may act as a natural yet noisy channel of information for the speech system, its contribution may be regulated as a function of the needs for perceptual completion.

It was argued early by Campbell (1992) that lip-reading is a natural ability that one may have difficulty improving (contrary to reading ability). In light of

previous reports on the robustness of visual speech influences on auditory speech processing, this also suggests that neural underpinnings of AV speech integration may rely on a neural architecture that is at least pre-determined (see also, for instance, reports on the individuals' differences in AV speech integration (Grant *et al*, 1998)). The status of visual speech and auditory-visual speech as a natural ability present early on in life is in agreement with prior developmental studies.

This notion is further compatible with a supramodal speech code, which is amodal in nature (i.e. accessible through different sensory modalities) and characterized by the abstract representation of speech units based on a motor metric (e.g. Liberman, 1996). The postulation of abstract representations in speech processing is naturally associated with *perceptual* redundancy (as opposed to physical redundancy or *sensory invariance*) providing a crucial common framework for the integration of auditory-visual speech.

Speech and internal world

i. Information-theoretic models

In the auditory speech domain, a wealth of studies has permitted the definition of specific representational stages of information extraction. The general processing stages in auditory speech figure in Appendix B. Classic stages of information extraction are: (1) the acoustic evaluation stage, where auditory inputs are evaluated based upon acoustic features, such as frequency and amplitude; (2) the phonetic evaluation stage where extracted acoustic features are matched against possible

phonetic prototypes; (3) the phonetic features combination stage; (4) the phonological stage where phonemic representation is categorized; and (5) accesses the mental lexicon. Higher-processing levels then follow that engage the grammaticality of sentences (syntactic and semantic modules). Note that the processing from acoustic input to mental lexicon is essentially *serial* (Norris *et al*., 2000). The question elicited by this general information-theoretic approach is whether articulatory constraints can interfere with speech-specific features at various stages of their processing ((i) the early phonetic, (ii) the sensory memory and/or (iii) the phonetic feature combination stages). This question is precisely at the core of AV speech integration because visual speech inputs provide the speech system with articulatory-based information.

AV speech models take into consideration the fact that an early stage of integration is pre-phonetic; visual speech information could interact at either or some of the three phonetic representational stages. A late stage of integration, also defined as 'post-phonetic', would take place after acoustic and visual information have been categorized in their phonetic and visemic forms, respectively. Similarly, the issue of dependency assumes that sensory-specific information occurs either (i) at the featural level (dependent processing) or after sensory-specific categorization (independent processing) and thus prior to or after sensory-specific categorization, respectively.

ii. Timing of AV speech integration

The timing of AV speech integration is an ongoing debate, which is highly dependent upon the theoretical assumptions. For instance, classic evidence shows that the phonemic categorization is influenced in the context of visual inputs. The

McGurk effect is but one example, and Appendix C provides two other instances of categorical shift boundary induced by the dubbing of an auditory place-of-articulation continuum [pa]-[ta] or [ba]-[da] onto a visual place-of articulation [ka] or [ga], respectively. Within the context of a classically staged processing such as the one described above, visual speech can either be considered (i) a 'natural' input channel (comparable to the auditory channel) at any of these stages or (ii) an 'added element' that follows its own type of processing up to completion and comes to be integrated with the completed auditory product. If (i) is implicitly contingent on a supramodal speech code or an abstract representation of the sort, (ii) explicitly adds a new stage of integration that may or may not be speech specific.

Initial processing of inputs, whether auditory or visual, are unquestionably modality specific. The problem is knowing how early the sensory-specificity is lost. As was mentioned, visual speech processing functionally differs from other types of facial motion (Campbell, 1992). Visual speech is also a particular case of biological motion that happens to be socially relevant for our species and particularly relevant in a speech context. It is also the case that in the natural ordering of events in speech production, the movements of the articulators usually precede the actual utterance and as such, the optic flow is conveyed to the eyes earlier (and faster) than the acoustic flow to the ear. Yet for all of the natural and rapid dynamics that speech involves, classic accounts of speech processing are highly static (they are unidirectional in the early stages of information processing and do not readily consider the realm of neuro- and electro-physiological evidence for the existence of temporal windows of integration in both the auditory and visual sensory systems. Again, the underlying

computations are to be distinguished from the (readily available sensation of) continuous perception.

A major implication in considering neural dynamics based on temporal windows of integration is that the serial nature of information processing is compromised; if the overall ('global') hierarchy remains, windows of information extraction ('local') can now overlap such that one process (say at the 'auditory feature extraction stage') is still undergoing while the previous output of that same stage has already reached the following one. The major distinction here is not the hierarchy, *per se*, but the actual *rendering of the discretization* process. In particular, the analytical stages exist as a functional ensemble of discretization modules rather than as a serial and continuously incremented buffer of information processing. These dynamics naturally lose temporal resolution as constrained by the 'width' of temporal windows of integration but not totally considering that the potential overlaps of integrative windows themselves provide a source of temporal information (which may not be originally contained in the signal).

iii. Internal prediction

This computational context shaped by experience, modulated by the level of arousal, yet architecturally constrained, constitutes the basis of the 'internal' world or "what the state of the neural system is at any instant regardless of the incoming inputs". Consequently, the brain is not a passive analyzer of physical inputs but an *active predictive* device, whose local and global internal states dynamically interface with the external world.

These assumptions are present throughout the literature (e.g. Barlow, 1994) and most explicitly in the field of sensorimotor integration, where the action-perception loop is more readily quantifiable. More precisely, the working hypothesis is that the nervous system is capable of simulating a motor plan internally. This 'enactment' constitutes an internal prediction, which is compared with the actual motor realization. The adjustments of the motor behavior are made available through a constant comparison of the 'internal prediction' with sensory feedback. This hypothesis has been empirically tested and followed by 'forward models' of sensorimotor integration (Wolpert, 1995; Wolpert *et al.,* 1998; Mehta & Schaal, 2002; vanRullen & Koch, 2003; for a general overview see Friston, 2002).

On the perceptual side 'alone', Harth and colleagues (1987) proposed a model where *feedback* connectivity shapes the extraction of information early on in the hierarchy of the visual pathway. This initial conception of 'top-down' regulation is now complemented by the notion that *feed-forward* connections may not carry the 'extracted information' *per se* but rather the residual error between the 'top-down' internal predictions and the incoming inputs (Rao & Ballard, 1999). The hypothesis that this regulation occurs early on in the analysis of sensory inputs is currently tested in vision - as early as V1- (Sharma *et al.,* 2003), and neural dynamics from population to synaptic level have provided promising results (Mehta, 2001; Shin, 2002).

The working hypothesis for auditory-visual speech integration

Early on, Stevens & Halle (1962, 1967) proposed such 'analysis-by-synthesis' mechanism in auditory speech perception, and it may well constitute the first explicit forward model in the literature, from a perceptual standpoint. The 'analysis-by-synthesis' proposal has thus far remained in the shadow of the more established Motor Theory of Speech Perception (Liberman *et al.*, 1967; Liberman & Mattingly, 1985). Both views borrow from the motor representations of speech used in production for the perception of acoustic inputs. However, a crucial difference between the 'analysis-by-synthesis' (AS) description and that of the Motor Theory (MST) resides in the very notion of *internal prediction* and its instantiation as a forward mechanism. The former explicitly posits a comparative mechanism interfacing between the motor speech representations and the evaluation of acoustic inputs, where the resulting output is the 'residual error' between the internal prediction and the sensory input, similar in concept to sensorimotor integration.[6]

In the following chapters, I will extend the original 'analysis-by-synthesis' model of Stevens & Halle (1962, 1967) to AV speech perception and provide evidence for a forward model of AV speech integration. The general theme of the reported experiments is that of auditory-visual speech in neural time.

[6] A neural account for such an action-perception interface has recently built on the 'mirror neurons' described in monkeys' prefrontal cortex and in premotor cortices (e.g. Kohler *et al.*, 2002; Ferrari *et al.*, 2003). The neural responses of mirror neurons correlate with the observation of *intended* motor commands rather than their explicit executive function; that is, mirror neurons respond upon observation of an intended gesture rather than to the actual action. Most recently AV mirror neurons were described in the monkey premotor cortex (Keysers *et al.*, 2003). These neurons respond to either heard or seen actions. The exact nature of information encoded by mirror neurons is under debate and their possible involvement in forward type of connectivity is being discussed (Miall, 2003).

In Chapter II, I characterize the temporal window of integration, thereby the temporal resolution of the AV speech integrative system.

Chapters III and IV provide electrophysiological (EEG) evidence for the neural dynamics underlying the integrative process of AV speech. The results will be interpreted in the context of a forward model of AV speech processing.

In Chapter V, I initiate a different approach in characterizing possible levels of multisensory interactions that may enable one to differentiate the neural underpinnings of speech versus non-speech processing.

Chapter VI will provide a more in depth commentary on forward modeling in AV speech perception in relation to current dominant models of auditory and auditory-visual speech perception.

Chapter 2: Temporal window of integration in bimodal speech perception

"Comment cette flèche (du temps) apparaît-elle quand on observe l'événement *dans son ensemble*, c'est-à-dire quand on considère la série des évènements simples? […] *La flèche du temps serait uniquement inscrite dans le passage du moins probable au plus probable*. […] Si tel est le cas, l'irréversibilité du temps psychologique serait évidemment une pure illusion."

"How does the arrow (of time) emerge when an event is observed *as a whole*, i.e. if we consider the series of simpler elements? […] *The arrow of time would only be a consequence from the least probable state to the most probable state […] If such is the case, the irreversibility of psychological time would be, of course, a pure illusion"*

Hubert Reeves

Forty-three normal hearing participants were tested in two experiments, that focused on temporal coincidence in auditory visual (AV) speech. In these experiments, audio recordings of [pa] and [ba] were dubbed onto video recordings of a face articulating [ka] or [ga], respectively (ApVk, AbVg), to produce the illusory "fusion" percepts [ta], or [da] (McGurk and McDonald, 1976). In Experiment 1, an identification task using McGurk pairs with asynchronies ranging from -467 ms (auditory lead) to +467 ms was conducted. Illusory fusion responses were prevalent over temporal asynchronies from -30 ms to +170 ms and more robust for audio lags. In Experiment 2, simultaneity judgments for incongruent and congruent audiovisual tokens (A_dV_d, A_tV_t) were collected. McGurk pairs were more readily judged as asynchronous than congruent pairs. Characteristics of the temporal window over which illusory fusion responses and subjective simultaneity judgments were maximal were quite similar. The 200ms duration and the asymmetric profile of the temporal window of integration are interpreted in the context of a forward model of AV speech integration based upon recent neurophysiological and brain imaging findings.

2.1 Introduction

In natural conversational settings, both auditory and visual information are important for speech perception. Although auditory information alone is usually sufficient to perceive spoken discourse, a congruent facial display (articulating the audio speech) provides critical cues in noisy environments (MacLeod &

Summerfield, 1990; Helfer, 1997) and benefits hearing-impaired listeners (Grant *et al.*, 1998). While typically supporting the perception of audio speech signals, visual speech information can also alter the expected perceptual interpretation of clear audio signals. The 'McGurk effect' demonstrates that adding conflicting (incongruent) visual information to an audio signal alters the auditory percept. The presentation of an audio /pa/ (bilabial) with a synchronized incongruent visual /ka/ (velar) often leads listeners to identify what they hear as /ta/ (alveolar), a phenomenon referred to as 'fusion' (McGurk & McDonald, 1976).

In the McGurk fusion, the nature and content of auditory-visual (AV) information are fundamentally different, yet sensory inputs converge on a unique percept clearly differing from the initial unimodal percepts. This illusion permits one to *quantify the degree of multisensory integration* in a domain specific context (speech) that is ecologically relevant to humans. In the present experiments, we take advantage of the McGurk illusion to explore the temporal boundaries of AV speech integration and their implications for underlying neural integrative processes.

On the basis of cortical mechanisms known to be involved in multisensory processing, two major neural implementations and predictions for AV speech integration can be stated.

The neural convergence hypothesis

First, the existence of subcortical and cortical multisensory convergence sites (e.g., the Superior Colliculus (SC) and the Superior Temporal Sulcus (STS),

respectively) has lead to the ‘neural convergence hypothesis’ (Stein & Meredith, 1993; Meredith, 2002). Sensory inputs from different modality feed forward to sites of neural convergence, where information is integrated as recorded through an enhanced or ‘supra-additive’ neural response. Calvert (2001) extends the neural convergence hypothesis by proposing that multisensory sites of integration feed back onto unisensory cortical sites - for instance onto auditory associations areas for speech-specific processing. In this implementation, multisensory neurons underlie the integration process, and one would predict that the extent of desynchronization tolerated by AV speech should reflect the temporal tuning properties of multisensory neurons. Specifically, multisensory neurons have been shown to be very tolerant to stimuli desynchronization -as much as 1.5 s- providing that the sources of information remain within the same spatial location (Meredith *et al*, 1987).

From a perceptual standpoint, very large tolerances to AV asynchronies have indeed been reported by Massaro *et al.*(1996). In their study, two groups of participants were tested on their ability to identify synthetic and natural AV speech tokens in a factorial design including syllables /ba/, /da/, /th-a/ and /va/. The first group was tested with AV asynchronies of +/-67, +/-167and +/-267ms, and the second group with AV asynchronies of +/-133, +/-267and +/-500ms. Congruent AV identification was overall less affected by asynchronies than the incongruent pairs. Although no clear boundary for AV integration could be drawn from the psychophysical measurements, a fit of the Fuzzy Logical Model of Perception (FLMP) to asynchronous AV tokens indicated that a slight decrease of the model performance occurred at +/-267ms, while a significant breakdown occurred at +/-

500ms. These results suggest that the integrative process tolerate asynchronies on a ~1 second scale, in agreement with the temporal resolution of multisensory neurons.

Importantly, the tuning properties of multisensory neurons -thus, their response characteristics - rely on the spatial relationships between multisensory events. The McGurk effect remains surprisingly robust under AV spatial disparities (Jones & Munhall, 1997) as well as under spectral manipulations such as filtering of facial information (Campbell & Massaro, 1996; McDonald *et al*, 2000) or a points - of-light facial display (Rosenblum & Saldaña, 1996). The prevailing contribution of facial kinematics in AV speech integration finds further support in a study reported by Jordan *et al.* (2000) indicating that luminance distribution may enhance information drawn from facial articulatory movements. Taken together, these results suggest that the neural mechanisms underlying AV speech integration are efficient at extracting dynamic cues that are potentially informative to the speech recognition system, in spite of poor spatial resolution. As such, specialized processing pathways should be considered that are capable of handling the complex dynamics of facial articulators (Munhall & Vaitikiosis, 1998). In particular, while the activation of multisensory sites to multisensory stimulation has been widely reported, the functional implication of such activation remains controversial in light of recent fMRI findings (e.g. Laurienti *et al.*, 2002). Multisensory neurons permit a rough estimate of coincident multisensory inputs, regardless of the specific nature of the stimulation and may enhance detectability of stimuli (Stein & Meredith, 1993, Stein *et al*., 1996). However, it is unlikely, as recent findings suggest, that multisensory sites of

integration stand alone in the computations of complex and perceptually constrained stimuli such as AV speech.

Large-scale neural synchronizations

For instance, a second neural implementation for AV speech integration finds support in recent brain imaging studies, where multisensory integration appears to involve a global subcortical and cortical neural network (e.g. Calvert, 2000, 2001). From a functional standpoint, the anatomical spread of activation seen in fMRI studies suggests an important role for dynamic large-scale neural computations. Indeed, recent electroencephalographic (EEG) recordings during AV speech presentation have shown that the gamma band (40-70Hz) power increases for a period of ~150ms (Callan *et al*, 2001). The 'gamma' sampling resolution (~30ms) and the sustained gamma activation over ~150ms emerge as two important times scales. The gamma frequency range is often reported as the response underlying 'cognitive binding' (e.g, Tallon-Baudry *et al*, 1997; Rodriguez *et al*, 1999), while ~150ms (low-frequency oscillations or theta range, 4-7Hz) also observed under bimodal stimulation (Sakowitz *et al.*, 2000), is implicated in temporal encoding (for review, see Buzsáki, 2002). Perceptual correlates in these time ranges have often been reported in speech-related studies and have provided the basis for the 'Asymmetric Sampling in Time' model proposed by Poeppel (2003). According to Poeppel's proposal, perceptual unit formation evolves on a ~200ms time-scale (i.e. the theta range), a time constant characteristic of the syllabic unit of speech (Arai and

Greenberg, 1999), while featural aspects, are processed on a ~30ms time-scale (i.e. the gamma range).

In this second neural implementation, one would predict that desynchronization of AV speech inputs should be tolerated as far as the synchronization of neural populations is not perturbed, i.e. within the time-constant of neural population co-activation, which underlies the integration process. We have previously mentioned that surface articulatory dynamics provide crucial temporal signals that may cue AV integration, as was also suggested by Summerfield (1987). One of the most salient features in natural AV speech conditions involves the lip movements, which correlate with the corresponding overall amplitude of the acoustic signal (e.g. Rosen, 1992; Grant *et al*, 2001). If such AV correspondence drives the integration process, tolerance to AV asynchrony may be governed by one's ability to estimate the synchronicity of acoustic amplitude fluctuations and facial kinematics, which evolves in the 3-4Hz range (Rosen, 1992, Grant *et al*, 2001). The time scale over which the audio and visual signals evolve is that of the speech unit formation, or syllable (Poeppel, 2003; Arai & Greenberg, 1998). In the hypothesis that temporal matching of acoustic amplitude and visual facial kinematics intervenes in the cross-modal binding of information, one would predict that incongruent AV stimuli (McGurked pairs) should show a decreased tolerance to AV asynchrony –i.e. the inherent temporal decorrelation function induced in mismatched stimuli is further accentuated by temporal misalignments. Possible neural mechanisms underlying this computation will be described in the discussion section.

Auditory-visual speech asynchronies revisited

When considering temporal relationships, both for synthetic and natural AV speech inputs, an interesting profile has emerged when the temporal coincidence of AV events is manipulated: AV integration of speech does not seem to require precise temporal alignment (Dixon and Spitz, 1980; McGrath and Summerfield, 1985; Pandey et *al.*, 1986; Massaro, *et al.*1996, 1998; Grant, 2001) , and accurate AV speech recognition is maintained over a temporal window ranging from approximately -40 ms audio lead to 240 ms audio lag. Additionally, in an extension of the Massaro *et al.* study (1996), Massaro (1998) shows that desynchronizations in the order of 150ms start perturbing the integration of AV speech, yet clear disruptions of the FLMP fit were obtained for asynchronies larger than ~500ms.

The specific goal of this study builds on results by Munhall *et al*. (1996). In the first of the reported set of experiments, Munhall *et al.*looked at the effect of asynchrony and vowel context in the McGurk illusion. Their stimuli consisted of audio utterances /aba/ and /ibi/ dubbed onto a video of a face articulating /aga/. The range of asynchronies tested spanned from -360ms auditory lead to +360ms auditory lag in steps of 60ms. Two main results reported by Munhall *et al.* are of particular interest for the present study. First, the response distributions obtained over all ranges of asynchronies show that the fusion rate (/d/ percepts) remained close to or below 10%. Auditorily driven responses dominated for most asynchronies, and visually driven responses (/g/) were prominent from -60ms audio lead to +240 ms audio lag. These results raise a crucial question regarding the effective occurrence of AV

integration in this experiment. A conservative definition of AV speech integration entails that a unitary integrated percept emerges as the result of the 'combination' of auditory and visual information at some stage of the speech-processing pathway. When the percept /g/ (visually-driven response) dominates near AV synchrony, it remains unclear whether a case of visual dominance or a manifest integrated percept is being instantiated. Although we recognize that most AV speech studies have considered any deviation between the percept in audio alone and in bimodal condition to be a case of AV integration, we here consider that the rate of unimodal error – particularly the error in the visual alone condition which may be identical to the fusion percept (such as a /d/ response to a visual alone /g/ and a combined audio /b/ and visual /g/) – needs to be taken into consideration to enable one to distinguish between AV integration and unimodal error. For example, consider an individual who has a fusion rate of 90% at synchrony (i.e. reports [ta] when presented with an audio [pa] dubbed onto a place of articulation [ka]). This same individual perceives an audio /pa/ as /ta/, 2% of the time (auditory error) and a video /ka/ as /ta/ 30% of the time (visual error). The initial fusion rate can be accounted for unimodal error rates 32% of the time, leaving 58% of [ta] reports non-accounted for but by AV interaction. Hence in the Munhall et al. study, the reported asynchrony function may be an example in which visual information dominates the auditory percept. If this was the case, it is the temporal resolution of visual speech processing that is being shown, not the influence of asynchronies on AV integration *per se*. This issue is essential when considering which neural system may underlie the multisensory integration process.

Secondly, Munhall et al. reported a V-shaped function for auditory driven responses (/b/) with a minimum around 60ms, suggesting that synchronous auditory and visual stimuli may not be optimal for AV integration. Pair-wise comparisons of the proportion of /b/ responses across the different temporal conditions revealed that the responses at synchrony were significantly different from those at -60ms (auditory lead) and at 240ms (visual lead). However, because temporal asynchronies were only tested in steps of 60ms, it is unclear whether temporal misalignments between 0ms and -60ms or between 180ms and 240ms would also significantly impact AV integration. Again, one needs to determine whether significant changes in perception occur in steps of 30ms for instance, as it specifically relates to the temporal resolution (i.e., the frequency range) of the neural integrative system.

While suggesting possible temporal limitations of AV integration, Munhall's study used a fairly coarse temporal granularity to investigate the effects of asynchronous AV input on speech recognition, and therefore clear boundaries and resolution for the influence of visual information on auditory speech processing could not be drawn. As we pointed out, a more fine-grained profiling of the effects of asynchronies on AV integration is necessary to connect perceptual effects with their neurophysiological processes.

Motivated by these considerations, two experiments were conducted which explored the tolerance of the McGurk effect to a broad range of AV temporal disparities. The first experiment investigated the effect of AV asynchrony on the identification of incongruent (McGurk) AV speech stimuli. The second experiment

focused on the subjective simultaneity judgments for congruent and incongruent AV speech stimuli tested in the same asynchrony conditions. The present studies extend the results of Munhall et al. (1996) by using smaller temporal step sizes, increasing the range of tested asynchronies, and determining the boundaries for subjective audiovisual simultaneity. Specifically, we addressed the following questions: (i) For which stimulus onset asynchronies (SOAs) was the fusion response dominant over the auditory or visual driven response? (ii) Is the temporal window for subjective audiovisual simultaneity equivalent to the temporal window for perceptual fusion, and is this the same for congruent (matched audio and visual stimuli) and incongruent (mismatched auditory and visual stimuli) speech input?

2.2 Materials and Methods

Participants

Participants (native speakers of American English) were recruited from the University of Maryland undergraduate population and provided written informed consent. Two groups of participants took part in this study. The first group included twenty-one participants (11 females, average 21 years) who were run in the voiced A_bV_g condition (A_bV_g: audio /ba/ and video /ga/). The second group consisted of twenty-two participants (8 females, average 22.5 years) who were run in the voiceless A_pV_k condition (A_pV_k: audio /pa/ and video /ka/). No participant had diagnosed hearing problems and all had normal or corrected-to-normal vision. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli

i. Video and Audio Processing

Movies drawn from a set of stimuli used in Grant *et al.* (1998) were digitized from an analog tape of a female speaker's face and voice. An iMovie file was created unchanged from the original with an iMac computer (Apple Computer, CA). The iMovie was then segmented into each token (A_bV_b, A_dV_d…) and compressed in a Cinepak format. Each stimulus was rendered into a 640x480 pixels movie with a digitization rate of 29.97 frames per second (1 frame = 33.33ms). The soundtracks were edited using Sound Edit (Macromedia, Inc.). Each soundtrack was modified to produce a fade-in and fade-out effect over the first and last 10 ms. Stereo soundtracks were digitized at 44.1 kHz, with 16-bit amplitude resolution.

ii. Generation of McGurk pairs

Audio /ba/ and /pa/ were extracted from natural utterances produced by the same female speaker and then dubbed onto video /ga/ and /ka/, respectively, to form the McGurk pairs. Both voiced and voiceless McGurk pairs were tested, in order to insure generalizability. For each McGurk pair, the consonantal burst of the digitized audio file (e.g., /ba/) was aligned with the consonantal burst of the underlying audio portion of the video file (e.g., /ga/) to within +/ - 5ms (temporal resolution limited by the editing software).

iii. Audiovisual alignment in asynchrony conditions

Audio-visual asynchronies were created by displacing the audio file in 33.33 ms increments (frame unit) with respect to the movie file. This process resulted in the creation of stimuli ranging from (+) 467 ms of auditory lag to (-) 467 ms of auditory lead. Thus, a total of twenty-nine stimulus conditions (28 asynchrony conditions and 1 synchrony condition) were used in the study.

Procedure

Both identification (Experiment 1) and simultaneity judgment (Experiment 2) were designed using Psyscope (version 1.1) together with QuickTime extension (QT OS 8). Responses were recorded using a button box connected to a Mac G4 through a USB Keyspan adapter (28X). Individual responses were recorded on-line.

Identification and subjective simultaneity experiments took place in a dimly lit, quiet room. Participants were seated at about 65cm from the visual display, with the movie subtending a visual angle of 7.5º in the vertical plane and 9.5º in the horizontal plane. Videos were displayed centered on a 17" G4 monitor on a black background. The luminance of the video display was suprathreshold for all stimuli insuring that no difference in response latency was artificially induced due to low luminance contrast. Sounds were presented through headphones (Sennheiser, HD520) directly connected to the computer at a level of approximately ~70 dB SPL.

The average duration of the AV stimuli used in both experiments was 2590 ms, including video fade-in (8 frames) and fade-out (5 frames). Interstimulus

intervals (ITIs) were randomly selected among 5 values (500 ms, 750 ms, 1000 ms, 1250 ms and 1500 ms). For both voiced and voiceless conditions, the identification task (Experiment 1) and simultaneity judgment task (Experiment 2) were run separately. Each participant took part, successively, in the identification experiment (e.g., A_bV_g) followed by the subjective simultaneity judgment experiment with the same McGurk pair and congruent counterpart (e.g., A_bV_g and A_dV_d). The task requirements were given prior to each experiment; importantly, participants were unaware of AV asynchronies prior to the identification task. For both identification and simultaneity judgment tasks, no feedback was provided and no training was given prior to testing.

i. Experiment 1 – Identification Task

The identification task contained 10 presentations of each timing condition (29 timing conditions x 10 repetitions/condition in both A_bV_g and A_pV_k blocks for a total of 290 trials per block). In addition, for A_pV_k identification, ten trials each of audio-alone /pa/ and visual-alone /ka/ were included to obtain an estimate of unimodal identification performance. Thus, for A_pV_k identification there was a total of 310 trials per subject. A single-trial 3 alternative-forced choice (3AFC) procedure was used. Participants were asked to make a choice as to "what they hear while looking at the face" in AV condition and additionally for A_pV_k, "what they heard" in A alone conditions and "what the talker said" in V alone conditions. Three choices were given. In the A_bV_g pair, participants could answer /ba/, /ga/, or /da/ or /th-a/. The

options /da/ or /th-a/ were mapped onto a single response button. In the A_pV_k pair, participants could answer /pa/, /ka/, or /ta/. Note that for both AV stimuli the first response category corresponds to the auditory stimulus, the second to the visual stimulus, and the third to the fused McGurk percept.

ii. Experiment 2 – Subjective Simultaneity Judgment Task

The simultaneity judgment task contained 6 repetitions of each timing condition for either McGurk pair (A_bV_g and A_pV_k) and for either natural congruent pair (A_dV_d and A_tV_t), for a total of 696 trials per subject. Stimuli were pseudo-randomly intermixed. A single-trial 2 alternative-forced choice (2AFC) procedure was used.

Following Experiment 1, participants were asked to give their impressions of the difficulty of the task. All participants reported being aware of some cases in which A and V stimuli were not aligned in time. Participants were informed that AV synchrony was, in fact, manipulated and that in a second experiment participants' sensitivity to AV asynchrony is explored. Participants were thus asked, in Experiment 2, to determine if the time alignment of A and V stimuli was accurately rendered during the dubbing process and whether the auditory and the visual utterances were synchronized. Participants were told not to pay attention to the identity of the stimuli but rather to focus on the temporal synchrony of the stimuli. Participants were given two choices: "simultaneous" or "successive". They were told that the order did not

matter in the ‘successive’ case and that they should press this button whether the auditory or the visual appeared to come first.

Analysis

Responses were sorted and averaged for each participant and each timing condition. A grand average of each possible response per timing condition was then computed across participants. The analysis of participants’ performance for each timing condition revealed that four participants (out of forty-three) showed a constant average fusion rate of *less than 40%* regardless of asynchrony. They were not considered for further analysis (three participants in the A_bV_g condition and one participant in the A_pV_k condition showed an average of 22% fusion rate for all asynchronies). Reported paired-t-tests for establishing the temporal window of integration boundaries were submitted to a Bonferroni correction.

2.3 Results

Experiment 1 : Identification Task

i. Voiced McGurk pair A_bV_g

Figure 1 shows the distribution (in percent) of each of the three possible response categories (/ba/, /ga/, /da/ or /th-a/) as a function of SOA (N=18). Auditory-

visual /ga/ responses (visually driven responses) were seldom given, whereas /ba/ (auditorily driven responses) and /da/ or /th-a/ fusion responses formed the majority of responses. The overall trend shows that as the asynchrony between the AV utterances increases, /ba/ judgments increase, whereas /da/ or /th-a/ judgments (fusion responses (FR)) decrease. An analysis of variance across SOAs shows a significant influence of asynchrony on fusion rate ($F(1, 28) = 9.242$, $p<0.0001$). Unimodal stimuli were not collected for this pair, and therefore correction of fusion rates by unimodal errors could not be calculated (cf. Results section 1.2). A Fisher's PLSD test applied to uncorrected fusion rate across SOAs showed a range of non-significantly different SOAs between -133 ms and +267 ms. The temporal boundaries of the fusion rate plateau (SOAs at which fusion was maximal) were calculated on the basis of an asymmetric double sigmoidal (ADS) curve fitted to the average fusion rate function. A confidence interval of 95% was chosen to determine the asynchrony values at which the fusion rate was significantly different from that obtained at synchrony. Using an ADS fit ($r^2 = 0.94$) and a 95% confidence limit, a fusion rate plateau was determined to be from -34 ms auditory lead to +173 ms auditory lag. Moreover, the ADS fit confirms the asymmetrical profile of fusion responses and also suggests an off-centered peak towards auditory lag at about + 69 ms (cf. Table 1).

ii. Voiceless McGurk pair A_pV_k

Figure 2 shows the proportions (in percent) of each of the three possible response alternatives (/pa/, /ka/, or /ta/) as a function of SOA (N=21). Comparable to

the A_bV_g condition, auditory-visual /ka/ (visually-driven) responses have the lowest probability of occurrence, whereas /pa/ (auditorily-driven responses) and /ta/ judgments (fusion) occur frequently and are clearly affected by audio delay. As the AV asynchrony increases, /pa/ judgments (auditorily driven responses) increase while /ta/ judgments (fusion responses) decrease.

In interpreting the bimodal responses to incongruent audio-visual stimuli, it is important to consider the particular errors that might be made by audio-alone and visual-alone processing. This is particularly relevant for visual-alone processing, where error rates can be quite high. Thus, since visual /ka/ is sometimes perceived as /ta/ it is possible that /ta/ responses to the audio-visual token A_pV_k may in fact be visual-alone driven responses rather than a fusion response representing true bimodal processing. One method for dealing with this potential confound is to use the unimodal error rates to normalize the bimodal fusion response rates. This procedure will generate a more conservative estimate of fusion. In the A_pV_k condition, audio alone and visual alone identifications were collected. Individual fusion rates for the A_pV_k condition were corrected on the basis of the individual's confusions in unimodal conditions (especially in the visual domain) in order to insure the bimodal nature of the fusion response. For example, consider an individual who has a fusion rate of 90% at synchrony. This same individual perceives an audio /pa/ as /ta/, 2% of the time (audio error) and a video /ka/ as /ta/ 30% of the time (video error). The corrected fusion rate (CFR) based upon of the individual's unimodal error rates becomes 58% (measured fusion rate minus audio error and visual error). The corrected fusion rates for each asynchrony value were averaged across participants and compared with the

averaged rate of /ta/ responses that would be expected solely on the summation of unimodal error responses /ta/ to an audio alone /pa/ (average of 0.05, N=21) and a visual alone /ka/ (average of 0.48, N=21). If the fusion rate is superior to the sum of error rates in unimodal conditions (i.e. superior to 0.53 (0.05+0.48)), unimodal error rates do not suffice to account for /ta/ responses in the bimodal condition.

Figure 2 illustrates that participants reported perceiving the 'fused' /ta/ over a wide range of audio-visual asynchronies. Auditory-visual /ta/ responses were compared to the unimodal /ta/ occurrences in auditory-alone and visual-alone conditions. The resulting values therefore indicate *true bimodal responses.* An analysis of variance across SOAs shows a significant influence of asynchrony on fusion rate ($F(1, 28) = 4.336$, $p<0.0001$). SOAs at which the fusion rate exceeds the averaged summation of error rate value (constant) correspond to the limits at which unimodal error responses /ta/ to an auditory /pa/ (5%) or to a visual /ka/ (48%) may account for the /ta/ response in bimodal condition A_pV_k. According to this definition, true bimodal fusion responses were observed from - 167ms of auditory lead to + 267ms of auditory lag. These same limits were obtained by applying a Fisher's PLSD at 95% confidence to the effect of SOAs on fusion rate ($p<0.0001$).

Fitting results ($r^2 = 0.98$) showed that the fusion rate (FR) at SOAs ranging from -25 ms of auditory lead to +136 ms of auditory lag did not significantly differ from the fusion rate obtained in the synchrony condition. The ADS fit also confirms the asymmetrical profile of fusion responses and suggests an off-centered peak, towards auditory lag of about + 55 ms (cf. Table 1).

Experiment 2 : Simultaneity Judgment Task

i. McGurk pair A_bV_g - Congruent pair A_dV_d

Figure 3 shows that the rate of simultaneity judgments for both the McGurk pair A_bV_g and the congruent pair A_dV_d decreased as the asynchrony between audio and video stimulus components increased. At synchrony (0ms SOA), the congruent pair A_dV_d was judged 98% of the time to be simultaneous whereas A_bV_g reached a simultaneity rate of only 74% (N=18). An ADS fit allowed defining the boundaries of the simultaneity plateau in both conditions with 95% confidence. The limits of the plateau, as defined by the ADS fitting procedure, resulted in a temporal window of integration ranging from -73 ms to +131 ms for the congruent pair ($r^2 = 0.98$) and from -36 ms to +121 ms for the incongruent pair ($r^2 = 0.98$).

A paired t-test between congruent and incongruent tokens across SOA's revealed a significant difference between the two simultaneity rate profiles ($p<0.0001$). The incongruent A_bV_g pair was associated with a smaller temporal window and an overall lower rate of simultaneity judgments compared to the congruent profile (cf. Table 1).

ii. McGurk pair A_pV_k - Congruent pair A_tV_t

As with the A_bV_g and A_dV_d conditions, Figure 3 shows that the percentage of simultaneity judgments on both the McGurk stimulus A_pV_k and the congruent

stimulus A_tV_t decreased as the asynchrony between audio and video stimulus components increased. At synchrony (0 ms SOA), the congruent pair A_tV_t was judged 95% of the time to be simultaneous whereas the incongruent A_pV_k reached a maximum simultaneity rate of only 80% (N=21). Using the ADS fitting procedure and a 95% confidence limit to define the boundaries of the simultaneity plateau for each stimulus condition resulted in a range from -80 ms of auditory lead to +123 ms of auditory lag for the congruent pair ($r^2 = 0.99$) and -44 ms to +117 ms for the incongruent pair ($r^2 = 0.98$). A paired t-test between the simultaneity rate for congruent and incongruent tokens across SOAs revealed a significant difference between the two data series ($p<0.0001$). Similar to the trend observed for the A_bV_g McGurk pair, the incongruent simultaneity profile revealed a smaller temporal window and an overall lower rate of simultaneity judgments as compared to the congruent profile (cf. Table 1).

2.4 Discussion

Two experiments were conducted to examine the effects of audiovisual temporal asynchrony on syllable identification and simultaneity judgment. The major finding was that AV speech inputs are extremely tolerant to asynchrony, and that bimodal information separated in time by as much as 200ms is usually perceived as simultaneous. Specifically, both the identification experiment and the subjective simultaneity judgment experiment revealed temporal windows of maximal AV integration of about 200ms. Information-processing windows of similar duration have

been suggested as a basis for perceptual unit formation in the auditory cortices (Näätänen, 1992; Yabe *et al.*, 1997; Winkler *et al.*1998; Loveless, 2001; Poeppel, 2001; Yabe *et al*, 2001a; Yabe *et al*, 2001b; Po eppel, 2003). Providing further evidence for the 'discrete perception' view (VanRullen & Koch, 2003), the emergence of a temporal integration window in AV speech can be accounted for by inherent dynamics of cortical neurons – such as long-range synchronization (Rodriguez *et al*, 1999) -, and neural convergence on multisensory sites are but one possible step in the integration of biologically complex spectro-temporal stimuli.

With regard to prior studies

The temporal boundaries are overall consistent with the observations of Munhall *et al.* (1996) and extend their findings by sampling many more asynchrony values. Our fusion rates in both McGurk conditions (with and without error rate correction) are well above the fusion rates reported by Munhall *et al.*(199 6). We can only hypothesize that these differences are stimuli-related and/or task-related, since both studies provided a closed-set response choices. Furthermore, Munhall *et al.* (1996) mentioned that their participants' responses showed great variability. In our study, all but four participants showed uniform responses (see reported standard errors in figs 1,2 and 3 and method section).

Although the fusion rate remains resilient outside the plateau of integration established by ADS fitting (e.g., from -167ms to 267ms for A_pV_k), maximal true bimodal fusions (i.e. corrected fusion rates) cluster within ~200ms. Both the

integration window and the larger range of true bimodal interaction remain well below the estimated 500ms breakdown suggested in an earlier study by Massaro *et al.* (1996). One possible difference may result from the conservative approach that was taken here, first in our choice of stimuli, by considering that an integrated AV percept results from two distinct unimodal inputs, and second, by our correcting the measured fusion rate, insuring that unimodal errors could not account for the integrated percept. Some methodological differences, such as our choice of variable inter-trial intervals, may also contribute to the discrepancies in the estimate of the temporal integration window boundaries, although we feel it is unlikely that these values have a significant impact on our results.

Bimodal speech (congruent and incongruent) appears to tolerate much larger asynchronies than has been reported for non-speech stimuli (Dixon and Spitz, 1980) and argues for temporal integration far beyond the classical notion of simultaneity and temporal order threshold established with simpler non-speech stimuli within and across sensory modalities (Hirsh & Sherrick, 1961; Zampini *et al., 2002).* Although estimates of the 'point of subjective simultaneity' (PSS) for simpler stimuli range within a 100ms window of asynchronies (Stone *et al*, 2001; Lewald et al., 2001), two important points should be raised. First, PSS values, as pointed out by Stone *et al* (2001), are highly variable across participants in contrast with the consistent pattern we found across participants in our experiments. Second, the inter-individual variability in PSS estimates is contained within the limits of the temporal window of integration for AV speech stimuli (i.e. approximates 100 ms) but reported PSS do not delimitate an individual's *plateau* of subjective simultaneity, rather, each participant

is characterized by one PSS value, before and after which AV asynchronies affect the performance.

AV coherence hypothesis

The subjective simultaneity judgment experiment comparing incongruent (A_bV_g and A_pV_k) and congruent (A_dV_d and A_tV_t) syllables allows one to evaluate the processing of illusory versus real speech percepts. According to a recent study by Grant and Greenberg (2001), the level of coherence between area of mouth opening and acoustic amplitude envelope can play a significant role in AV speech integration. In particular, the acoustic dynamic envelope and facial kinematics are correlated to a greater degree in the congruent than in the incongruent case. If such AV coherence is computed at the neural level, one would predict that, for equivalent SOAs, incongruent speech would be less tolerant to desynchronization than congruent speech. This is indeed what was found. The congruent tokens -A_dV_d and A_tV_t - were more readily considered 'simultaneous' than the incongruent tokens (~95%). In the McGurk case, simultaneity judgments never exceeded 80%, and remained maximal within a plateau narrower than the congruent tokens. The AV incongruency of the speech tokens impinges, as predicted, on the subjective simultaneity judgment.

Simultaneity rating versus identification

Interestingly, the temporal window found in the subjective simultaneity task approximates that of the identification task. The correspondence between two different domains of perceptual analysis is intriguing from a classical viewpoint, where subjective time perception may be mediated by 'internal clocks' (e.g., Treisman, 1994). A possible alternative, in line with recent neurophysiological findings, is that temporal perceptual phenomena on the millisecond scale are implicit to neural computations (e.g., Van Rullen & Koch, 2003). As such, the perceptual resolution of short-range subjective time is closely associated with the time scales of underlying neural computations (for a thorough review on oscillatory brain mechanisms, see Başar, 1998).

Perceptual unit of speech

It is noteworthy that the 200ms temporal window of integration for AV speech shown in both experiments (Figure 4) corresponds to average syllable duration across languages (Arai & Greenberg, 1998). Insofar as the syllable is considered a basic and critical unit for the perceptual analysis of speech, temporal analysis on the syllabic scale is desirable and quite probably necessary (Greenberg, 1996). Indeed, compromising the syllable integrity is one the few variables that leads to drastic reduction in prosody comprehension (Lakshimarayan *et al*, 2003). Moreover, the dynamics of AV utterances in production are on the syllabic scale and an important

aspect of a possible supramodal speech code (Liberman and Whalen, 2000). The temporal evaluation mechanism of auditory and visual information streams appears essential to the processing of AV syllabic speech, as suggested by the overall decrease in simultaneity rate, the narrowing of the simultaneity plateau in incongruent AV syllables together with the width of the fusion rate plateau. Importantly, our results suggest that subjective simultaneity and integration of AV speech events share basic processing steps on a ~200 ms time scale, a time constant approximating the non-modality specific theta range (4-7Hz) (Buzsáki, 2002; Yordanova et al, 2002).

Speech specificity

A crucial aspect of natural (and synchronized) AV speech is that preparatory movements of the facial articulators usually precede the onset of the acoustic signal – for as much as few hundreds of milliseconds. In natural conditions, visual information can thus be extracted earlier than the auditory information. The precedence of visual cues may not only facilitate the detection of the auditory signals (e.g., Kinchla *et al.*, 1966) but crucially for ecologically relevant events such as in AV speech, also predict the identity of the produced auditory utterance to follow. In addition to the Callan *et al.* (2001) study described earlier, recent EEG findings show that visual speech influences auditory-specific cortical potentials in two major ways: (i) by constraining the auditory processing on a ~200 ms time scale (theta range) and (ii) by speeding up the neural processing of auditory speech on the ~20ms scale (gamma range scale) (cf. Chapter 3).

Forward Model

A forward model of AV speech integration is being proposed, in which visual speech initiates an abstract representation predictive of the auditory inputs (figure 5). It is critical to distinguish feed-forward (or "bottom-up") and feed-back (or "top-down") flows of neural information from forward connectivity. Forward mechanisms essentially posit that internal representations of the world are built-in constraints of the neural system, modulated by incoming inputs and overall brain states (i.e. level of arousal, expectations or 'state-dependent' activation). In particular, a growing body of evidence suggests the existence of *internal articulatory-based cortical representations* in speech (for review, see Jones (in press)). How does this new framework account for the reported experimental data?

Visual precedence and abstract representation

First, in natural conditions, precedence of visual information is proposed to initiate the speech processing system. Thus, any auditory inputs presented within the initiated perceptual window of 200 ms are unable to be evaluated against the visually-induced prediction. The window size observed does not only converge with general electrophysiological data on perceptual unit formation (e.g., Näätänen, 1992; Yabe *et al.*, 1997, 2001a, 2001b; Winkler *et al.*1998; Loveless, 2001; Poeppel, 2001, 2003) but is also relevant to AV speech, which unfolds on the syllabic scale as constrained

by the movements of the articulators (as was pointed out in the introduction, also available through global kinematics).

Additionally, the informational content (or speech identity) of the inputs are crucial in distinguishing two types of possible AV interactions, namely, 'fusion' (as reported here) and 'combination' (audio [pa] dubbed onto visual [ka]). Whereas the former results in a single fused percept 'ta', the latter results in any combinations and ordering of [pa] and [ka] (e.g. 'pka', 'kappa', etc.) (McGurk and McDonald, 1976). The difference in perceptual output for mismatched AV speech presentation suggests that both the informational content and the dynamics of AV speech act as constraints on the integration process.

Perceptual information and asymmetry

First, there exists a crucial difference across modalities regarding the informational content of the signals (i.e. the propensity of the incoming signal to lead to a robust perceptual categorization). The information *content* provided by speechreading constrains the speech categorization level to *visemes* (i.e. a representation based on place-of-articulation, where for instance [ba] and [pa] will be easily confused because they belong to the same viseme class 'bilabial'). In contrast, information provided by the auditory signal can easily single out a phoneme. Thus, the sensory modality initiating the speech-processing pathway will most likely play a defining role in the integration of AV speech. In the context of the proposed forward model of AV speech integration (Figure 5), leading visual speech information

(including the range of natural visual precedence) initiates the speech processing pathway. Gating studies of AV speech (e.g. Munhall *et al.*(1998)) have suggested that visual speech information accumulates with time of presentation –e.g. update and retention of visual information in memory store - while the auditory system uptakes information in a more categorical way. On this basis, one would predict that when the process is initiated by visual inputs, incoming auditory inputs can influence visually-induced representation within the limits of visual retention.

Second, the reported temporal integration window for AV speech inputs is characterized by a marked asymmetry. Leading auditory information decreases integration, while leading visual information tends to enhance it. This trend has previously been reported in connected speech (Grant and Greenberg, 2001) and is typically accounted for by an inherent adaptation of the central nervous system to differences in the speed of light and sound (Massaro, 1996). Empirical data suggest that a compensatory mechanism may act upon the perceived synchrony of auditory-visual events up to distances of 20 meters (Engel, 1971). However, in laboratory experiments, the physical distance between video and audio sources, and the subject, are small enough (~1m) to make any differences in the speed of light and sound negligible (~5 ms) (e.g., Sugita and Suzuki, 2003). From an AV speech standpoint however, one would expect visual leads to be more beneficial to the integration than auditory leads, as visual categorization remains incomplete. The observed asymmetry indicates that 30 to 100 ms of auditory processing are sufficient to 'suppress' the

influence of visual inputs, - a reminder of rapid temporal encoding the auditory system.

Auditory information reaches the primary auditory cortex as early as 10 ms (Celesia, 1976; Liégeois-Chauvel *et al.*,1991; Howard *et al*, 1996), and voicing information is already neurophysiologically realized at ~60ms (Steinschneider *et al.*, 1994, 1999), i.e. through 2 or 3 cycles of the gamma-band. Visual information can be recorded as early as 30 ms in the primary (V1) and motion (MT/V5) visual cortex (Ffytche *et al*, 1995; Buchner *et al*, 1997). In light of the temporal segmenting of informational streams in the cortex, the complex interplay of first arrival time latencies in cortical areas is likely to bring but a partial explanation to the asymmetric profile we observed in AV speech integration. The dynamic complexity is increased by plastic properties of neural systems: for instance, in multisensory sites, the precedence of unisensory inputs not only inhibits sensory-specific cortices (Laurienti, 2002) but also regulate the integrative function of multisensory cells in various subcortical and cortical areas (e.g., Benevento, 1977). What neural latencies provide us with is a reliable indication of when a temporal integration window is potentially initiated, i.e. the onset of 'permissible' duration for AV interaction, but they do not account for the temporal resolution of perceptual processes.

Our results suggest that both amodal internal representations ('abstract mutual information') and stimulus correlations ('spatio-temporal mutual information') need to be considered for the neural basis of AV speech integration. Our model takes into account two types of information redundancy likely to constrain the integration of

multisensory information, namely, external constraints of stimuli properties such as spatio-temporal correlation of the auditory and visual signals and internal constraints defined by the degree of perceptual saliency of the stimuli. Visually-based predictions of the auditory inputs proposed in our model suggests that abstract and spatio-temporal mutual information is accessible cross-modally. The different temporal profile of congruent versus incongruent tokens complements previous neurophysiological evidence that a congruent A_dV_d or A_tV_t is not equivalent to an illusory A_bV_g or A_pV_k (e.g., Sams *et al*, 1991). Despite the percep tual equivalence in categorical labeling, the decrease of spatio-temporal AV correlation in signal dynamics for incongruent pairs appears to be detected by the neural system while at the same time, the abstract (articulatory based) information content remains ambiguous enough to permit fusion. Accordingly, very similar simultaneity and fusion profiles for the incongruent stimuli were obtained, while a larger permissible temporal window of integration was found for congruent AV stimuli, in agreement with the common temporal and speech encoding processing steps.

Global cortical dynamics observed in multisensory processing further suggest that long-range synchronization in the theta range can mediate the integration process. Sensory-specific (auditory and visual cortices) and multisensory sites (STS, SC) have consistently been involved in the multisensory processing of AV speech, along with right prefrontal cortices, also shown to be activated in silent lip-reading (e.g., Bernstein *et al*, 2000) and temporal p erception of non-speech events (e.g., Harrington *et al.*, 1998; Bushara *et al.*, 2001; Calvert *et al*, 2001). The involvement of prefrontal cortices further supports the hypothesis that the “amodal theta range” could carry

perceptual unit formation on the 200ms time-scale, while orchestrating computations from separate cortical areas.

TABLE 2.1

STIMULUS	TASK	A LEAD Left Boundary (ms)	A LEAD Right boundary (ms)	PLATEAU CENTER (ms)	WINDOW SIZE (ms)
A_pV_k	ID	-25	+136	+56	161
	S	-44	+117	+37	161
A_tV_t	S	-80	+125	+23	205
A_bV_g	ID	-34	+174	+70	208
	S	-37	+122	+43	159
A_dV_d	S	-74	+131	+29	205

Table 2.1: Temporal integration windows parameters across conditions and stimuli.

Measures extracted from ADS fits (r^2>0.9) and a 95 % confidence limit on fusion or simultaneity rate at synchrony condition (SOA = 0 ms). ID is the identification experiment (Experiment 1). S is the subjective simultaneity experiment (Experiment 2).

FIGURE 2.1

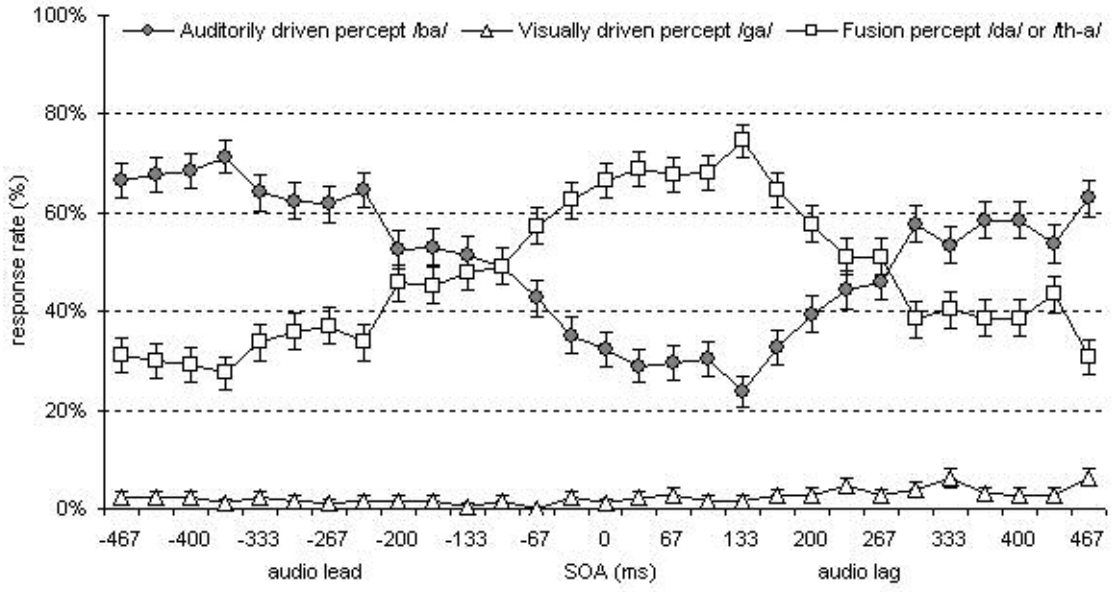


Figure 2. 1: Response rate as a function of SOA (ms) in the A_bV_g McGurk pair.

Mean responses (N=18) and standard errors. Auditorily driven responses (filled circles) are /ba/, visually driven responses (open triangles) are /ga/ and fusion responses (open squares) are /da/ or /th-a/.

FIGURE 2.2

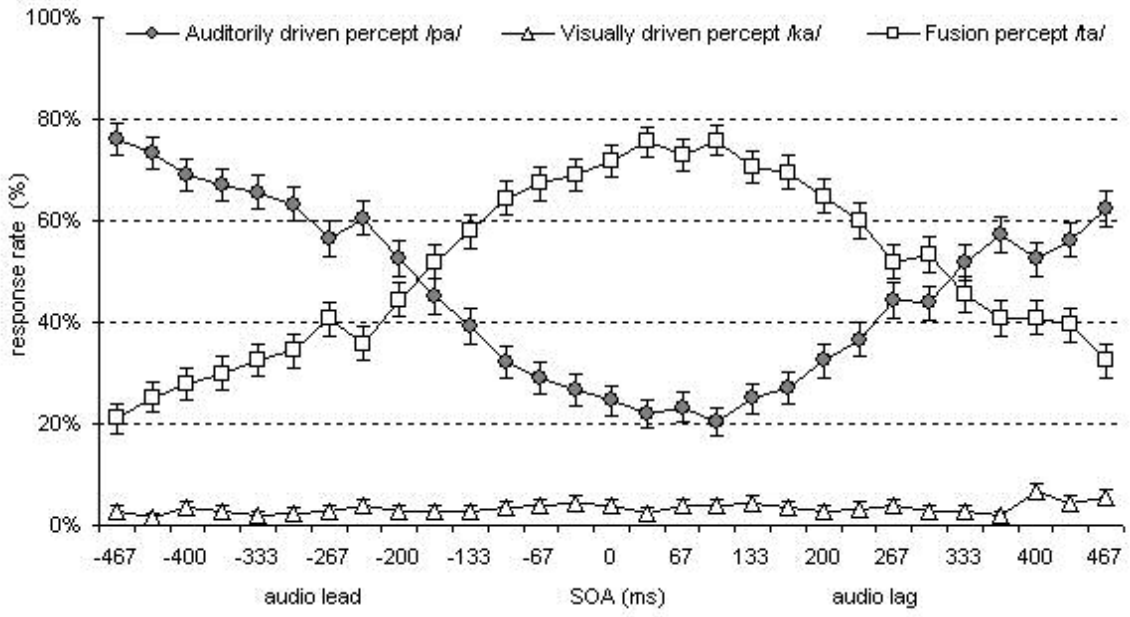


Figure 2. 2: Response rate as a function of SOA (ms) in the A_pV_k McGurk pair.

Mean responses (N=21) and standard errors. Auditorily driven responses (filled circles) are /pa/, visually driven responses (open triangles) are /ka/, and fusion responses (open squares) are /ta/. The sum of unimodal responses /ta/ to auditory alone /pa/ or visual alone /ka/ equals 53%. Fusion rates lower than 53% cannot be accounted for by unimodal errors. Fusion rates exceeding 53% constitute the true bimodal responses and can be observed from -167ms of audio lead to 267ms of audio lag.

FIGURE 2.3

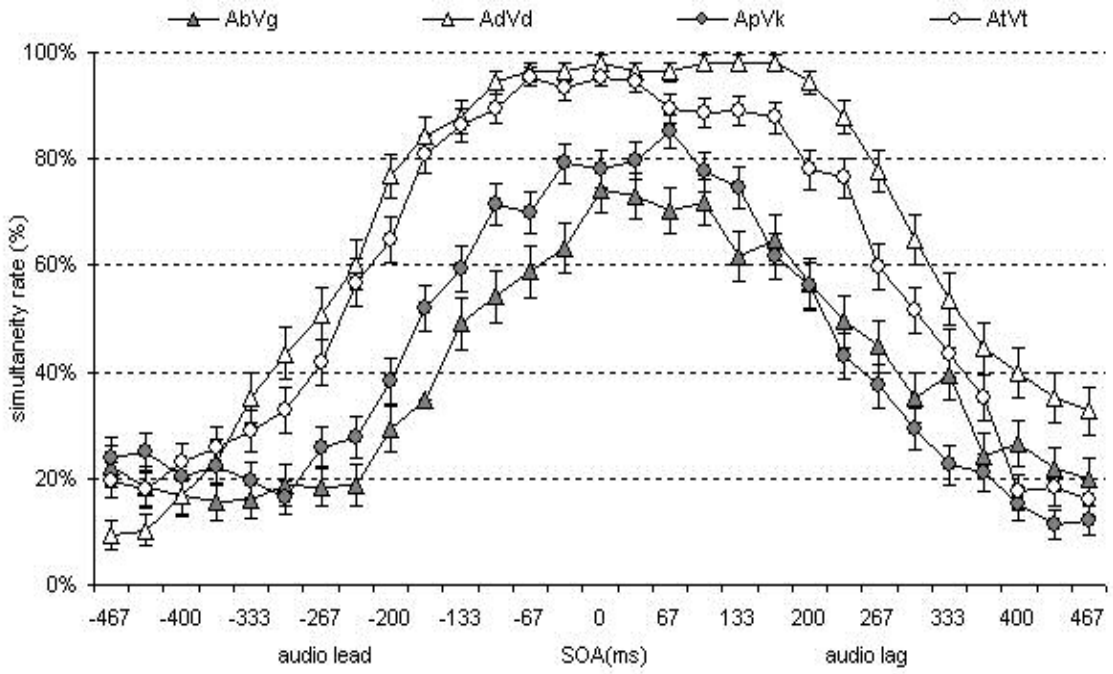


Figure 2. 3 Simultaneity judgment task.

Simultaneity judgment as a function of SOA (ms) in incongruent and congruent conditions (A_pV_k and A_tV_t N=21; A_bV_g and A_dV_d N=18). The congruent conditions (open symbols) are associated with broader and higher simultaneity judgment profile than the incongruent conditions (filled symbols). See Table 1and Figure 4 for further analysis of integration constants across conditions.

FIGURE 2.4

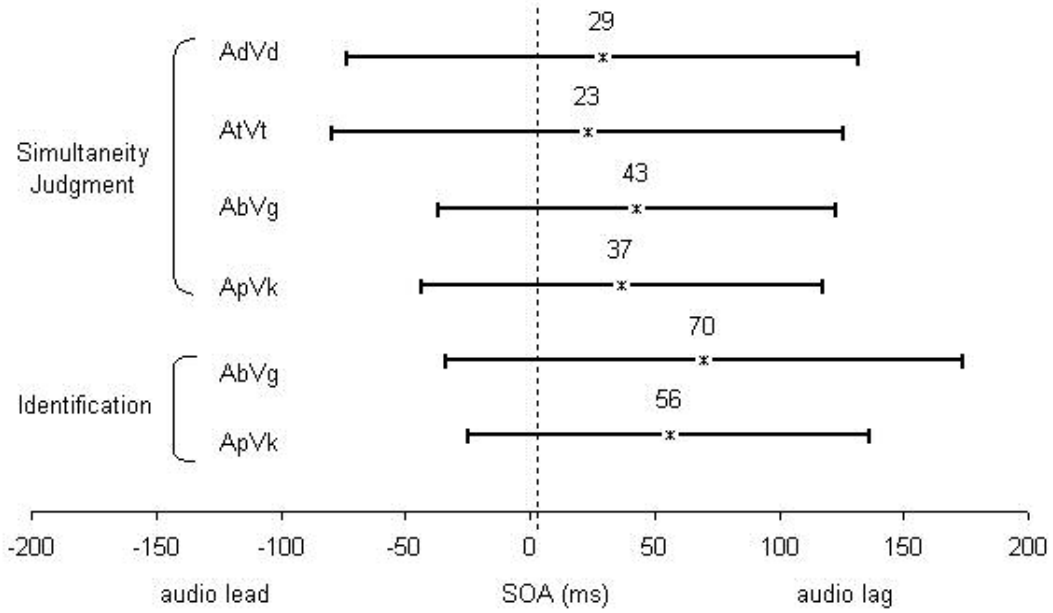


Figure 2. 4 Temporal integration windows across conditions and stimuli.

Temporal integration windows obtained across conditions show similar characteristics in width (~200ms) and in the existence of a displacement towards auditory lag. The plateau observed in the simultaneity judgment tasks for congruent tokens (A_dV_d and A_tV_t) is larger than for incongruent tokens (A_bV_g and A_pV_k). The cross marks the center of the plateau defined by the ADS fitting.

FIGURE 2.5

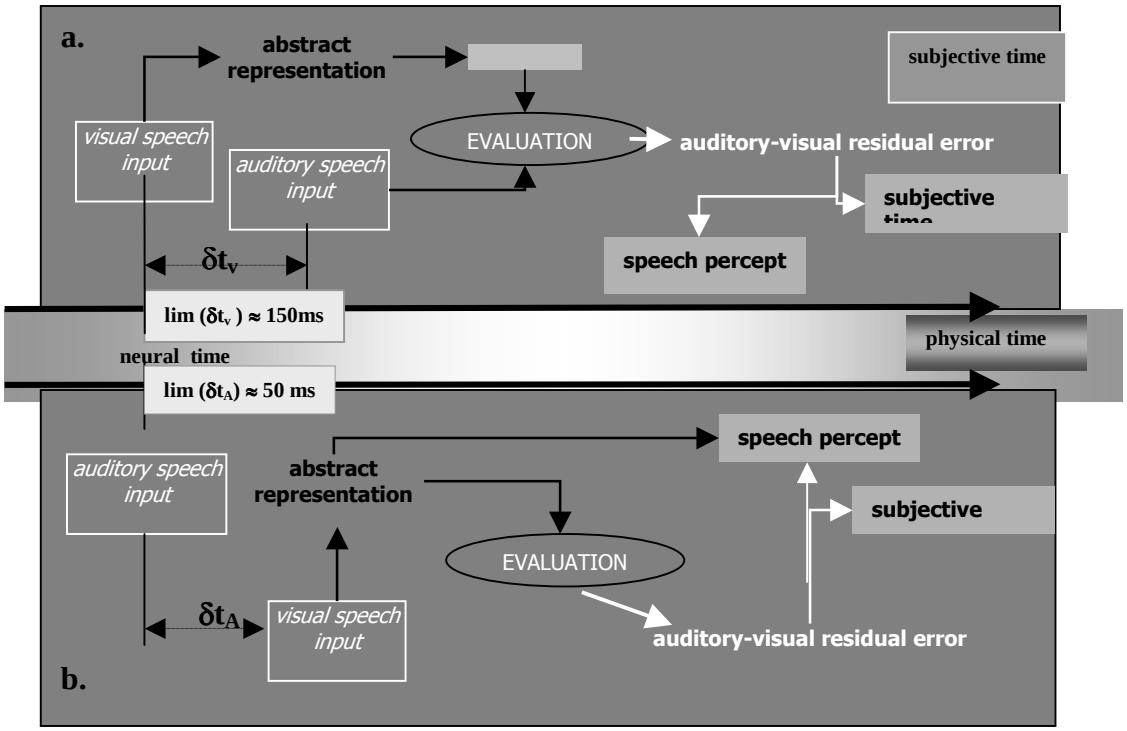


Figure 2. 5 Forward model of auditory-visual speech integration in time.

a. In natural *and* audio lag conditions, visual inputs initiate a speech representation based on place-of articulation. This abstract representation acquires a predictive value for expected auditory inputs. Note however, that visual categorization is limited to visemic (place-of-articulation) representation, which may undergo continuous update of sensory memory storage until audio inputs.

b. In artificial audio lead conditions, audio input allows complete categorization unless visual inputs interfere within ~50ms of audio processing. In the latter case, visual information is evaluated against audio input prior to perceptual completion.

Chapter 3: Visual speech speeds up the neural processing of auditory speech

"[…] information and knowledge are important to the brain: it must recognize the structure and regularity both to distinguish what is new information and to make useful interpretations and predictions about the world."

Horace Barlow

It would be as useless to perceive how things 'actually look' as it would be to watch the random dots on untuned television screens.

Marvin Minsky

Synchronous presentation of stimuli to the auditory and visual systems can modify the formation of a percept in either modality. For example, perception of auditory speech is improved when the speaker's facial articulatory movements are visible. Neural convergence onto multisensory sites exhibiting supra-additivity has been proposed as the principal mechanism for integration. Recent findings have suggested, however, that sensory-specific cortices are responsive to inputs presented via a different modality. Consequently, when and where audio-visual representations emerge remains unsettled. n combined psychophysical and electroencephalography (EEG) experiments we show that visual speech *speeds up* the processing of auditory signals early (within 100ms of signal-onset) and is reflected as temporal facilitation and response reduction. Crucially, the latency facilitation is systematically dependent on the degree to which the visual signal predicts possible auditory targets The observed AV interaction (latency facilitation and amplitude reduction) challenges the supra-additivity model for AV speech. The data support the view that there exist abstract internal representations that constrain the analysis of subsequent speech inputs. To our knowledge, this is the first evidence for the existence of a 'forward model' (or 'analysis-by synthesis') in auditory-visual speech perception.

3.1 Introduction

Studies of auditory-visual (AV) speech highlight critical issues in multisensory perception, including the key question of how the brain combines different incoming signals from segregated sensory processing streams into a single

perceptual representation. In AV speech research, an example of a phenomenon challenging neuroscience-based accounts as well as speech theories is the classic McGurk effect (McGurk & McDonald, 1976) in which an audio [pa] dubbed onto a facial display articulating [ka] elicits the 'fused' percept [ta]. A major question raised by the McGurk illusion is *when* in the processing stream (i.e. at which *representational stage)* sensory-specific information fuses to yield unified percepts.

The standard explanation for multisensory integration has been the existence of convergent neural pathways onto multisensory neurons (Stein & Meredith, 1993) that are argued to provide the substrate for 'multisensory binding (Meredith, 2002). A typical signature of multisensory neurons is the enhanced response (or 'supra-additivity') to the presentation of co-occurring events. Consistent with this concept, fMRI studies of AV speech have shown that auditory and polysensory cortices, specifically the Superior Temporal Sulcus (STS) and the Superior Temporal Gyrus (STG), in fact show an enhanced activation when compared to unimodal (auditory or visual) speech (Calvert *et al.*, 1999, 2000). The involvement of polysensory cortices has suggested a possible computational route for AV speech processing whereby signals integrated in multisensory cortical sites feed back onto primary sensory fields (Calvert, 2000). This feedback hypothesis (Calvert, 2000) predicts the enhanced activation of auditory cortices, the assumption being that the connectivity is excitatory and driven by the supra-additive output of STS.

While this explanation has appealing properties, there are complicating factors. For example, neurophysiological recordings in non-human primates show that classic multisensory integration sites such as STS demonstrate little specificity

for stimulus attributes (Bruce *et al*, 1981; Watanabe & Iwai, 1991). Second, in human studies, response enhancements have been observed in different perceptual contexts, and also shown in primary sensory cortices when stimuli are presented to a different sensory modality (e.g. Calvert *et al*, 1997, 1999). Third, the activation of multisensory cortices is observed for both biologically relevant (e.g. face matched with speech sound) and arbitrary multisensory pairings (e.g. tone with flashed light) and does not appear specific to AV speech processing. With regard to the issue of representational stage of integration, it therefore remains difficult to establish what the nature of information fed back (i.e. STS output) onto auditory cortices may be.

Recent findings also present new empirical challenges for the standard convergence model. Rather than supra-additivity, *suppression* (or 'deactivation') of sensory-specific cortices has been reported (Raij *et al.*, 2000; Laurienti *et al.*, 2002) in conjunction with an enhanced activation of multisensory cortical sites and, more recently, congruent AV speech has also been shown to elicit sub-additive interactions in polysensory regions (Laurienti *et al*, 2002).

Indeed, a growing body of anatomical evidence shows that primary sensory areas are directly interconnected (Falchier *et al*, 2002; Rockland & Ojima, 2003), suggesting that sensory streams can interact early on and that multisensory neural convergence is only one possible route for intersensory interactions. For instance, activation of auditory association areas in primates has been observed under somatosensory stimulation (Fu *et al*, 2003), and eye position modulates the respon se properties of neurons primary auditory cortex (Werner-Reiss *et al*, 2003). Additionally, intracranial recordings in primates have shown the existence of direct

somatosensory inputs in the auditory cortices (Schroeder *et al*, 2001). One proposed functional implication of intersensory cortico-cortical connectivity is to mediate cross-modal plasticity when one sensory system is compromised (Bavelier & Neville, 2002), but the functional benefit for non-impaired systems remains unknown.

In the context of AV speech perception, the anatomic intersensory connectivity data predict early interactions among processing streams. Although most recent studies are based on hemodynamic experiments that cannot speak directly to the timing issues (Logothetis *et al*, 2001; Attwell & Iadecola, 2002), there exist electrophysiological studies that have the appropriate temporal resolution to deal with timing. These EEG and MEG studies (Sams *et al*, 1991; Colin *et al*, 2002; Möttönen *et al.*, 2002) have typically used an oddball or mismatch negativity paradigm, and the earliest AV interactions have been reported for the 150-250ms latency mismatch response. We depart from the (passive and preattentive) mismatch negativity design and focus on the early ERPs elicited by participants explicitly discriminating AV speech.

Particular properties of AV speech play a crucial role for our design. Natural AV speech represents ecologically valid stimuli for humans (DeGelder & Betelson, 2003), and one would predict an involvement of functionally specialized neural computations capable of handling the spectro-temporal complexity of AV speech inputs - say as compared to a tone-flash pairing, for which no natural unified discrete representation can be assumed. For instance, natural AV speech is characterized by particular dynamics such as (i) the temporal precedence of visual speech (the movement of the facial articulators typically precedes the onset of the acoustic signal

by tens to a few hundred milliseconds) which provides a specific context for AV speech integration and (ii) a tolerance to desynchronization of the acoustic and visual signals of about 250ms (Munhall *et al*, 1996) - a time constant characteristic of syllables across languages (Arai & Greenberg, 1998) and which relates closely to a proposed temporal constant of neural integration underlying perceptual unit formation (Näätänen, 1992; Poeppel, 2003).

Furthermore, in the speech domain, abstract representations have been postulated to derive from the (intended) motor commands (articulatory gestures) for articulatory movements involved in speech production (Stevens & Halle, 1967; Liberman & Mattingly, 1985). Visual speech provides direct - but impoverished - evidence for particular articulatory targets; in contrast, the auditory utterance alone usually permits complete perceptual categorization (say on the phone). For instance, while an audio-alone /pa/ leads to a clear percept /pa/, its visual-alone counterpart (i.e. seeing a mouth articulating [pa]) is limited to the recognition of a visual place-of-articulation class, or the 'viseme' category *bilabials*, which comprises the possible articulations [pV], [bV], and [mV] (Massaro, 1998) (where the [V] stands for any possible vowel).

Within this general framework, we investigated the cortical dynamics of perceptual fusion for ecologically natural multisensory events and focused on the timing of AV speech integration, which remains an open question. We conducted three behavioral and EEG experiments to characterize the influence of visual speech on the most robust auditory event-related potentials (ERP) N1/P2 (negativity peaking at ~100ms post-auditory stimulus onset followed by a positivity peaking at ~200ms)

and focused our analysis on systematic variations of the auditory ERP as a function of visual speech information. We used both congruent stimuli (AV syllables [ka], [pa] and [ta]) and incongruent McGurk stimuli (McGurk & McDonals, 1976), in which the dubbing of an audio [pa] onto a visual place-of-articulation [ka] elicits the illusory or fused percept [ta]. In all experiments, participants identified (3 alternative-forced choice paradigm) syllables in auditory (A), visual (V) and AV conditions during EEG recording. We show that the visual information systematically influences key timing properties of the auditory responses and argue for a 'forward model' for AV speech integration.

3.2 Materials and Methods

Participants

Twenty-six native speakers of American English (13 females, mean 21.5 years) were recruited from the University of Maryland population. No participant had diagnosed hearing problems, all had normal or corrected-to-normal vision and were right-handed. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli and Procedure

To preserve the natural relation between auditory and visual inputs, we used natural speech consisting of a woman's face articulating the syllables [ka], [pa] and [ta]. The average duration of the AV stimuli was 2590 ms, including video fade-in (8 frames), neutral still face (10 frames), place-of-articulation (variable) and fade-out (5 frames). Interstimulus intervals (ITIs) were pseudo-randomly selected among 5 values (500 ms, 750 ms, 1000 ms, 1250 ms and 1500 ms). Stimuli were pseudo-randomly intermixed and presented per recording period as follows: in Experiment 4, sixteen participants were presented with two blocks of 200 AV stimuli each (congruent AV [ka], [pa], [ta] and incongruent audio [pa] dubbed onto visual [ka] were presented 50 times per block) and 2.5 blocks of 240 unimodal (auditory and visual alone [ka], [pa], and [ta] presented 40 times per block) stimuli (for a total of 1000 trials, 100 presentations per stimulus). Ten participants took part in Experiment 5, consisting of the stimuli used in Experiment 4 (A, V, AV) presented 100 times per stimulus (for a total of 1000 trials). Ten participants, who took part in Experiment 4, also participated in Experiment 6, which consisted of 200 incongruent AV stimuli (McGurk fusion and combination pairs (audio [ka] dubbed onto visual [pa]), only fusion are reported here).

Participants were placed about 1m from the visual display, with the movie subtending a visual angle of 8.5º in the vertical plane and 10.5º in the horizontal plane. Videos were displayed centered on a 17" G4 monitor on a black background. Sounds were presented through Etymotic ER3A earphones connected to the computer through a sound mixer table at a comfortable level of approximately 70 dB SPL. Lights were dimmed before recordings.

In all conditions, a single-trial 3 alternative-forced choice (3AFC) procedure was used. For all experiments, the three choices were [ka], [pa], or [ta]. In the AV conditions (Experiments 1 and 2), participants were asked to make a choice as to "what they hear while looking at the face". In the unimodal conditions (A, V), participants were asked to make a choice as to what they hear or see, for the A or V conditions respectively. In Experiment 6, participants were asked to report, "what they see and neglect what they hear". No feedback was provided.

Electroencephalographic recordings

EEG recordings were made using a Neuroscan system (Neurosoft Systems, Acquire 4.2b), using 32 Ag/AgCl sintered electrodes mounted on an elastic cap (Electrocap, 10-20 enhanced montage). Data were continuously acquired in AC mode at a sampling rate of 1kHz. References electrodes were left and right mastoids and grounded to AFz. A band-pass filter from 1Hz to 100Hz was applied online. Two electrodes monitored horizontal eye movements (HEOG) and two others recorded vertical eye movements (VEOG) for off-line artifact reduction and rejection. Impedances were kept below 5kOhm per channel.

Data analysis

After artifact rejection and ocular artifact reduction (linear detrending), epochs were baseline corrected on the basis of a pre-stimulus interval of 400 ms chosen prior

to either auditory (A condition) or visual onset (V alone and AV conditions). A threshold of ± 100μV was used to reject residual artifacts. Approximately 75-80% of the original recordings were preserved (about 20 trials per stimulus condition were rejected). Individual averages were made for each stimulus-response combination. Only correct responses were further analyzed (false alarm, correct rejection and error rates did not provide enough samples for comparison). For McGurk conditions, fusion responses [ta] were considered 'correct' in all experiments. A zero-phase-shift double-pass Butterworth band-pass filter (1-55Hz, 48dB) was applied for event-related potentials peak analysis and reported traces.

An automatic peak detection procedure was used for common ERP parameterization (peak latency and peak amplitude) and corrected manually for each electrode and each participant when necessary. A bootstrapping method (Efron, 1979) was used to resample the data 300 times for each individual, each condition and each electrode (6 electrodes were used: FC3, FC4, FCz, CPz, P7, and P8). Unprocessed and bootstrapped ERP values were submitted to repeated measures analyses of variance with factors of modality (2 levels, since in V condition, no auditory-event related potential was observed), stimuli (6 levels; audio and congruent audio-visual [ka], [pa], and [ta]), event-related potentials (3 levels; P1, the small positivity at ~50ms post-auditory onset was included in the analysis), and electrodes (6 levels). Electrodes comparisons were submitted to Greenhouse-Geisser corrections when sphericity could not be assumed. Unpaired t-tests were used to test predicted contrasts. Reported P values in text are for unprocessed ERP values (bootstrapped data lead to similar significant effects but are not reported in this manuscript).

3.3 Results

In the first experiment, unimodal (A, V) and bimodal (AV) stimuli were tested in separate blocks. Figure 1 shows the grand averaged responses obtained for each syllable (place-of-articulation condition) tested in A, V and AV conditions. The presence of visual speech inputs (AV condition) significantly reduced the amplitude of the N1/ P2 auditory ERP compared to auditory alone conditions (A), in agreement with the deactivation hypothesis (Laurienti *et al*, 2002; Bushara *et al*, 2003) and contrary to the expectation of supra-additivity (Calvert *et al.* , 1999).

Analyses of variance showed a significant interaction of modality (A, AV) on the amplitude of the N1/P2 response component ($F(1.304, 19.553) = 49.53$, $p < 0.0001$). Additionally, we observed a significant shortening of response peak in AV syllables compared to A alone conditions. (Note that no N1/P2 – i.e. typical auditory ERP - was elicited in visual alone conditions; because we focus here on auditory ERP, V alone conditions will not be further reported.) Repeated measures ANOVA testing modality (A, AV) and stimulus identity (/ka/, /pa/, or /ta/) showed a significant interaction ($F(1.834, 27.508) = 14.996$, $p< 0.0001$). One can observe this effect on the N1, and it is even more articulated for the P2. These results argue (i) for an early AV interaction that is evident as early as the N1 and (ii) for a manifestation of AV interaction not as response supra-additivity but rather as deactivation and latency facilitation.

Because unimodal and bimodal conditions were run in separate blocks in the first experiment, participants knew at the start of a visual trial whether to expect an auditory stimulus or not. To control for participants' expectancy, the same

experimental items were pseudo-randomly presented in a second experiment. A similar amplitude reduction was observed affecting the auditory N1/P2 complex in all AV conditions ($F(1.507, 13.567) = 17.476$, $p < 0.0001$). The temporal facilitation effect was also observed again. Repeated measures ANOVA showed a significant effect of presentation modality (A, AV) ($F(1, 9) = 21.782$, $p < 0.001$) and a marginally significant interaction of presentation modality and stimulus identity ($F(1.938, 17.443) = 3.246$, $p < 0.06$).

The overall effects of visual speech on auditory ERP amplitude and latency were similar for Experiments 1 and 2 (blocked versus randomized designs). Crucially, the *temporal facilitation* effects, unlike the *amplitude reduction* effects, varied with stimulus identity (i.e. [pa], [ta], or [ka]). Whereas visual modulation of auditory ERP *amplitude* did not significantly vary with stimulus identity (Experiment 4, $F(1.884, 28.265) = 1.22$, $p<0.308$; Experiment 5, $F(1.565, 14.088) = 0.033$, $p<0.94$), the temporal facilitation *was* a function of stimulus identity (Experiment 4, $F(1.908, 28.62) = 13.588$, $p<0.0001$; Experiment 5, $F(1.808, 16.269) = 20.594$, $p<0.0001$). As mentioned, articulator movement in natural speech precedes the auditory signal and may therefore predict aspects of the auditory signal insofar as the speech recognition system incorporates a forward (or 'analysis-by synthesis') model (Stevens & Halle, 1967; Wolpert *et al*., 1995). Thus, if a visual input is ambiguous (e.g. V [ka] was correctly identified only ~65% of the time), the predictability of the auditory signal should be *lower* than if the visual stimulus is salient and predictable (e.g. V [pa] ~100% correct identification), and facilitation effects should vary accordingly: the more salient and predictable the visual input, the more the auditory processing is

facilitated (or, the more visual and auditory information are redundant, the more facilitated the auditory processing).

Consistent with this hypothesis, we observed articulator-specific latency facilitation, and Figure 2 shows the grand averaged (across Experiments 1 and 2) visual modulatory effects on N1/P2 latency (2a) and N1/P2 amplitude (2b) as a function of correct identification (C.I.) in the visual alone condition. For example, [ka] was identified correctly only ~65% and associated with a 5 to 10 ms latency facilitation on the N1 and the P2; the syllable [pa], in contrast, was identified correctly more than 95% and was associated with a latency facilitation of ~10ms at the N1 and ~25ms at the P2. These results suggest that the degree to which visual speech predicts possible auditory signals affects the amount of temporal facilitation in the N1/P2 transition (Figure 2a) but does not affect its amplitude differentially (Figure 2b).

For the McGurk fusion, an audio [pa] was dubbed onto a visual [ka]. If the rules of integration in AV speech are based upon the saliency and redundancy of inputs across sensory channels, one predicts that in McGurk fusion, the ambiguity of the visual speech input [ka] will not facilitate the latency of the auditory ERP (i.e. the amount of latency facilitation observed in McGurk fusion should be less than for a natural AV [pa], where redundant information is being provided). A similar amplitude reduction should however be observed that is independent from the informational content of visual speech input (as shown in Experiments 1 and 2). Figure 3 summarizes the latency (3a) and amplitude effects (3b) observed in Experiments 1 and 2 (filled bars) for congruent AV /pa/ and the McGurk ‘fusion’

token. As predicted, little-to-no temporal facilitation was observed for the McGurk condition (Figure 3a, congruent AV /pa/ versus 'fusion'), while the amplitude decrease of the auditory ERP is comparable to that of a congruent AV /pa/ (Figure 3b, congruent AV /pa/ versus 'fusion').

One hypothesis about the observation that both congruent and incongruent (fusion) AV stimuli produced equivalent reduction in amplitude, independent of stimulus identity or perceptual redundancy, is that the visual modality divides the attention participants focus on the auditory modality. This possibility lead to a third experiment, in which we tested the effects of attending to the visual modality when auditory and visual inputs were *incongruent* (so as to know which modality the reported percept is associated with). If attending to the visual modality (i.e. non-attended auditory modality) underlies the observed amplitude reduction, one would predict that explicitly directing the participants' attention on the visual inputs would further attenuate the auditory ERP (Woods *et al*, 1992). Participants were presented with the same McGurk stimuli and answered according to what they *saw* instead of what they *heard*. Figure 3b (right bar, 'fusion in visual attention') shows that there was little to no amplitude difference between "visually attended" incongruent stimuli and either congruent (AV [pa]) or incongruent AV stimuli tested in Experiments 1 and 2, suggesting that in AV speech, the deactivation of the auditory cortex is automatic and independent of attended modality. Figure 3a reports the temporal facilitation observed for McGurk fusion. Surprisingly, under visual attention, the incongruent AV stimulus showed similar temporal facilitation as observed earlier for congruent AV [pa] (Figure 3a), i.e. the auditory ERPs were temporally facilitated

despite the ambiguity in the visual domain. This result will be discussed below within a forward model of AV speech integration.

3.4 Discussion

Our results show two major electrophysiological features of AV speech integration. First, the degree of perceptual ambiguity in the visual domain predicts the speed of neural processing in the auditory domain, consistent with a forward model view of speech processing. Second, contrary to the predictions of enhanced activation, AV speech results in reduced auditory evoked related potentials when compared to auditory speech alone. This amplitude reduction is independent of AV speech congruency, participant's expectancy (Experiment 4 and 2), and attended modality (Experiment 6). Our findings suggest (i) that AV *speech* processing follows specific rules of integration not solely accounted for by general principles of multisensory integration and (ii) that at least two distinct time scales underlie the integration process.

EEG studies of multisensory integration for *artificial* AV pairings have thus far supported the response enhancement results observed with fMRI, showing supra-additivity of unisensory responses to the presentation of co-occurent AV stimuli (Giard & Peronnet, 1999). In particular, the amplitude of the auditory N1/P2 complex was increased in AV conditions (tones paired with circles) and preceded by an early-enhanced component (40 to 90 ms post-stimulation). We believe the differences in results reflect the effect of ecologically valid stimulation in the study of multisensory

perception (DeGelder & Bertelson, 2003). No clear perceptual categorization of 'tones-circles' can be assumed, because no single perceptual entity 'tone-circle' is available in the natural environment.

Temporal Facilitation

Building on the notion of the ecological validity of the signal, we interpret our EEG results as supporting the notion of predictive coding in the context of a forward (Wolpert *et al*, 1995) or 'analysis-by synthesis' model (Stevens & Halle, 1967) of AV speech integration. Figure 4 illustrates the proposed model, in which the perceptual outcomes depend on (1) the saliency of visual inputs and (2) the redundancy of visual and auditory inputs. The notion of predictive coding, first used in motor systems (Wolpert *et al*, 1995) has more recently been tested and extended to sensory systems (Rao & Ballard, 1999). A key assumption of forward models is that an internal representation of the world is intrinsically present, built on prior experiences ('nurture') and inherent cortical properties ('nature') (Barlow, 1994). On the forward view, sensory inputs are not solely processed in a feed-forward fashion but are constrained *early on* by the internal predictions of the system. A major consequence is that early sensory processing can specialize in computing the *residual error* between the sensory input and the internal prediction, which characterizes the forward nature of the system.

We propose that the natural dynamics of AV speech (e.g. precedence of visual speech inputs) allow the speech processing system to build an online-prediction of the

auditory signal. The temporal facilitation of auditory ERPs suggests that interactions of AV speech inputs are constrained early on by preceding visual information. In particular, AV speech syllables used in this study naturally provide visible co-articulatory movements few hundreds of milliseconds prior to the acoustic signal. The amount and nature of visual information extracted during this period is proposed to initiate the speech processing system, in which the formation of an abstract representation is continuously updated through visual inputs, up to the point of explicitly auditory input. The set of possible visemic (i.e. articulatory-based) representations initiated in the visual domain is considered to provide the 'context' in which auditory inputs are being processed. The abstract representations triggered by the visual signals, in turn, provide internal predictions whose strength correlates with the saliency of visual inputs - i.e. the ease of perceptual categorization in visual alone condition - and against which the auditory inputs are being evaluated.

On this view, the temporal facilitation observed in the auditory N1/P2 complex under AV speech conditions reflects the residual errors of the auditory inputs matched against the internal predictor. Consistent with the neural generators of the N1/P2 (Näätänen & Picton, 1987) and prior reports of multisensory interactions (Calvert, 2000; Wright *et al.*, 2003), a possible locus for such prediction-to-auditory-input evaluation is the Superior Temporal Gyrus (STG).

The surprising neutralization of the temporal facilitation observed in the McGurk fusion condition points to a possible role of attention in the proposed model. In particular, perhaps the weight of the visually initiated predictor can be regulated by attention in the evaluation process. It has previously been suggested that in

conflicting multisensory presentation (such as in the McGurk fusion case), directing one's attention to a particular modality tends to increase the bias of the attended modality over the unattended modality (Welch & Warren, 1980). In our Experiment 6, this attentional biasing effect is observed as temporal facilitation regardless of the degree of saliency, i.e. the visual-based prediction is here proposed to dominate the auditory input in the evaluation process.

Reduction or supra-additivity

The amplitude reduction of the auditory N1/P2 complex challenges classic supra-additive effects reported for multisensory events in the brain imaging literature. While *non-speech* stimuli have been found to enhance the amplitude of classic auditory ERPs, such supra-additivity, characteristic of subcortical and cortical multisensory neurons (Stein & Meredith, 1993), is not *a priori* suitable for unisensory cortices. fMRI recordings have, of course, shown activation of auditory cortices to AV speech presentation (Calvert, 1997). However, because fMRI results must be interpreted over much longer time scales (hundreds to thousands of milliseconds), potential supra-additive effects may occur in the summation of the signal over time. Nevertheless, specific image acquisition and data analysis strategies can compensate for these issues (Logothetis *et al.*, 2001; Attwell & Iadecola, 2002), and recent fMRI studies (Wright *et al.*, 2003), too, have reported *decreased* activation of sensory cortices when stimulation targeted a different modality (e.g. decreased auditory cortex activation to the presentation of visual stimuli). This finding has been proposed to

result from a deactivation mechanism in which stimulation of one modality inhibits the non-stimulated modality (laurienti *et al.,* 2002).

Consistent with this last proposal, deactivation mechanisms may provide a way to minimize the processing of redundant information cross-modally. In our model, the internal prediction deriving from visual inputs essentially reduces the informational content to place-of-articulation ('viseme'). In the incoming acoustic information, information pertaining to place of articulation is confined, roughly, to the 2nd and 3rd formants. Following the assumption that the system acts upon incoming inputs to reduce signal redundancy in order to extract 'novel' information, the deactivation of auditory cortices by preceding visual inputs could target the auditory neural population extracting information only in the relevant frequency range.

Temporal integration and early interaction

The effect of visual speech inputs on early auditory evoked responses raises the issue of the temporal locus of AV speech integration. Previous electrophysiological studies using the mismatch negativity paradigm in the context of AV speech reported that the auditory mismatch component (MMN), typically peaking between 150ms and 200ms, could be elicited when a visual signal incongruent with the auditory speech syllables was presented, suggesting that visual speech accesses auditory sensory memory (Sams *et al*, 1991; Colin *et al*, 2002; Möttönen *et al*, 2002)

The type and timing of first cross-modal interaction, however, has remained, speculative.

We observe that the processing of auditory speech depends on visual inputs as early as 100ms (cf. N1 effects both in amplitude and in time), suggesting that the first systematic AV speech interaction occurs prior to N1 elicitation. Additionally, amplitude and temporal effects evolve on two different time scales: while the latency facilitation occurs in the 25ms range and depends upon visual saliency (and thereby informational content of the visual input), the amplitude reduction is independent of visual speech information and spreads over ~200ms. These time constants have been reported throughout the auditory neuroscience literature (Näätänen, 1992). In particular, 200-300ms has been hypothesized to underlie perceptual unit formation (Poeppel, 2003) while 20-40ms is related to auditory feature extraction. Consistent with this perspective, our results suggest that at least two computational stages of multisensory interactions are in effect in AV speech integration: first, as reflected in the auditory ERP latency facilitation, a featural stage in which visual informational content enables the prediction of the auditory input, and second, as reflected in the amplitude decrease, a perceptual unit stage in which the system is engaged in a bimodal processing mode, independently of the featural content and attended modality. The range of temporal phenomena observed electrophysiologically (~20ms of temporal facilitation and ~200ms of amplitude reduction) may relate to speech features associated with (1) phonetic-based analysis and (2) syllabicity**,** respectively. The time constants found with AV speech coincide with recent hypotheses (Poeppel, 2003) that speech is simultaneously processed with both shorter (25-50ms) and longer

(150-250ms) temporal integration windows in cortex. Our results show that visual speech modulates early stages of auditory processing (~50-100ms), probably prior to phonetic perception. This observation is in line with early integration. We do not find support for supra-additivity at this stage of auditory processing. We observe the early interaction of auditory and visual signals as a latency facilitation of the N1/P2 evoked responses, conditioned by the saliency of visual inputs. This suggests that visual inputs carry a specific predictive value for the auditory utterance. These data are most naturally interpreted in the context of speech perception theories that incorporate a forward model.

Acknowledgments

We thank Dr. Jonathan B. Fritz and Dr. Anita Bowles for their critical and helpful comments on earlier versions of the manuscript. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense. This work was supported by a grant from the National Institutes of Health to DP (NIH R01DC05660). During the preparation of this manuscript, DP was in residence as a Fellow at the Wissenschaftskolleg zu Berlin.

FIGURE 3.1

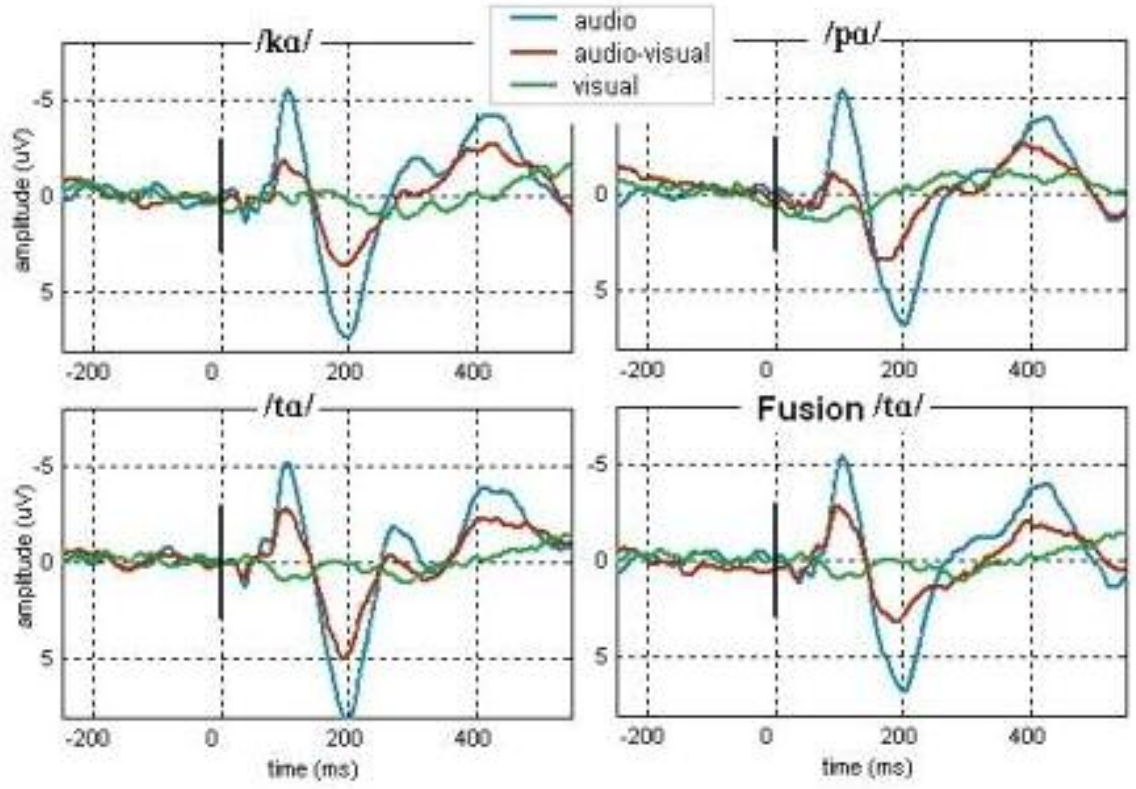


Figure 1.1 Average ERPs across conditions.

Grand averaged auditory, visual and auditory-visual speech event-related potentials at a centro-parietal recording site (CPz, data band-pass filtered 1 to 55Hz). The black vertical line indicates the onset of the auditory signal. AV speech (red trace) produced a faster but smaller auditory-event related potential compared to the auditory alone condition (blue trace). Visual speech (green trace) onset occurred ~400ms prior to auditory onset and did not elicit an auditory-event related potential but did produce typical visual ERPs at temporo-occipital electrode sites. The three distinct places of articulation tested as well as the McGurk case are tested separately.

FIGURE 3.2

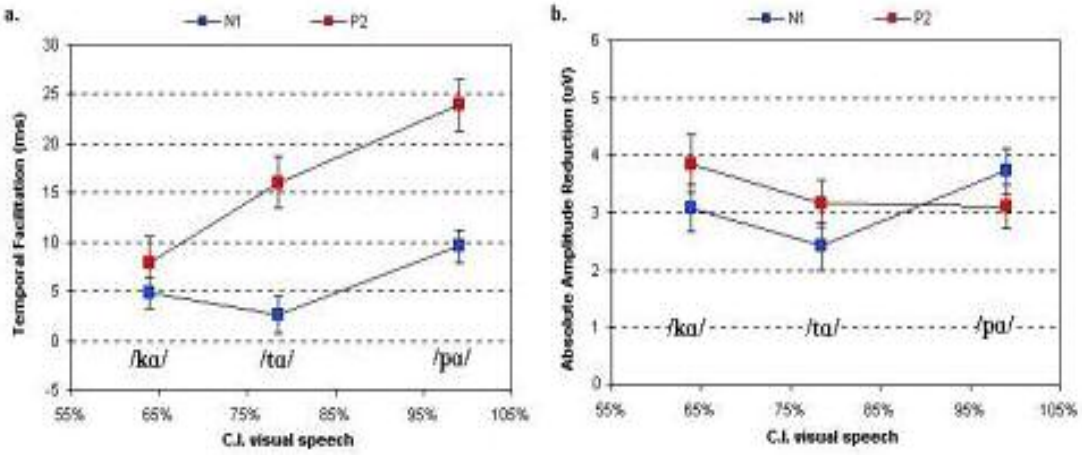


Figure 3.2 Latency facilitation and amplitude reduction.

Latency and amplitude difference of N1/P2 in auditory-visual syllables as a function of correct identification (C.I.) in the visual alone condition (Experiments 3 and 4, n=26). The latency (a) and amplitude (b) differences are the latency (or amplitude) values for the A condition minus the latency (or amplitude) for the AV condition for the N1 (blue) and P2 (red) ERPs. A positive value means AV is faster than A. The temporal facilitation of the N1 and P2 increased as the saliency (correct identification, C.I.) of visual inputs improved. The amplitude reduction in AV speech (b) remained constant across syllables and independent of visual saliency.

FIGURE 3.3

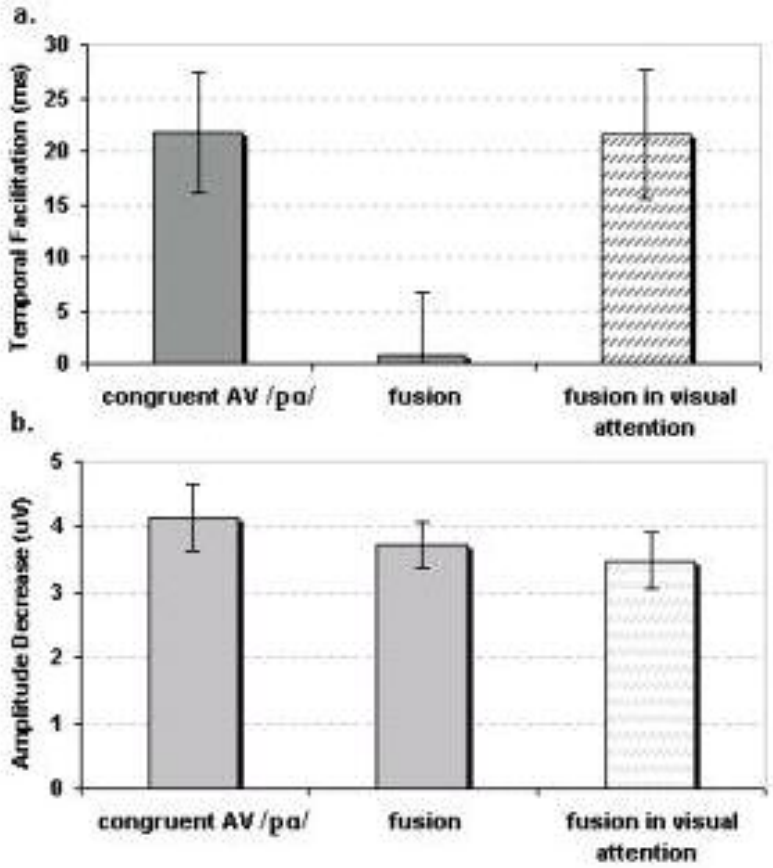


Figure 3.3 P2 latency facilitation and intersensory bias.

Compared to congruent AV /pa/ (3a, left), no latency facilitation was observed for fusion (3a, middle). However, when attention was directed to visual inputs in AV conditions, temporal facilitation was recovered in fusion (3a, right) suggesting that visual attention can enhance the biasing effect of the weak predictor. The amplitude decrease (3b) was consistent across all stimuli and independent of attended modality, pointing to the automaticity of AV speech integration.

FIGURE 3.4

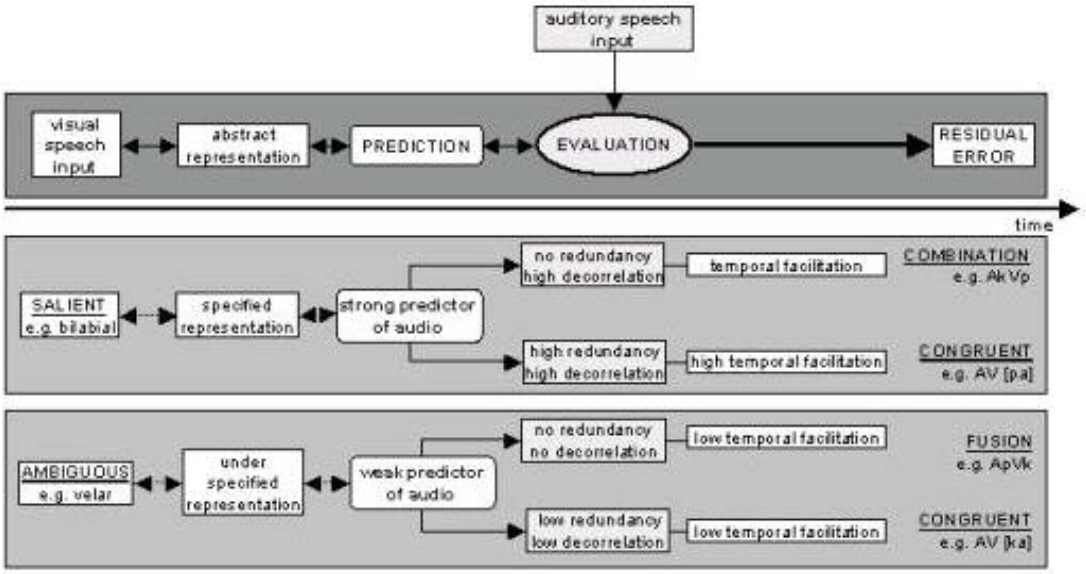


Figure 3.4: Forward model in auditory-visual speech integration.

Visual speech inputs (that typically *precede* the auditory signals) elicit an abstract speech representation. Its predictive value varies as a function of visual saliency and is updated as more visual information is made available. Incoming auditory speech inputs are then evaluated against the prediction. Redundancy between predictor and auditory inputs is decorrelated such that greater redundancy leads to greater decorrelation. Thus, the stronger the predictor, the faster the auditory speech processing. On this interpretation, the N1/P2 reflects the residual error of the evaluation process.

3.5 Controls on the origin of the amplitude reduction and latency facilitation

In this section, I provide further EEG data controlling for the modulatory effects reported in Experiments 3 and 4; various hypotheses regarding the origins of the amplitude reduction and the latency shifts of auditory-evoked potentials observed in AV speech conditions are tested.

Methods

The following control experiment was conducted with the same EEG apparatus, the same experimental settings and analyzed in the same manner as the stimuli tested in Experiment 4 and 5 (Chapter III).

Eleven naive participants took part in this experiment (5 females, mean 22.27 years). Participants were asked to identify the stimuli as being either a clear *[pa]*, *speech* (i.e. speech utterance different from [pa]) or *I don't know* (i.e. could not determine if it was speech or not). Participants were asked to respond to what they 'hear while looking at the computer screen' in auditory-visual conditions (AV), what they 'hear' or 'see' in audio alone (A) and visual alone (V) conditions, respectively.

Thirteen conditions were tested in this experiment. All stimuli were drawn from original AV speech token [pa], which provided the most salient effects in previous experiments (i.e. all timings are the same as original AV [pa] unless otherwise stated). Table 1 describes the characteristics of the tested stimuli.

Stimulus (abbreviation)	Description
Audio alone [pa] (**A_p**)	Identical to Experiment 4 and 5.
Visual alone [pa] (**V_p**)	Identical to Experiment 4 and 5.
Audio-visual [pa] (**A_pV_p**)	Identical to Experiment 4 and 5.
Audio [pa] with truncated visual [pa] (**trA_pV_p**)	Identical to Experiment 4 and 5 but visual fade-in onsets 2 frames (60ms) prior to audio onset. All visual motion information was eliminated prior to audio onset.
Audio [pa] dubbed onto truncated visual noise (**trA_pV_{noise}**)	Timing identical to (**trA_pV_p**). Face has been replaced by noise.
Reversed audio [pa] (**$_pA$**)	The original A_p was reversed in time.
Reversed audio [pa] dubbed onto visual [pa] (**$_pAV_p$**)	$_pA$ was dubbed onto the original V_p and started at the same time as the original audio [pa].
Audio [pa] dubbed onto backwards going visual [pa] (**$A_{pp}V$**)	A_p was dubbed onto the backward V_p . The timing respected the original A_pV_p (i.e. as much time with video alone preceded the onset of audio stimulus).
Audio [pa] dubbed onto a still visual face (**A_pV_{still}**)	A neutral face was chosen in the movie sequence (closed mouth) and replaced the entire length of the original movie.
Audio white noise (**A_{noise}**)	Same duration as the original $A_{p.}$
Visual noise (**A_{noise}**)	Same visual noise as in $trA_pV_{noise.}$
Audio-visual noise (**AV_{noise}**)	A_{noise} was dubbed at time of the original A_p but on $V_{noise.}$
Audio noise dubbed onto visual [pa] (**$A_{noise}V_p$**)	A_{noise} was dubbed at time of the original A_p on $V_{p.}$

Table 3.1 Control stimuli for Experiments 4 and 5

Stimuli were pseudo-randomized and each was presented 100 times in 5 blocks of 180 trials each (for a total of 1300 trials). Inter-trials durations were pseudo-randomly chosen among 5 values (500 ms, 750 ms, 1000 ms, 1250 ms and 1500 ms).

All values for auditory-evoked potentials amplitude and latency are being reported from observation of a centro-parietal site of recording (CPz) in order to provide a better comparison with Experiments 4 and 5.

Replication of Experiments 4 and 5: amplitude decrease and latency shift in natural AV speech

Results for A_p, V_p and A_p V_p replicate the findings of Experiment 4 and 5 with another set of naïve participants. A comparable amplitude reduction and latency shift of the auditory N1/P2 complex in AV conditions was thus predicted. Figure 3.5 reports the Global Field Power (GFP) across all participants (N=11) over all electrodes obtained for each stimulus. The latency of the N1/P2 complex in AV (yellow) condition was shorter than in A condition (blue). An amplitude reduction of the N1/P2 complex in AV condition as compared to A condition was also observed.

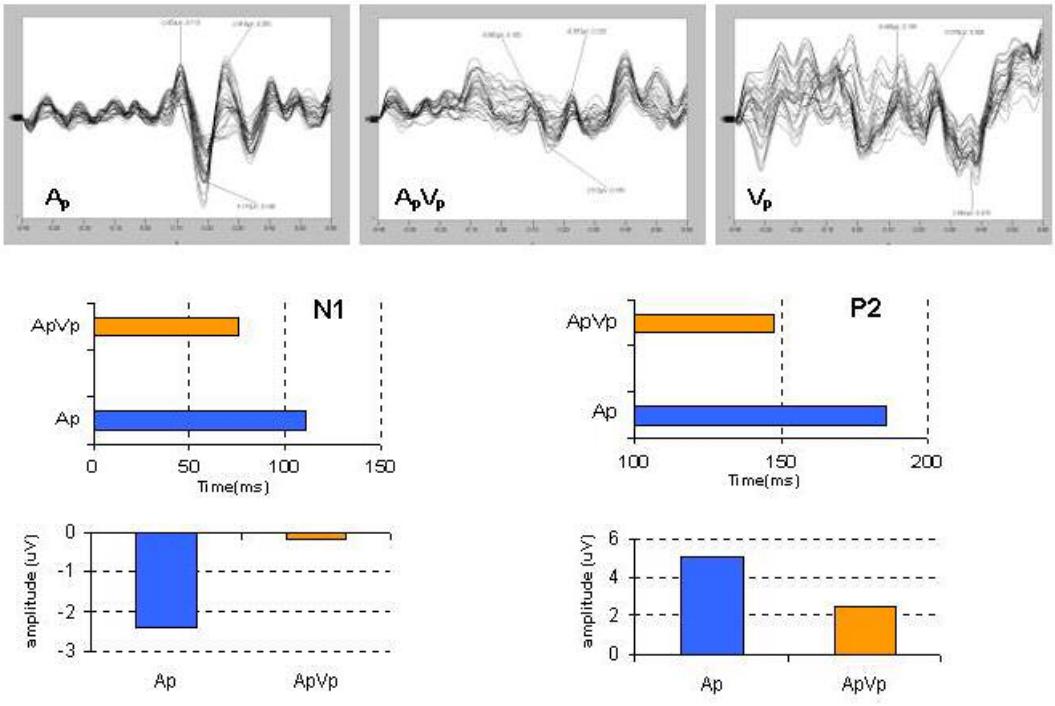


Figure 3.5: Temporal facilitation and amplitude reduction of the auditory N1/P2 complex in ApVp condition as compared to Ap

Suppression of visual precedence: contamination of auditory-evoked potentials by visual onset potentials

Second, in the hypothesis that precedence of visual inputs is at the origin of the amplitude and temporal facilitations effects, taking out visual information that precedes the audio speech onset should cancel out these effects.

Thus (everything else being equal), trA_pV_p (no visual precedence - same informational content) and trA_pV_{noise} (no visual precedence – no informational content) should suppress both amplitude reduction and temporal facilitation of the auditory N1/P2 complex. However, this procedure was predicted to contaminate the auditory evoked potentials due to *superposition effects*

(i) Large visual evoked potentials occurring at the same time as the auditory evoked potentials (~40 to 150ms) may cause a spread of activation that can potentially create artifacts. If such is the case, the recordings may not reflect accurately underlying neural interactions. This issue was already considered in Experiments 1 and 2, where fade-in frames were added in order to avoid abrupt visual onsets.

(ii) The predominance of large visual-evoked potentials may hide / superimpose on potential neural signals of interest that are of smaller amplitude.

Figure 3.6 reports the Global Field Power (GFP) across all participants (N=11) over all electrodes obtained for A_pV_p, trA_pV_p and trA_pV_{noise}. As predicted, no N1 could readily be distinguished from visual potentials.

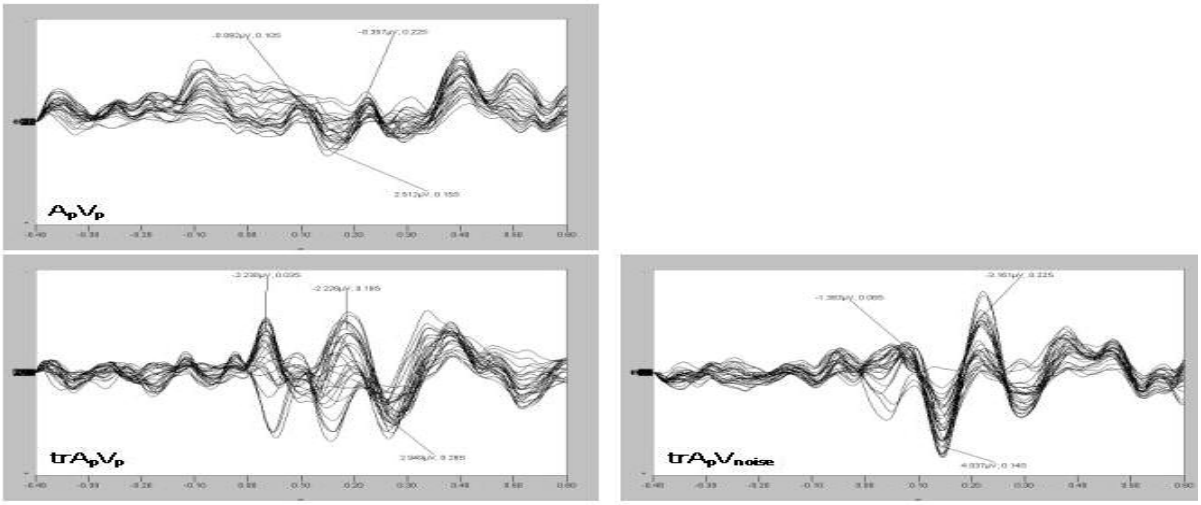


Figure 3.6: Contamination of auditory evoked-potentials by abrupt visual onsets.

Partial suppression of visual precedence: amplitude decrease but no temporal facilitation

As an alternative, two different stimuli were created ($A_{pp}V$ and A_pV_{still}) that do not provide explicit speech information in the visual domain but should avoid visual contamination of the auditory evoked potentials.

In conditions $A_{pp}V$ and A_pV_{still}, no significant temporal facilitation was predicted because visual information is ambiguous and does not provide a clear indication of visemic class (as compared to the original visual [pa]). However, reduced amplitude of the N1/P2 complex was hypothesized when facial information was presented. Precedence of face possibly elicits a syllabic-based or bimodal mode of processing

independently of the speech content. Figure 3.7 reports the results for these two conditions in comparison with the effects observed for A_pV_p and A_p.

Results show that both amplitude reduction and latency facilitation were present with still face and backwards going visual speech presentations but to a lesser extent than congruent A_pV_p when compared to A_p alone. These results indicate that the presence of facial information alone impacts the processing of auditory speech.

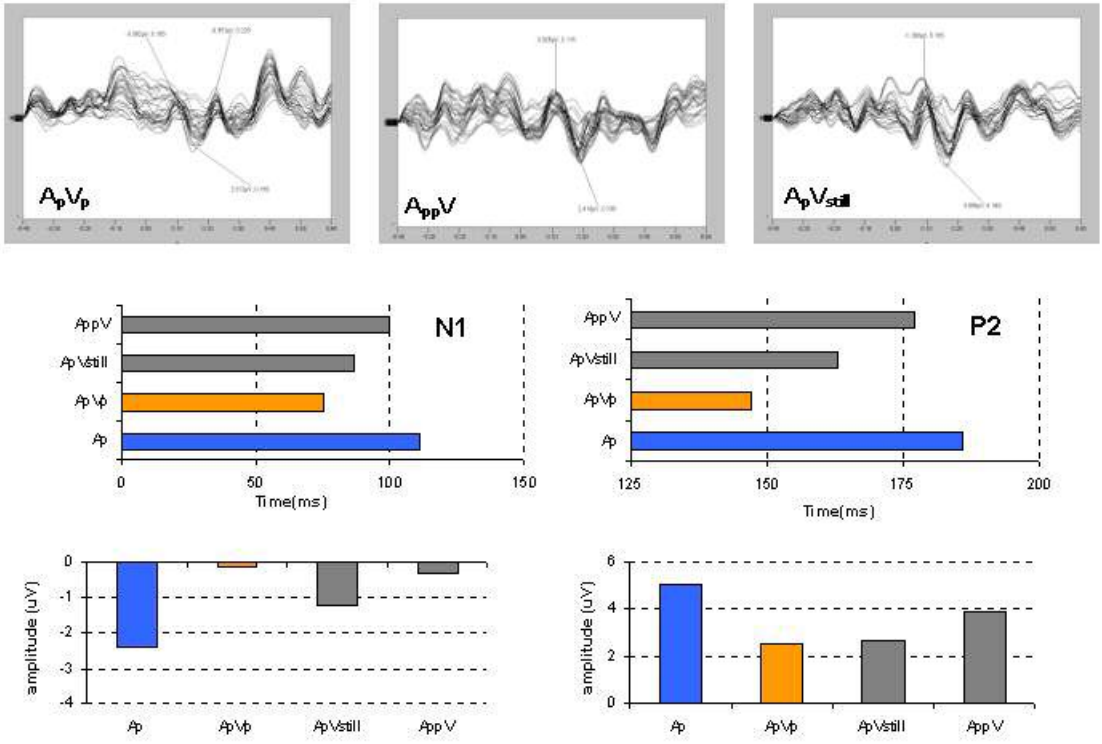


Figure 3.7: Effect of backward and still visual inputs: variable effects

<u>Reversed audio speech: amplitude decrease but no temporal facilitation</u>

In the $_pAV_p$ (reversed audio) condition, the amplitude reduction but not the temporal facilitation of the N1/P2 audio complex was predicted. Again, the natural

precedence of the face and speech informational content it provides engages the speech system in bimodal mode of speech processing. Second, the basis of redundancy in reversed speech is naturally contained in the dynamics of auditory speech, which are now perturbed. Figure 3.8 provides a comparison of the auditory evoked-potentials observed at a centro-parietal recording site (CPz) between reversed audio alone (black trace) and paired reversed audio and visual [pa] (red trace).

As predicted, no temporal facilitation but amplitude reduction of the N1/P2 auditory complex was observed.

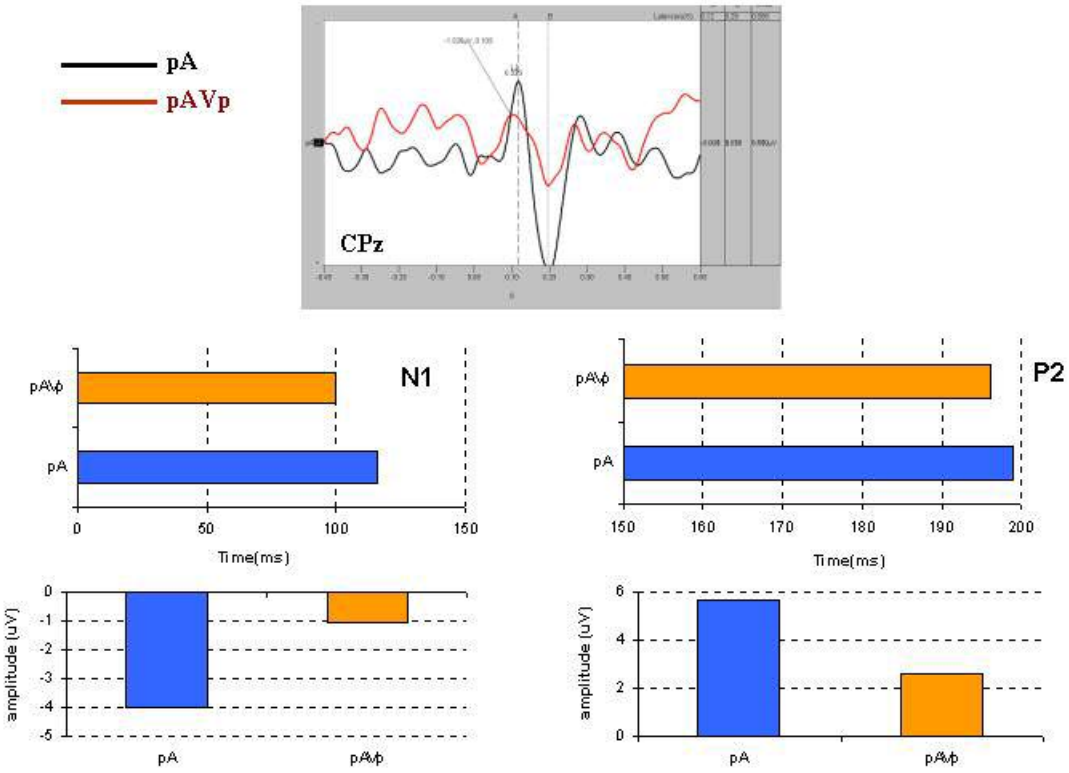


Figure 3.8: Effect of visual [pa] on reversed auditory speech: no temporal facilitation, decreased amplitude

Auditory-visual noise pairing: no temporal facilitation but amplitude increase is observed

An arbitrary AV noise pairing was used that follows the same dynamics of AV speech (i.e. precedence of visual noise). AV noise was predicted to lead to equal-to-enhanced amplitude of the N1/P2 complex and to little-to-no temporal facilitation. Enhanced amplitude was predicted on the basis of the artificial nature of AV pairings and in agreement with prior findings for arbitrary pairings (e.g. Giard and Peronnet, 1999). Second, slight temporal facilitation was predicted because visual precedence may facilitate the detection of auditory input. *{Note*: Chapter V will provide further empirical findings and suggestions as to what can be expected in terms of electrophysiological recording depending upon the saliency and the nature of the AV stimuli. Here, I only provide a comparison and control stimulation that is explicitly uninformative in the speech domain.}

As predicted, no temporal facilitation was observed but an enhanced amplitude of the N1/P2 complex was observed in bimodal conditions.

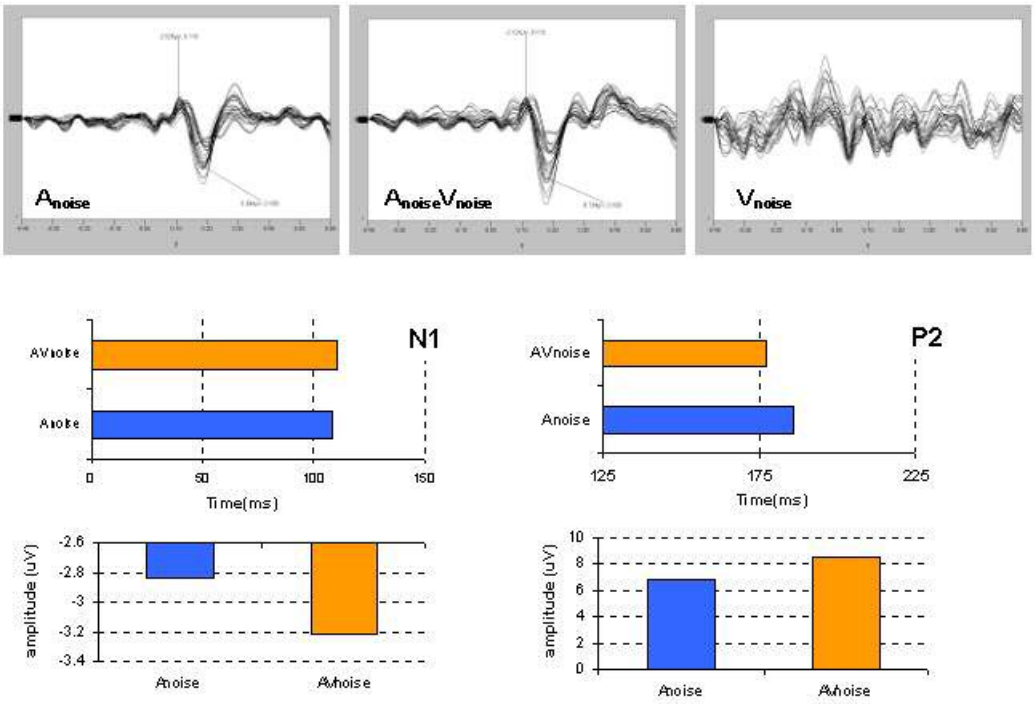


Figure 3.9: Arbitrary AV noise pairing: no temporal facilitation, enhanced amplitude

As predicted no temporal facilitation was observed and both the N1 and the P2 components were enhanced in amplitude as compared to the audio alone condition.

<u>Audio noise and visual place-of-articulation: temporal facilitation but no amplitude decrease</u>

The visual context of auditory noise was further tested with a visual [pa]. Figure 3.10 gives a summary of A_{noise} paired with noise or visual [pa] in comparison with effects reported for natural congruent AV speech. While temporal facilitation was observed with audio noise dubbed onto visual [pa], no amplitude decrease was

observed.

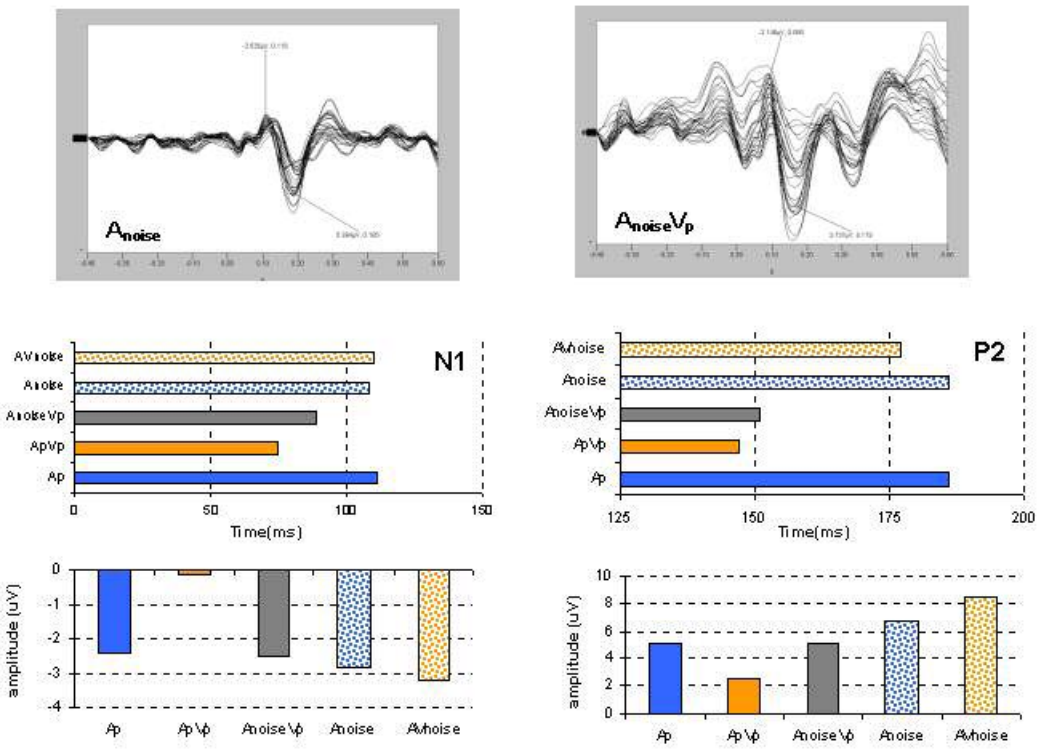


Figure 3.10: Visual context effects on auditory evoked-potentials to auditory noise

Chapter 4: Neural correlates of auditory-visual speech desynchronization

The temporal window of integration for auditory-visual (AV) speech processing has consistently been estimated to be ~250ms across a variety of psychophysical studies. Additionally, fMRI studies have described cortical sites involved in the integration of AV speech, but their dynamics and specific functions remain mostly unknown.

In this chapter, the spectral dynamics of electroencephalographic (EEG) recordings in response to the presentation of desynchronized auditory-visual (AV) speech syllables are described. Two different tasks were compared: a temporal order judgment task and an identification task. Both tasks employed a similar set of AV speech stimuli (consonant-/a/ syllables, e.g. [pa]). EEG recordings during these tasks showed that the pattern of cortical activation varied with the cognitive demands of the task. Specifically, theta (4-7Hz, ~200ms period) power was observed to be larger in the temporal judgment task (TOJ) while gamma power (25-55Hz, ~25ms period) was more pronounced in the identification task (ID). Interestingly, the ratio of gamma over theta power varied systematically with auditory-visual speech desynchronization in the TOJ task and was lateralized to the right-hemisphere. This pattern of lateralization became apparent as the AV speech stimuli got closer to synchrony. Results are discussed in the context of a forward model of AV speech perception.

4.1 Introduction

Psychophysical evidence suggests that multisensory interactions occur at various levels of perceptual analysis. In the auditory and visual modalities, numerous biasing effects of inter-sensory spatial and temporal sensitivity have been shown (e.g. Sekuler & Sekuler, 1997, Shams *et al*, 2002; Lovelace *et al*, 2003; Recanzone, 2003; Wada *et al*, 2003; Zwiers *et al.*, 2003; Bertelson & Aschersleben, 2003). Striking examples of AV interactions are found in the speech domain, where the dubbing of an audio stimulus [pa] onto a visual place of articulation [ka] results in the unique percept [ta], an illusion classically referred to as 'McGurk fusion' (McGurk & McDonald, 1976). Although McGurk fusion is often cited as evidence for AV integration, it is not the only example within the speech domain. In fact, interactions among audio and video speech components appear to be a natural adaptation of the speech system, one which emerges early on during infancy (e.g. Dodd, 1974; Kuhl & Meltzoff, 1984).

Temporal resolution of multisensory convergence sites

The neural mechanisms underlying the integration of auditory and visual speech inputs are not well understood. The convergence of multisensory inputs to neural sites of multisensory convergence is often described as the prevailing mechanism for multisensory integration. Sensory pathways converge onto sub-cortical (e.g. Superior Colliculus (Stein & Meredith, 1993)) and cortical sites [e.g. Superior Temporal Sulcus (STS) (Bruce et al., 1986), Prefrontal cortex (PFC)

(Benevento *et al*, 1977; Fuster *et al*, 2000)), Temporo -Parietal cortex (TPC) (Leinonen *et al*, 1980)], where multisensory neurons respond in a 'supra -additive' manner to spatio-temporally coincident inputs. Supra-additivity is defined as a higher spiking rate and a longer duration of the multisensory cell's output that which would be predicted by summing its output responses to the same stimuli presented unimodally (Stein & Meredith, 1983).

The response characteristics of a multisensory neuron are a function of the cell's receptive field; supra-additive responses are observed when multisensory inputs originate form the same spatial location and at the same time ('spatio-temporal coincidence principle', Stein and Meredith, 1993). The simultaneity threshold for multisensory cells has been seldom studied. As a rule, supra-additivity is observed when an input reaches a multisensory neuron while the cell is still responding to the preceding stimulus (Meredith *et al*, 1987). In other words, the spiking duration of a neuron's output establishes the *temporal window of supra-additivity* for a subsequent stimulation. Supra-additivity is observed for asynchrony values as large as 1.5 seconds in the SC (Meredith *et al*, 1987) and in the PFC (Benevento *et al*, 1977).

Temporal window of integration in AV speech

Building on the spatio-temporal coincidence principle, AV speech provides a multisensory event where the speaker's face is the common spatio temporal source of the signals. Specifically, it is in the 3-4 Hz range that acoustic amplitude envelope fluctuations and the movement of the lips are most correlated (Grant, 2001; Grant &

Greenberg, 2001). If AV speech integration is mediated by multisensory neurons, a straightforward prediction is that *decorrelating* the AV inputs should lead to less integration. Two signal manipulations can be used to this effect: desynchronization of the AV inputs and/or dubbing of an audio speech token onto discrepant visual speech inputs.

Studies of congruent and incongruent (McGurk & McDonald, 1976) AV speech desynchronization have shown that the degree of AV integration remains unaffected by asynchronies less than or equal to ~250ms (Massaro *et al.*, 1994; Munhall *et al*, 1996). This ~250ms window was observed whether participants were asked to identify the AV speech tokens (Massaro *et al*, 1994; Munhall *et al*, 1996) or to judge their temporal relationships (see Chapter 2; Conrey & Pisoni, 2003). These results suggest that (i) early on, the neural computations underlying the integration of AV speech operate on a ~250ms time scale and that (ii) the temporal resolution of multisensory neurons does not appear to provide a sufficient constraint for this process.

Speech non-invariance

The discretization of information that enters the system as a continuous flow of physical inputs is a fundamental instantiation of brain function (VanRullen & Koch, 2002). In the speech domain, the spectral complexity of acoustic inputs is affected by the rapid rate of information flow, such that preceding and following articulatory movements (i.e. co-articulatory movements) may influence the acoustic

pattern at any instant (Liberman *et al*, 1965). These backward and f orward spectro-temporal perturbations lead to the classic problem of speech segmentation, where it has been argued that the lack of clear acoustic invariance should be compensated for by an early restructuring of acoustics inputs by the auditory system (Liberman *et al.*, 1967; Studdert-Kennedy, 1981; Liberman & Mattingly, 1985; Liberman, 1995).

Multiple temporal resolutions

The organization of sensory systems in the cortex appear to be involved in the presence of 'global' and 'local' analytic processing streams (e.g. Ungerleider & Mishkin, 1982; Rauschecker & Tian, 2000; Alain *et al*, 2001). The notion of 'global' versus 'local' refers here to the level of information processing. The 'local' level is associated with a more hierarchical type of processing (high-temporal resolution but slower) while the 'global' level is low-temporal resolution (but faster) (e.g. Sergent, 1982). The notion of multiple temporal resolutions windows in the analysis of sensory inputs as the basis for hemispheric functional differentiation has been put forward in the visual domain (Brown & Kosslyn, 1993) and proposed as basis for general perceptual processing (Ivry & Robertson, 1998).

By analogy, in the speech domain, the restructuring of acoustic inputs may implicate simultaneous windows of temporal integration at various time scales, in particular on a phonetic and sub-segmental (fine-grained temporal analysis) and a syllabic-scales (coarser temporal analysis) (Poeppel, 2003). Specifically, in the 'asymmetric sampling in time' (AST) framework, Poeppel (2003) proposes that the

cortical dynamics of speech processing operate on two time constants: a fine-grained temporal analysis (~25ms) and a coarser temporal analysis (~250ms). The shorter time scale (~25ms) is slightly left lateralized whereas the longer time scale (~250ms) is right lateralized. Hence, each temporal resolution is *simultaneously*present. The AST highlights two time-scales that are of particular importance in speech: 250ms, which corresponds to syllabicity (Arai & Greenberg, 1998) and ~25 ms which relates to phonetic or featural processing (e.g. differences along a voicing or place of articulation continuum) (Rosen, 1992; Greenberg, 1996).

Electrophysiological correlates of AV speech integration

In the present series of experiments, electrophysiological (EEG) recordings using synchronized congruent and incongruent AV speech showed a modulation of in the amplitude of early auditory specific evoked-related potentials (N1/P2 complex) in amplitude, which lasted ~250ms (see Chapter 3). The duration of this amplitude decrease suggests a possible mediation of AV speech integration in the theta band, a low-frequency brain oscillation with a period of ~150-250ms (4 7Hz). In addition to this amplitude decrease, a latency facilitation of the N1/P2 complex was observed in the tens of milliseconds, suggesting the involvement of a finer-temporal resolution mechanism possibly evolving in the gamma band; the gamma band is a high frequency component with a period of ~25ms (~40Hz) which has been associated with local computations in sensory systems (e.g. Tallon-Baudry, 1999). These two temporal resolutions (~250ms and ~25ms) also appear in a previous EEG study of

AV speech perception (Callan *et al.*, 2001), where a sustained power increase of the gamma band (30-80Hz, ~25ms) was found that lasted ~150ms. The source of this gamma activation was localized in the left Superior Temporal Gyrus (STG).

Together these results suggest that the early (i.e. within ~250ms post-stimulation) neural underpinnings for AV speech integration evolve on two different time-scales: a coarse-temporal resolution (theta) and a fine temporal resolution (gamma).

In this context, we wanted to determine whether, as predicted by the AST (Poeppel, 2003), AV speech integration would involve specific hemispheric patterns of lateralization in the theta and gamma frequency ranges. Studies of AV speech integration in brain damaged patients and normal populations have not shown clear hemispheric contribution (Campbell *et al*, 1990; Soroker *et al*., 1995), yet the integrity of the corpus callosum appears necessary for the AV speech integration process (Baynes *et al*, 1994).

We conducted an EEG study using desynchronized congruent and incongruent (McGurk) AV speech and pursued two major questions.

First, as suggested by prior psychophysics, the early stage of AV speech stimuli (i.e. the first ~250ms) appears to constrain the integration process independent of cognitive task. In particular, whereas a fine-grained analysis of speech inputs is desirable for *identifying* the stimuli (~25ms), a coarser analysis of inputs may suffice in a *temporal judgment task* (~250ms). Hence, our first goal was to establish whether

a clear modulation in the theta and gamma ranges could be established as a function of cognitive demands (i.e. identification (ID) versus temporal order judgment (TOJ)).

Second, on the basis of the AST proposal (Poeppel, 2003), we sought to determine whether gamma (25-55Hz, ~25ms) and theta (4-7Hz, ~250ms) power were sensitive to the synchronization of AV speech tokens (i.e. if the temporal resolution of AV speech integration observed psychophysically would be reflected cortically in these frequency ranges). For both questions, hemispheric differentiations in these frequency ranges were investigated.

4.2 Materials and methods

Participants

Ten native speakers of American English (7 males, mean age = 22.7 years old) were recruited from the University of Maryland population. Participants reported no hearing problems; all had normal or corrected-to-normal vision and were right-handed. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli

To preserve the natural relationship between auditory (A) and visual (V) inputs, we used natural AV speech stimuli consisting of a woman's face articulating

nonsense syllables. Original stimuli consisted of syllables [ta] and a McGurk fusion stimulus created by dubbing an audio [pa] onto the woman's face articulating [ka] (the audio [pa] was aligned with the original [ka]). The average duration of the AV stimuli was 2590 ms, including video fade-in (8 frames) - added at the onset of the video to avoid abrupt visual onset in the electrophysiological recordings-, neutral still face (10 frames) -during which classic face processing related potentials were expected to occur-, natural production of the syllable [ka] or [ta] (variable number of frames) and fade-out (5 frames).

Desynchronization of AV stimuli was created by displacing the audio file by 10 (+/- 333ms), 4 (+/- 133ms) and 2 (+67ms) frames prior to (-) or after (+) the natural occurrence of the auditory utterance with respect to the movie file. These asynchrony values were chosen on the basis of the response curves obtained in a prior study of identification and simultaneity judgment of AV speech desynchronization (see Chapter 2). In that study, participants could detect +/- 333ms of desynchronization, but could not detect +67 and +133ms of desynchronization and were at chance at -133ms of desynchronization (i.e. perceptual bistability) both in identifying the stimuli or detecting stimuli asynchrony. Our reasoning in choosing those values was that both tasks: detection of asynchrony (studied in the previous experiment) and temporal order judgment (the task in the current experiment) in AV speech would lead to similar psychophysical results -based upon the time-scale of the perceptual phenomenon. Thus, five asynchrony conditions per stimulus type (real [ta] vs. McGurk [ta]) were chosen for the EEG experiment on AV speech temporal order judgment (TOJ). Interstimulus intervals (ITIs) were pseudo-randomly selected among

5 values (500 ms, 750 ms, 1000 ms, 1250 ms and 1500 ms). Stimuli were pseudo-randomly intermixed.

Procedure

Participants took part in two three-hour EEG sessions. In the first session (see Chapter 3), participants were presented with synchronized congruent and incongruent (McGurk [ta]) bimodal (AV) and unimodal (A, V) [ka], [pa], [ta] stimuli (for a total of 1000 trials, 100 presentations per stimulus). A single-trial three alternative forced-choice procedure was used (3 AFC) and the three choices were [ka], [pa], or [ta]. Participants signaled their responses by button-press. In bimodal conditions, participants were asked to make a choice as to "what they hear while looking at the face". In the unimodal conditions (A, V), participants were asked to make a choice as to what they hear or see, for the A or V conditions respectively.

In the second session (Experiment 6), the same participants were presented with desynchronised congruent and McGurk [ta] in five blocks of 200 stimuli each. Asynchrony conditions included audio leads (–333ms and -133ms) and audio lags (+67ms, +133ms, and +333ms).

The two EEG sessions were scheduled a minimum of 1 week apart. In the temporal order judgment task (Experiment 6, TOJ), participants were asked to judge whether the audio came first, whether the video came first, or whether the audio and video stimulus components were synchronized. In both sessions, a single-trial 3 alternative-forced choice (3AFC) procedure was used. No feedback was provided and

participants were not asked to follow a particular strategy (for instance, they were not told to lip-read) nor were they provided with training.

Participants were placed about 1m from the visual display, with the movie subtending a visual angle of 8.5º in the vertical plane and 10.5º in the horizontal plane. Videos were displayed centered on a 17" G4 monitor on a black background. Sounds were presented through Etymotic ER3A earphones connected to the G4 computer through a sound mixer table at a comfortable level of approximately 70 dB SPL. Lights were dimmed before recordings.

Electroencephalographic Recordings

EEG recordings were made using a Neuroscan system (Neurosoft Systems, Acquire 4.2b), with 32 Ag/AgCl sintered electrodes mounted on an elastic cap (Electrocap, 10-20 montage) individually chosen according to head size (distance inion-nasion and circumference). Data were continuously collected in AC mode at a sampling rate of 1kHz. The pre-amplifier gain was 150 and the amplifier gain was set to 1000 for all thirty-two recording channels. Reference electrodes were left and right mastoids grounded to AFz. A band-pass filter from 1Hz to 100Hz was applied online. Two electrodes recorded horizontal eye movements (HEOG) and two others recorded vertical eye movements (VEOG) for off-line data artifact reduction and rejection.

EEG data pre-processing

After artifact rejection (extraction of noisy data originating from non-ocular movements) and ocular artifact reduction (linear detrending based upon blink template), epochs were baseline corrected on a pre-stimulus interval of 400 ms chosen prior to stimulus onset (i.e. prior to auditory onset (Experiment 3, A condition) or prior to visual onset (Experiment 3, V alone and AV, and Experiment 6, all conditions). A threshold of ± 100μV was used to reject residual artifacts. Approximately 75-80% of the original recordings were preserved (about 20 trials per stimulus condition were rejected). Trials were sorted by stimulus and response. In Experiment 3 (ID task), only [ta] responses to congruent AV [ta] and McGurk [ta] were considered (the rate of McGurk fusion was ~85%, thus most data samples were preserved, see Results section). In Experiment 6 (TOJ task), EEG recordings were analysed based upon the maximum judgments for each stimulus type. Thus, for -333ms audio leads, only "audio first" responses were considered, while for +333ms audio lags, only "visual first" responses were considered" and for the remaining 3 conditions (-133ms audio leads and+67ms, +133ms audio lags), "simultaneous" responses were considered. In the McGurk condition, -133ms audio lead provided an equal number of data for "audio first" and "simultaneous". Both were separately considered.

Time-frequency analysis

Epoched EEG data were imported into Matlab using EEGlab (Salk Institute). Matlab routines were created for the remaining analyses. Regions of interests

comprising three to five electrodes were defined in order to increase the signal to noise ratio. The five regions of interest (ROI) were: frontal (F, which includes electrodes FC3, FC4 and FCz), occipital (O, which includes O1, O2 and Oz), centro-parietal (CP, which includes CP3, CP4 and CPz), right hemisphere (RH, which includes P8, TP8, T8, FT8 and F8) and left hemisphere (LH, which includes P7, TP7, T7, FT7 and F7). A Morlet wavelet transform was separately applied for each ROI on a single trial basis and normalized on the 400ms pre-stimulus interval. Wavelet coefficients were then averaged across trials for each stimulus for each participant. Time-frequency spectra were produced for each stimulus condition and each individual.

Statistical analysis

Frequency bands of interest (FBI) were defined as theta (θ, 4-7Hz), alpha1 (α_1, 8-10Hz), alpha 2(α_2, 10-12Hz), alpha (α, 8-12Hz), beta1 (β_1, 13-18 Hz), beta2 (β_2, 18-25Hz) and gamma (γ, 25-55Hz). Wavelet coefficients comprised within the defined frequency bands were averaged for each ROI both by stimulus and by individual. We here report results for theta and gamma only. Non-overlapping windows of 25 ms were then applied to provide a power estimate per FBI and per ROI and exported for statistical analysis in SPSS (SPSS Inc., Illinois). Reported p values are Greenhouse-Geisser corrected when sphericity could not be assumed.

4.3 Results

Behavioral results

Behavioral results replicate prior findings, where AV speech desynchronization within ~250ms did not significantly affect performance in identification or subjective simultaneity ratings (see Chapter 2). Figure 1 (panel a) reports the rate of correct identification for audio alone (A), video alone (V) and synchronized AV conditions recorded during the first session (Experiment1, ID). The McGurk stimulus induced ~85% of fusion and congruent AV [ta] was correctly identified ~95%. Figure 1 (panel b) shows the grand average (n=10) rate of simultaneity judgment as a function of asynchrony condition. Note that the spread of the 'plateau of simultaneity' does not result from inter-individual variability but rather reflects a consistent pattern of individual temporal window of integration, in agreement with prior reports of AV speech desynchronization studies. Figure 1 (panels c and d) shows the distribution of individuals' responses ('audio first', 'simultaneous', and 'visual first') as a function of asynchrony in congruent AV [ta] (filled squares) and McGurk [ta] (open squares) conditions, respectively. Each square corresponds to an individual's response and the line represents the grand average data.

Although we used a temporal order judgment paradigm, a window of temporal integration similar to that obtained in a detection paradigm emerged. The boundaries of the temporal window of integration also fit the asynchrony values at which a decrease of subjective simultaneity judgment was predicted to occur.

EEG results

Figure 2 (left panel) provides an example of grand average event-related potentials (n=10) for a synchronized (black) and desynchronized (blue, audio lag of 67ms) congruent AV [ta] stimulus and for three regions of interest (top is frontal, middle is centro-parietal and bottom is occipital). These recordings include a 400ms pre-stimulation baseline and were band-passed filtered (1-55HZ). Event-related potentials elicited by visual inputs (naturally preceding the auditory inputs in synchronized *and* in audio lead conditions) are clearly apparent but will not be considered in this report.

Figure 2 (right panel) depicts examples of time-frequency spectra obtained via Morlet wavelet transform for one individual. Wavelet transforms were applied on single-trials of similar duration as those shown in the grand-average event-related potentials and were baseline corrected on the 400ms pre-stimulus interval power spectra (top is frontal, middle is centro-parietal and bottom is occipital).

Because we are here interested in the early integrative mechanisms of AV speech perception, our analysis focused on a window centered from -100ms to 400ms post-auditory onset.

Theta and gamma power across tasks

We first compared the theta (4-7Hz) and gamma band power (25-55Hz) in the identification (ID, synchronized tokens) and temporal order judgment (TOJ, +67ms perceived as simultaneous) paradigms.

A repeated measures ANOVA with factors of frequency band (2: theta, gamma), ROI (5: O, CP, F, LH, and RH), task (2: ID, TOJ), stimulus type (2: real [ta], McGurk [ta]), and time window (21) showed a significant two-way interaction of task with time window ($F(3.714, 33.423)=9.151$, $p \leq 0.0001$) and a significant three-way interaction of task with frequency band and time ($F(3.078, 27.702) = 5.309$, $p \leq 0.005$).

Figure 3a an 3b show the power in the theta band collapsed over 400 ms post-auditory onset in the left hemisphere (LH) and in the right hemisphere (RH), respectively. The theta band power in the TOJ task was found to be higher in both hemispheres as compared to the ID task from -100ms to 400ms post auditory onset on. A repeated measures ANOVA with factors of ROI (2: LH, RH), task (2: ID, TOJ), stimulus type (2: real [ta], McGurk [ta]), and time window (21) was conducted with theta band power as the dependent variable. This analysis revealed a significant two-way interaction of hemisphere with time window ($F(20, 180) = 2.27$, $p \leq 0.02$), and of task with time window ($F(20, 180) = 4.087$, $p \leq 0.0001$). In addition, there was a four-way interaction of task with ROI (LH, RH), stimulus and time window ($F (20, 180) = 2.048$, $p \leq 0.007$).

An additional repeated measures ANOVA was conducted with factors of ROI (3: F, CP, O), task (2: ID, TOJ), stimulus type (2: real [ta], McGurk [ta]), and time window (21), again with theta band power as dependent variable. This analysis showed significant interactions, in particular a two-way interaction of task with time window ($F(2.924, 26.318) = 6.79$, $p \leq 0.002$) and a three-way interaction of task with ROI and time window ($F (2.779, 25.008) = 4.891$, $p \leq 0.009$).

Figures 3c and 3d report the power in the gamma range over 400ms post-auditory onset in the LH and the RH, respectively. In this frequency range, power was found to be greater for the ID task as compared to the TOJ task. A repeated measures ANOVA with factors of ROI (2: LH, RH), task (2: ID, TOJ), stimulus type (2: real [ta], McGurk [ta]) and time window (21) with gamma band as the dependent variable showed a marginally significant one-way interaction of task ($F(1, 9) = 4.508$, $p \leq 0.063$). Because significant one-way interactions of both theta and gamma frequency band with task and time were found ($F(3.714, 33.423) = 9.151$, $p \leq 0.0001$), we further analyzed the ratio of gamma and theta power bands.

Gamma-Theta power ratio and patterns of hemispheric lateralization

Following the proposal by Poeppel (2003), we looked at possible interactions of theta and gamma frequency ranges, in particular across hemispheres. When analyzing the ratio of gamma over theta bands, a trend toward lateralization was obtained in both tasks with a higher ratio in the left hemisphere as compared to the right. However this hemispheric difference failed to reach significance ($F < 1$) when examined via repeated measures ANOVA with factors of ROI (2: LH, RH), task (2: ID, TOJ), stimulus type (2: real [ta], McGurk [ta]), and time window (21) for the gamma-theta ratio. A possible reason for not reaching significance may stem from the low signal-to-noise ratio given that only one condition could be used for each stimulus in this analysis. A second possibility may be due to the reduction of the temporal resolution -or dynamics- to non-overlapping 25 ms window bins. Paired t-

tests do produce marginally significant results in the real [ta] condition, in both ID ($p \leq 0.039$) and TOJ (, $p \leq 0.04$), but not in the McGurk case, which showed overall a greater variability ($p \leq 0.108$) and TOJ ($p \leq 0.57$).

Third and importantly, while collapsing across conditions may have provided more power, in particular for the TOJ task, the hemispheric gamma-theta ratio was most differentiated at +67ms and faded as the asynchrony value increased. We thus looked at the variations of lateralization across asynchrony values.

We found that for both congruent and incongruent stimuli, the right hemisphere (RH) gamma/theta ratio was correlated with the desynchronization of stimuli. In particular, the RH gamma-theta ratio was observed to decrease as the stimuli were synchronized whereas in the left hemisphere, no consistent relation between gamma/theta ratio and temporal asynchrony was observed. Figure 4 reports the hemispheric profiles of the gamma-theta ratio for real AV [ta] and McGurk [ta]. Reported p-values were obtained using paired t-tests paired comparisons over a 500ms period (-100ms to +400 ms around the auditory onset) and between hemispheric gamma-theta ratios.

4.4 Discussion

An EEG study of AV speech desynchronization was undertaken to determine the neural correlates of the temporal window of integration previously observed in AV speech integration (see Chapter 2). Two experimental paradigms were used, and a

time-frequency analysis of the EEG data was performed. There were three primary findings.

First, a cortical state-dependent activation in the theta (4-7Hz) and the gamma (25-55Hz) frequency ranges was observed that varied as a function of the cognitive demands in each task. Specifically, whereas a greater activation in the gamma range was observed in the ID task as compared to the TOJ task, the opposite pattern was found in the theta range. These different states of activation suggest that AV speech stimuli undergo a different mode of information extraction depending on the cognitive context.

Second, within this pattern of activation, a trend towards hemispheric differentiation was observed, where the gamma/theta power ratio tended to be higher in the left-hemisphere than in the right hemisphere. This result is in agreement with the AST predictions (Poeppel, 2003).

Third, the lateralization of the gamma-theta power ratio was a function of the desynchronization of the stimuli: a greater hemispheric differentiation was observed for desynchronization values within the temporal window of integration i.e. when "simultaneous" judgments were dominant. This result reflects the width of the temporal window of integration observed previously in psychophysics (see Chapter 2; Conrey & Pisoni, 2003). In particular, systematic variations of the gamma/theta power ratio in the right hemisphere suggest a right lateralization of 'global' or syllabic based cross-modal binding of information. The right hemispheric pattern of lateralization in the context of a temporal judgment task is also in agreement with prior PET findings (e.g., Bushara *et al*, 2001).

Theta and gamma functional trade-off: temporal information versus percept formation

The level of general arousal is a crucial (and natural) variable to consider in evaluating how the cortex extracts information and builds up a perceptual representation. Recent studies have shown that top-down modulation can be observed as early as primary sensory-specific cortices (Schulman *et al*, 1997; Pessoa *et al.*, 2003). This finding suggests that information extraction is constrained very early on by the internal state of the system -i.e. inputs at time t+1 depend on the state of the system at time t.

In the ID task, AV speech stimuli are processed until a unique perceptual representation has been achieved. Thus, the task focuses on the *integration* (i.e. unification) of sensory-specific inputs. Regardless of the underlying integrative mechanism, the integration of AV speech inputs seems to occur 'automatically' (McGurk & McDonald, 1976, 1978; Campbell *et al*, 2000; Colin *et al*, 2002) and the very nature of multisensory fusion implies that modality-tagging is lost after information has been unified into a single percept. For discrepant stimulation however, such as in the McGurk effect, identification may be affected by attentional drive (Summerfield & McGrath, 1984; Massaro, 1998; Soto-Farraco *et al*., 2003; Tiippana *et al*, 2004).

In contrast to the ID task, the TOJ task emphasizes the temporal comparison of inputs in the time domain. Thus, assuming that only one (completed) percept and not three (A, V, and AV) are present at once in the speech system, auditory and visual

inputs should be compared at least prior to unification (i.e. prior to losing their respective sensory-temporal specificity). In this context, several hypotheses can be suggested regarding the early stages of information processing in AV speech (i.e. within ~250ms).

First, the TOJ and ID tasks may follow two segregated routes of information processing. For instance ID may follow a speech-specific pathway while TOJ may engage a parallel 'time' pathway, in which only temporal information is being extracted. The first hypothesis would be in line with the notion of an independent 'internal clock' mechanism or a central time processor (e.g. Treisman *et al*, 1990, 1994). However, this option would suggest that the ~250ms temporal window of integration observed in identification and temporal judgment tasks in AV speech originate from two independent neural processes. While theoretically possible, this hypothesis does not offer the most parsimonious implementation of AV speech processing particularly in the earliest stages of information processing.

A second possibility, in line with the concept of 'spatio temporal coincidence' is that activation of the module responsible for the temporal evaluation of incoming AV speech inputs precedes the speech-specific evaluation stage –i.e. spatio-temporal coincidence is computed by multisensory neurons. However, the temporal tuning of multisensory cells is very lax (Meredith *et al*, 1999). As mentioned earlier, the position of multisensory integration by convergence in the hierarchy of neural processing is difficult to establish (cf. Chapter I, working hypothesis regarding multisensory-by convergence). The precedence of temporal evaluation over AV

integration could hypothetically be mediated by early inter-sensory connectivity (e.g. Falchier *et al*, 2001). However, this option seems inconsistent with the ~250ms width of the temporal window of integration, for one would expect a better temporal resolution perhaps approximating that observed in non-speech stimuli and in the order of ~20-50ms (e.g. Hirsh & Scherrick, 1962; Lewald *et al*, 2001; Zampini *et al*, 2002).

A third hypothesis is that the neural mechanisms operating in AV speech integration also constrain the temporal resolution of the AV speech system. According to this view, *both the identification of and the temporal evaluation of sensory-specific inputs rely on the early dynamics of AV speech interactions*. More precisely, it is here proposed that the neural processes within ~250ms post-stimulation in the ID and in the TOJ tasks are identical, but *they operate in different computational regimes*. In both tasks, the gamma/theta power ratio tended to be stronger in the left hemisphere. This observation is inline with prior reports in the literature conferring a major role to the left hemisphere for rapid processing and in particular with the AST proposal (Poeppel, 2003). In speech, this lateralization may underlie two different scales of simultaneous information processing, one at the sub-segmental level in the left hemisphere and the other on the syllabic scale in the right hemisphere (Poeppel, 2003).

The temporal locus of AV speech integration in cortex

Determining when –i.e. at which representational stage- visual speech inputs can modulate auditory processing is crucial for accounts of multisensory perception in general, and for theories of speech perception in particular. From a computational standpoint, the timing of integration determines whether auditory and visual speech information are being evaluated (perceptually categorized) independently prior to being integrated or whether categorization occurs only after integration (Massaro, 1998; Grant, 2002). The time reference used in AV speech integration is the phonetic processing stage, which precedes the phonological categorization of acoustic inputs. Two possible loci of AV speech integration are thus prior to (and at) or after the phonetic stage. The former is commonly referred to as 'early integration' and the latter as 'late integration'.

Multisensory-by convergence and feedback hypothesis

In recent fMRI studies of AV speech processing, Calvert and colleagues (1999, 2000) found a supra-additive activation of the STS. This activation was interpreted as evidence for the mediation of AV speech integration by multisensory cells, whose integrated output would then feedback onto primary sensory cortices (where smaller supra- and sub-additive effects were also observed). This interpretation is consistent with the 'multisensory convergence hypothesis' (Stein and Meredith, 1993).

As Schroeder *et al.*(2003) pointed out, multisensory convergence sites (specifically those comprising multisensory cells that demonstrate supra additivity) are relatively late in the hierarchy of sensory processing streams. Yet, prior EEG findings of AV speech have shown that auditory specific potentials differ in the presence of visual speech inputs as early as 50ms post-stimulation (Lebib *et al.*, 2003) and were systematically speeded up as early as 100ms (Chapter 3). One source of early auditory-evoked potentials is the STG, from which activation is recorded as early as 40 ms post-auditory onset with the magnetoencephalographic (MEG) technique (Näätänen & Picton, 1987; Yvert *et al*, 2001).

In fact, fMRI studies of AV speech have not only reported a supra-additive activation of the STS, but also that of the Superior Temporal Gyrus (STG). The posterior STG has also shown activation in the absence of auditory inputs, i.e. in lip-reading conditions (Calvert *et al.*, 1997; Ludman *et al*, 2000; Bernstein *et al.*, 2001; Calvert & Campbell, 2003) and is involved in the processing of visual biological motion (Grossman *et al.*, 2000; Vania *et al*, 2001; Servos*et al*., 2002) . Because, fMRI does not provide sufficient temporal resolution, it is, however, impossible to determine which area is being activated first.

Hence, an integrative mechanism of AV speech via multisensory STS (whose output provides *feedback* onto auditory cortices) may occur too late to account for the early effects observed in EEG studies. Rather, visual speech information may modulate the *feed-forward* processing of auditory speech.

Temporal window of integration as temporal locus of AV speech perception

In the present study, a cortical state-dependency related to the cognitive demands of the task is reported. This difference is reflected in the global power of the theta (4-7Hz) and gamma (25-55Hz) bands. Specifically, the ID task is associated with a higher power in the gamma band while the TOJ was associated with a higher power in the theta band. High-frequency components such as the gamma band have been associated with 'local' computations and the binding of fine-structured information (e.g. Tallon-Baudry *et al*, 1999; Tallon -Baudry, 2001), whereas low-frequency components such as theta involve more 'global' computations in a possibly amodal mode (Yordanova *et al.*, 2002). Our results suggest that a more analytic ('local') processing mode was engaged in the ID task, as opposed to the TOJ task, where a more synthetic ('global') mode of processing is observed.

The predominance of a ~200ms periodicity (theta) in the TOJ task converges with the temporal window of integration of ~250 ms described psychophysically (Chapter 2; Conrey & Pisoni, 2003) and is in line with the duration of amplitude reduction in auditory evoked potential (Chapter 3). The predominance of ~25ms periodicity (gamma) in the ID task is in agreement with prior reports of temporal facilitation (Chapter 3).

Crucially, the interaction of gamma and theta frequency bands quantified here as gamma/theta power ratio is a means to quantify the simultaneous effects of amplitude modulation (~250ms) and temporal facilitation (~tens of milliseconds) observed in auditory-evoked potentials (Chapter 3). Specifically, the AST proposes

that these two temporal resolution windows (phonetic, ~25ms and syllabic, ~250ms) evolve (i) simultaneously and (ii) are the basis for functional hemispheric differentiation. Under this premise, if AV speech processing follows the predictions of AST, it is in the hemispheric differentiation of the gamma/theta power ratio that the temporal window of integration should be observed (Figure 4, arrows). In other words, the extent of hemispheric differentiation observed when AV speech stimuli are getting closer to synchrony suggests the extent of the temporal window of integration observed psychophysically. From the multiple temporal resolutions view of the AST, the gamma/theta power ratio quantifies the *temporal sensitivity of the integrative mechanism across a phonetic-based and a syllabic-based coding scheme.*

From an AV speech integration viewpoint, the information provided by visual speech is dominantly articulatory-based and in the auditory domain, a major cue for place of articulation is specified by formant transitions, which evolve on the syllabic time-scale (Liberman, 1967). It was mentioned earlier that the ~250ms temporal window of integration emerges intrinsically from the neural integrative mechanisms of AV speech, regardless of the cognitive demands (i.e. ID or TOJ). Thus, the informational content of the first-input modality (i.e. audio or visual leading) remains fundamental in determining the mode of information extraction. In particular, in visual leads condition, the syllabic-scale (theta periodicity, ~250ms) may act as a major time-scale for the processing time of (intended) articulatory gesture, whereas in the audio leading condition, the phonetic-scale (gamma periodicity, ~25ms) may be preferred. If such were the case, the asymmetric profile of the AV speech temporal

integration window observed in a previous study (Chapter 2) may originate from this preferential mode of processing as induced by the initial input-modality. This hypothesis should be reflected in our results as an increased left-hemispheric gamma/theta power ratio for audio leads as compared to video leads. However, the signal-to-noise ratio obtained in our study does not permit us to test this prediction and further experiments may help clarify this matter.

Our results suggest that the *temporal locus of integration* in AV speech cannot be considered a 'point' in time but rather a ~250 ms moment. This ~250ms window is intrinsically defined by the hemispheric variation of the gamma/theta power ratio. Crucially, this integrative window emerges as a non-linearity of the neural dynamics and encompasses both fine-grained (phonetic) and coarse-grained (syllabic) analysis of speech. These results are best interpreted within a speech theory that incorporates an articulatory-based processing of speech, in particular an 'analysis-by-synthesis' model (Halle & Stevens, 1962) of AV speech integration.

Dynamics of a forward model of AV speech integration

In a prior EEG study, we (Chapter 3) proposed a forward or 'analysis-by-synthesis' model of AV speech integration, where the saliency of visual speech information and the redundancy of auditory and visual inputs were considered important constraints on AV speech integration mechanisms. This model is described

in Figure 5. The present results are incorporated in this figure to account for the hypothesized dynamics of the model.

The precedence of visual speech information naturally initiates the speech processing system prior to incoming auditory inputs. This precedence is proposed to provide the *context* in which auditory speech will be processed. In particular, and in line with the Motor Theory of Speech Perception and the 'Analysis by Synthesis' models (Liberman *et al*, 1995; Halle & Stevens, 1962), the processing of visual speech information is dominantly articulatory-based and engages a more global or syllabic type of processing operating over a ~250ms time scale. Thus, note that in our study, visual speech leading conditions (+67, +133, +333 ms) are 'more natural' than the auditory leading conditions (-333, -133 ms).

The processing of visual speech inputs is assumed to be partial or incomplete because visual speech is, in large measure, ambiguous and limited mostly to a place of articulation information. The processing of visual speech leads to an abstract (amodal) representation that originates from this articulatory-based evaluation of the visual signal *prior* to auditory onset. Such articulatory-based computations are in line with the implication of the theta rhythm, which has been shown to be amodal (Yordanova *et al*, 2002), and to be implicated in memory function (e.g. Basar *et al*., 2000; Buszáki, 2002) and, in spatial learning and sensorimotor integration (Caplan *et al.*, 2003). This initial representation implicitly acquires a *predictive* value of the auditory speech signal for two reasons. First, linguistic representations are accessible regardless of the input modality and second, the *state of activation* in the neural population underlying the visemic representation provides the *context* in which

auditory inputs will be analyzed. Note that the speech system holds an online abstract representation in storage which is at a *featural* stage, precisely that of the place of articulation (again, the involvement of theta (~250ms) at this stage is most likely). From a functional standpoint, this operation is considered 'global' or, from a speech standpoint, 'syllabic'.

At the onset of auditory inputs, an incomplete evaluation of the AV speech token is present that crucially depends on the saliency of visual inputs (i.e. a bilabial [pa] is more informative than a velar [ka]). Thus, the processing demands in the auditory domain are contingent upon how much information can be provided by the visual speech domain. It is proposed that auditory speech inputs, in this context, will be evaluated against the internal predictive value of the visually-initiated abstract representation so as to extract information missing for categorization of AV speech information. In particular, a salient visual input (e.g. bilabial) was shown to induce faster auditory processing than an ambiguous visual input (e.g. velar) (Chapter 3). This particular matching process is proposed to evolve on the gamma-scale (~25ms) in line with the implication of gamma-scale activations in perceptual object formation (e.g. Tallon-Baudry *et al*, 1999; Tallon -Baudry, 2001).

Crucially, *this operation is neither serial, nor independent from the syllabic processing stage*. It is here formulated as such to render the description of stages more explicit. From a brain's eye view however, the *analytical*stage (i.e. the gamma based processes) is simultaneous or *parallel* with the global stream described previously (i.e. the theta based computations). The emergent temporal window of integration observed here from the gamma/theta power ratio supports this hypothesis.

This model is thus an 'early' model of AV speech integration that nevertheless posits initial evaluation of visual speech based upon the natural dynamics of AV speech. Bimodal integration is based on the place-of-articulation feature, which predicts that visual speech can interact with any inputs cueing for place-articulation in the auditory domain. The sub-segmental stages in the auditory domain are preserved but the evaluation of place-of-articulation can be weighted by either acoustic or visual inputs prior to phonologic categorization. Further studies are however essential to permit a better specification of the *nature* of representations provided by visual speech inputs. In particular, as suggested by Summerfield (1987), voicing and manner cues that could be crucial for the processing of running speech may be provided by visual speech inputs.

In line with prior findings (Callan *et al.*,2001; Chapter 3) and with the present report, two frequency ranges are crucial in the 'analysis-by synthesis' of AV speech. In this view, the ~250ms time constraint is an emergent and non-linear temporal property of the integration process. The notion that on-line storage is a natural property of neural processing was suggested early on by Craik & Lockheart (1972), and has recently been hypothesized in neurophysiology (e.g. Colombo *et al*., 1996). This notion is fundamental because it suggests that the integrative time window naturally emerges from the processing of AV speech integration and is accessible and available to 'awareness' as a temporal percept independent of the speech percept. In particular, this model predicts that the characteristics of the temporal window of integration (Chapter 2) will be similar regardless of the level of perceptual access (i.e.

in both ID or TOJ task). Hence, whether in the ID or in the TOJ task, the informational content of the first-input modality determines the profile of the temporal window of integration.

The gamma range has been proposed to underlie featural binding in the formation of a percept (Tallon-Baudry *et al*, 1999). Interestingly, th e sustained gamma observed in Callan *et al.* study (2001) spanned over a ~150ms time window and our results suggest a coupling of theta and gamma. Note also that the sustained gamma activity in AV speech contrasts with the transient burst of gamma activity observed in non-speech stimulation (e.g. Bhattacharya, 2002). This sustained activation is in line with the proposed model in that both processes are present simultaneously in the speech system, an observation that can also be predicted from the AST framework (Poeppel, 2003).

Hemispheric specialization and time perception

Studies that have focused on a possible hemispheric lateralization in AV speech integration converge toward the contribution of both hemispheres (Campbell *et al.*, 1990; Soroker *et al*., 1995) but also the integrity of hemispheric transfer of information (Baynes *et al.*, 1994). We report here a trend for a higher gamma/theta power ratio in the left hemisphere and an increased hemispheric differentiation in AV speech pairs within the temporal window of integration. In particular, it is in the right-hemisphere (RH) that a systematic variation of the gamma/theta ratio was observed.

This dynamic suggests that the RH may mediate the binding of AV speech information on a syllabic time-scale.

A growing body of evidence shows that both hemispheres contribute to the processing of speech. For instance, recent MEG studies suggest that place-of-articulation and segmental information may be mediated by the right-hemisphere (Gage *et al*, 2002; Hertrich *et al.*, 2002). These new findings also support earlier evidence obtained with EEG (Segalowitz & Cohen, 1989, Molfese *et al*, 1985).

The processing of visual speech information is naturally constrained by the articulatory movements provided by the face. If the right-hemisphere mediates the processing of place-of-articulation, the processing of visual speech information may involve motor re-enactment in agreement with a growing body of evidence on the perception of intentions in action (Blackemore & Decety, 2001; Miall, 2003). fMRI studies of AV and V speech report not only STS or STG activation but also activation of the Inferior Frontal Gyrus (IFG) and recent fMRI studies further suggest that this network may underlie a motor-based representation of visual speech (Callan *et al.,* (in press); Skipper *et al*, 2003). Although our results do not provide anatomical substrates, prior fMRI research on AV speech desynchronization showed right-hemispheric activation of the IFG (Calvert *et al*, 2000). The 'analysis -by synthesis' of AV speech integration (Figure 4) could be implemented via a forward model implicating 'mirror neurons' in the processing of visual speech inputs (Miall, 2003).

Under the assumption that the temporal relationship of AV speech events is implicitly coded by neural integrative processes, the use of different cognitive tasks has here permitted us to describe two states of activation, where either gamma or

theta dominates according to the cognitive demands of the task. In AV speech, the temporal resolution of the system is syllable-based (Chapter 2) and the temporal window of integration of ~250ms observed here in the TOJ task naturally emerges from the decrease in the gamma/theta power ratio (diminished gamma over higher theta) in the RH. This result is consistent with (i) an articulatory-based mode of processing in the presence of visual speech and (ii) the weighting of theta versus gamma according to cognitive demands.

A revised functionality of multisensory STS in AV speech processing

The evidence for supra-additivity in STS (e.g. Calvert, 1999, 2000) is compatible with a forward model of AV speech integration. Two working hypotheses are here developed that could guide future experimentation.

First STS is not *only* a multisensory site of convergence but also a large cortical area that includes functionally distinct sub-regions. The anatomical separation of these sub-regions is not easily studied with fMRI and relies on the degree of resolution used in analysis, in particular in the image filtering process. Anatomical studies show that STS surrounds the auditory cortex (Pandya, 1995; Hackett *et al.*, 1998) and it is therefore difficult to differentiate activation from STG or MTG from surrounding STS. Neurophysiological studies further show that visual-only cells are readily found in this area. In fact, this region stands at the convergence of two major visual functional pathways and enhanced responses in the anterior Superior Temporal Polysensory area of monkeys have been recorded in visual cells (i.e. *not* multisensory

cells) that are proposed to derive form from motion (Oram & Perett, 1996). This region is also implicated in biological motion processing, for which cells specific to mouth and eye movements have been described (Puce *et al*, 1998). Evident ly, visual speech is also a form of biological motion and fMRI studies of lipreading show activation in this area as well (Calvert & Campbell, 2003). Hence, the functionality of STS is multipartite and supra-additive effects recorded previously may originate from visual-only as well as multisensory processing. Further neurophysiological investigations may help to disambiguate the precise locus of activation.

Although AV speech involves multisensory stimuli, it is argued that multisensory cells lack the specificity needed to mediate AV speech integration at a featural level of representation. Rather, it is proposed that the functionality of multisensory cells in AV speech integration -and in multisensory perception in general- is to enhance signal detectability. Multisensory neurons may provide a modulatory mechanism by which the saliency of bimodal stimulation is enhanced – i.e. relevant stimulation is detected-, a proposal that remains consistent with the neurophysiology of multisensory neurons (Stein & Meredith, 1993). According to this hypothesis, multisensory neurons play a central yet non-domain specific role in maximizing information (over noise). Support for this information-theoretic approach has recently been published (Patton *et al*, 2002). Recent ne urophysiological recordings have further shown that multisensory neurons display oscillatory mechanisms (Brecht & Singer, 1998, 200; Saito & Isa, 2003), a result that extends their role as ‘units of convergence’ to *intersensory relays*. From a computational standpoint then, the output of multisensory cells is not an integrated percept; rather,

multisensory cells may help to maintain a particular mode of information processing, namely multisensory.

The cortical dynamics of AV speech integration were shown to depend on the cognitive context in which information is processed. Hemispheric differentiation increased as AV inputs became more synchronized and was maximal at synchrony (i.e. in natural conditions with visual precedence). This differentiation was observed as a gamma/theta ratio correlate. Together, these results suggest that two temporal resolutions are crucial in AV speech perception, namely the global-syllabic (~250ms) and the featural-phonemic (~25ms) scales.

FIGURE 4.1

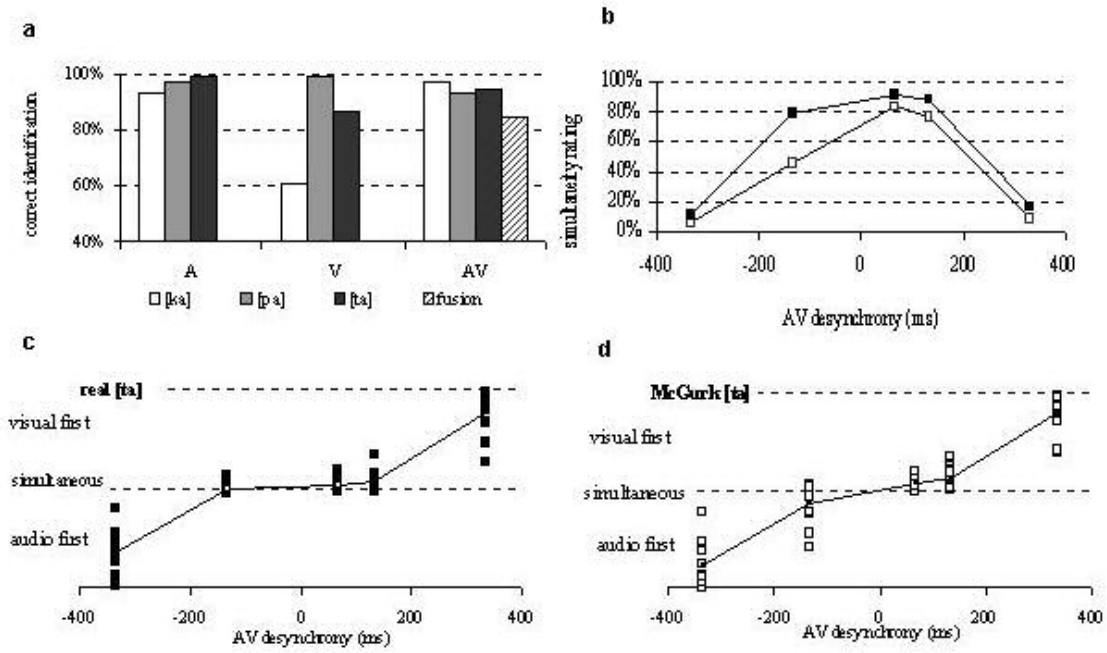


Figure 4.1: Identification and temporal order judgment of desynchronized auditory-visual speech

Panel a reports the percent correct identification obtained in Experiment 3. In particular, fusion responses in incongruent conditions were obtained ~85% of the time. Panel b, c and d report the temporal judgments obtained in Experiment 6. Filled squares represent congruent AV [ta] and open squares represent McGurk [ta] (audio [pa] dubbed onto visual [ka]). Note in particular the plateau obtained for congruent AV [ta], which is reduced in the McGurk conditions in agreement with prior findings (Chapter 2).

FIGURE 4.2

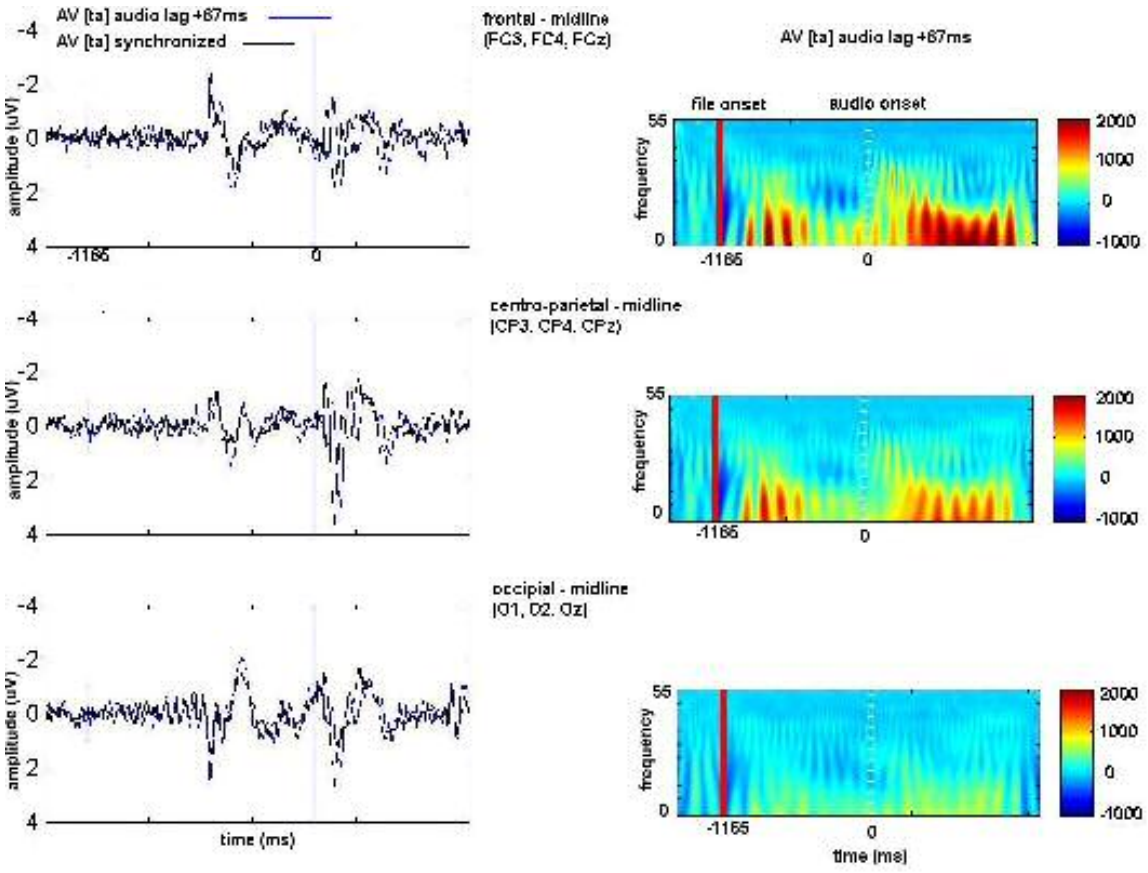


Figure 4.2: Event-related potentials and time-frequency spectrograms obtained for the presentation of synchronized and 67ms audio lag auditory-visual speech.

On the left, grand-average (N=10) data are reported collapsed across regions of interests Frontal (top), Centro-Parietal (middle) and Occipital (bottom). Potentials before 0 were elicited by face stimuli and movements of the face onsets. On the right, wavelet analysis (frequency as a function of time) quantifies the power in each frequency band of interest. Note that these spectrograms represent the data of only one participant.

FIGURE 4.3

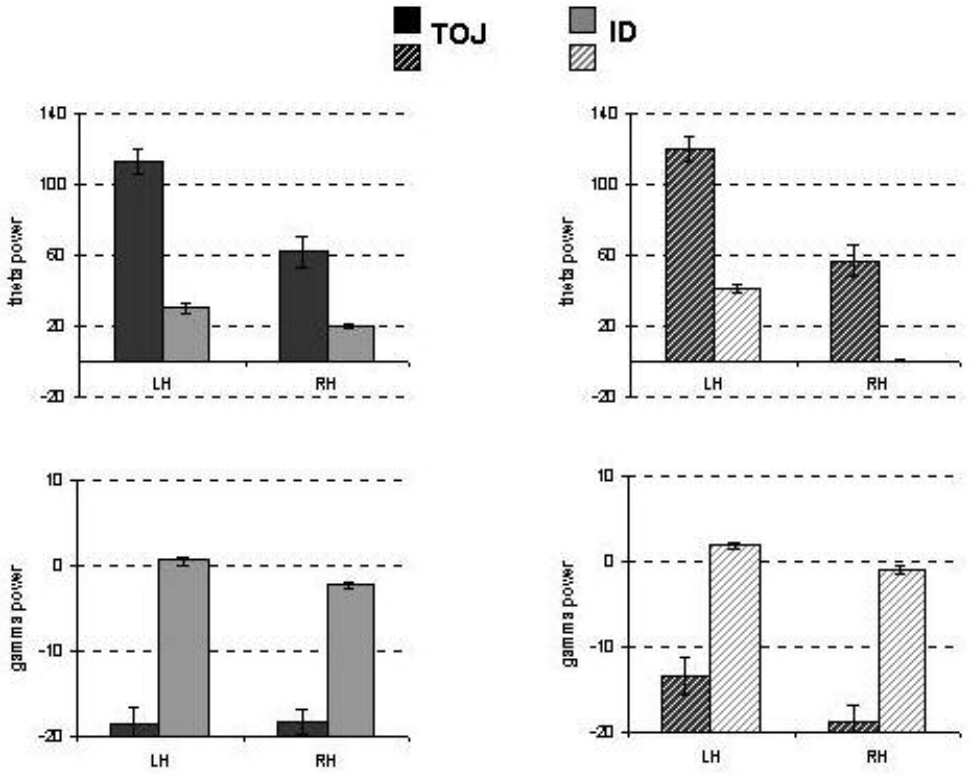


Figure 4.3: Hemispheric theta and gamma power in the identification task (ID) and in the temporal order judgment task (TOJ) for AV [ta] (left) and McGurk [ta] (right) (N = 10; [-100:+400ms] from audio onset)

In both hemispheres and in both AV pairings (i.e. AV [ta], top left or McGurk [ta], top right) a higher power in the theta band (4-7 Hz) was observed for the TOJ task (+67ms audio lead, judged 'simultaneous' more than 90% of the time) as compared to the ID task (synchronized). In contrast, the gamma range dominated in the ID task as compared to the TOJ task.

FIGURE 4.4

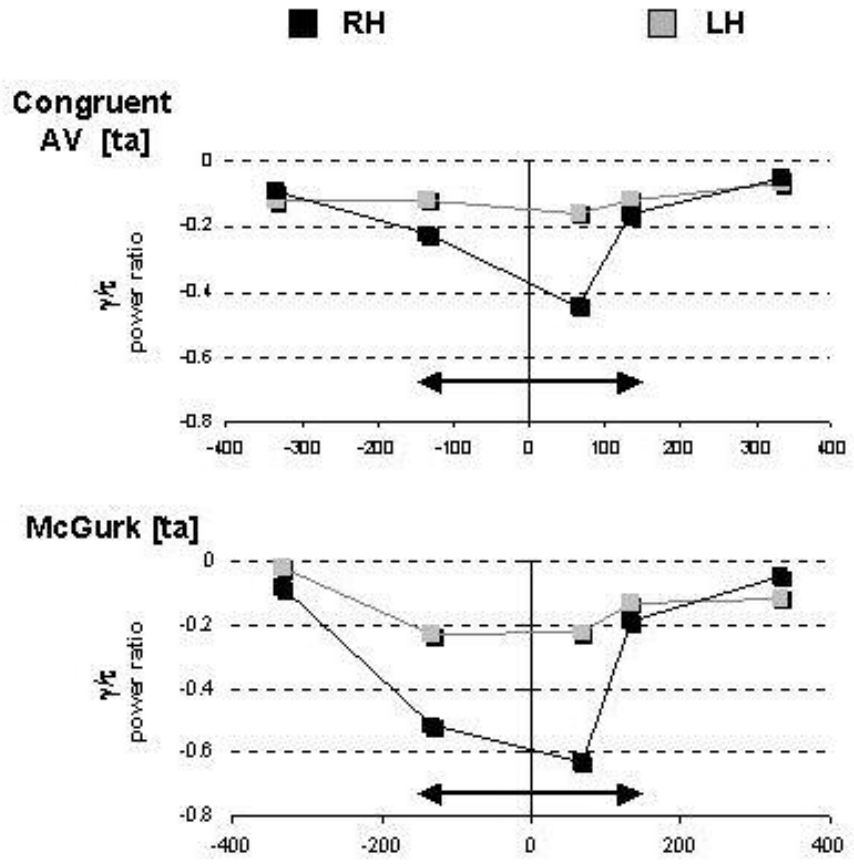


Figure 4.4: Hemispheric differentiation of the gamma/theta power ratio as a function of AV speech desynchronization. (N = 10; [-100:+400ms] from audio onset)

The gamma/theta ratio in the right hemisphere (RH, black) decreases as the AV stimuli become more synchronized (i.e. the contribution of gamma decreases and that of theta increases). This pattern of hemispheric lateralization is observed in both congruent (AV [ta], upper panel) and incongruent (McGurk [ta], lower panel) stimulation. The arrows indicate the emergent temporal window of integration as defined by the pattern of hemispheric lateralization in the gamma/theta power ratio.

FIGURE 4.5

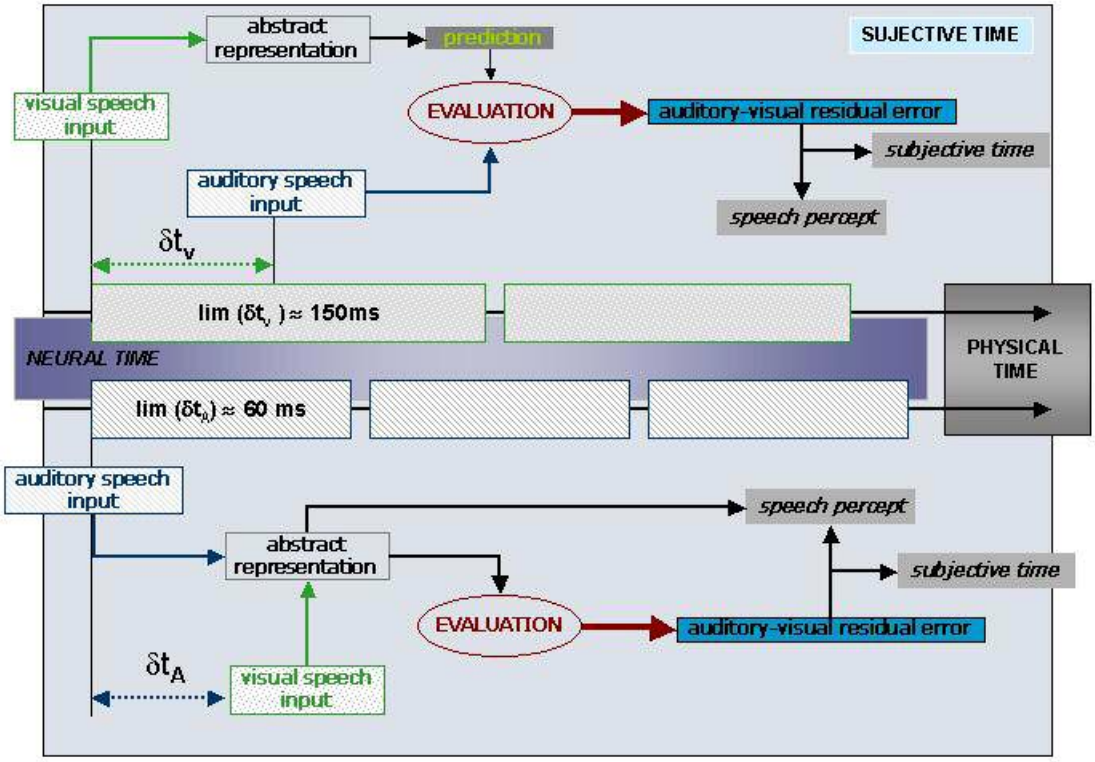


Figure 4.5: Auditory-visual speech desynchronization from a forward model viewpoint.

Precedence of visual input (in both natural and induced conditions) initiates the evaluation of visual inputs and leads to an articulatory-based representation or 'abstract representation'. In particular, it provides an internal prediction of the auditory input, the evaluation of the two resulting in the 'auditory-visual residual error'. The dynamics are constrained in the ~200ms window (see text) whether in the identification of or in the temporal evaluation of AV speech. When auditory inputs lead (artificial condition), the dynamics are constrained by the faster temporal resolution of the auditory system, and the evaluation stage is 'forced' by the temporal judgment task.

Chapter 5: Cortical dynamics of auditory-visual interactions: spectro-temporal complexity and saliency of stimulation

In previous chapters, two major levels of multisensory interactions were hypothesized to be involved in multisensory percept formation. It was suggested that a first level of auditory-visual (AV) interactions operates in a *domain general* or sensory-invariant mode, and a second in a *domain-specific* mode (here, speech). Evidence for speech-specific modulation of auditory evoked potentials in the presence of visual speech was provided in Chapter III. In Chapter IV, cortical dynamics in the theta and gamma ranges showed systematic variability in response to the presentation of desynchronized AV speech (*temporal* congruency of the AV speech tokens).

Here, I build on the hypothesis that the co-modulation of acoustic amplitude envelope and visual stimuli in the AV speech domain (e.g. Grant, 2001) may serve as a sensory invariant feature for multisensory processing. In Experiment 9, co-modulated arbitrary AV pairings were used and the perception of sound quality ('static' versus 'modulated') as a function of visual input was measured psychophysically. Participants were recorded with EEG while undertaking the psychophysical task, and a time-frequency analysis similar to that reported in Chapter IV was used for the EEG data processing.

Visual signals modulate static auditory percepts (Experiment 9)

5.1 Introduction

In multisensory research, the 'spatiotemporal coincidence' principle (Stein & Meredith, 1993) predicts that inputs from separate sensory modalities occurring in close temporal and spatial proximity will be integrated by neurons located in neural structures on which multisensory inputs converge. In AV speech, the co-modulation of the acoustic amplitude envelope with the movements of the mouth during the production of speech (Grant, 2001) provides a naturally spatio-temporal coincident input to the brain. Specifically, the source of auditory and visual signals is located around the mouth area and the modulation rate at which auditory and visual signals in natural speech are related approximates 3Hz (Grant, 2001; Grant & Greenberg, 2001). Here, auditory-visual co-modulations in the 3Hz range are hypothesized to provide a cue for spatio-temporal invariance in multisensory integration.

Space, time and the perceptual dimension of multisensory events

Changes in perceptual outcome within one sensory modality (e.g. auditory) in the presence of inputs from another modality (e.g. visual) have been reported in various contexts. For example, visual stimulation presented at the same time as an auditory stimulus but located away from it, has been shown to influence *auditory*

spatial localization (e.g. Radeau, 1974; Lewald *et al.*, 2001; Slutsky & Recanzone, 2001). In the visual domain, the temporal properties of auditory inputs bias the perception of *visual temporal rate* (Recanzone, 2003). In agreement with the 'modality precision hypothesis' (Welch &Warren, 1980), the contribution of the most *precise* modality often dominates in the registration of a multisensory event. In the McGurk effect (McGurk & MacDonald, 1976), the *quality* of the auditory percept is affected i.e. the problem is not reducible to overt *spatial* or *temporal* perception. The McGurk illusion is characterized by the integration or 'fusion' of two speech percepts (e.g. an audio [pa] and a visual [ka]) into a unified speech representation. Hence, the speech representation in one sensory domain (here, visual) affects the speech quality in another (here, auditory). In the McGurk effect, it is the perceptual dimension 'place-of-articulation' that is being modified.

Domain specificity in multisensory integration

To account for these different types of multisensory interactions, a recent review of the functional anatomy of multisensory processing (Calvert, 2001) put forward a tripartite view of the multisensory system. By analogy with the 'what' and 'where' pathways in the visual system, 'what', 'where' and 'when' (and potentially a 'novel') pathways were proposed that involve different multisensory neural populations. In particular, it was suggested that each multisensory anatomical area independently interfaces with unisensory cortices; for instance, the Superior

Colliculus (SC) interfaces the 'where' pathway with auditory and visual cortices, while the Superior Temporal Sulcus (STS) interfaces the 'what' pathway.

While a modular approach of the multisensory integrative system establishes a convenient and explicit framework to investigate the neural substrates of multisensory integration, psychophysical studies do not easily fit this dichotomy. For example, it was recently suggested that AV temporal ('when') and spatial ('where') information are interchangeable in the crossmodal association of inputs (Spence & Squire, 2003; Bertelson & Aschersleben, 2003). Another instance showing that the temporal and spatial characteristics of multisensory integration are not independent is found in a study by Radeau & Bertelson (1987), where visual biasing of auditory spatial localization was found to be stronger when visual stimulation was periodic and auditory stimuli were continuous.

Although multisensory integration affects perception in various ways, it remains unclear whether the reasons for these differences will to be found solely within the multisensory integrative system. In fact, sensory-specific and feedback processes are likely to play a major role in the integration of AV speech.

Multisensory supra-additivity

i. In arbitrary AV pairings

EEG studies using non-speech stimuli have previously used *arbitrary* pairings (e.g. a tone and circle) for which prior training was necessary (Giard & Peronnet, 1999). The analysis of evoked-potentials revealed an early supra-additive effect

(~40ms) in occipital regions, which was interpreted as originating from the recruitment of neurons in the striate cortex which are also responsive to auditory inputs (see also Molholm *et al.*, 2001). Enhanced AV interactions were also observed at ~100ms and ~170ms post-stimulus onset at temporal and central sites, originating from the Superior Colliculus (SC) and Superior Temporal Sulcus (STS), respectively (Fort *et al*., 2002).

Other EEG studies of arbitrary AV pairings have looked at a sound-induced illusion (Shams *et al*., 2000). In this illusion, two continuously moving visual disks are paired with a sound at the time where they meet. In a visual-alone condition, the percept is that of two circles passing by each other. When paired with a sound at the crossing time, the disks seem to bounce back. A follow-up study using EEG methodology showed a modulation of visual evoked potentials at ~170ms characterized by supra-additive effects in the gamma band in participants experiencing this illusory effect (Shams *et al*, 2001; Bhattacharya *et al*, 2002).

ii. In AV speech processing

fMRI studies of multisensory integration using arbitrary and natural AV pairings have reported supra-additivity in multisensory and sensory-specific cortices (e.g. Calvert, 1997; Calvert *et al.*, 1999; Fort *et al*., 2002; Beauchamp *et al.*, 2003). However, not all studies have reported supra-additive effects, and complex modulatory effects in the same cortices have also been described (e.g. Bushara *et al.*, 1999; Laurienti *et al*, 2001).

Supra-additive activation to the presentation of biological motion for visual and auditory-visual speech has also been observed in STS (Calvert, 1997; Calvert *et al.*, 2000). The term 'biological motion' is used in reference to movements performed by others and/or animals but also to their kinematics directly made available by points-of-light displays (Johansson, 1973). Regions in the STS have also shown selectivity to eye and mouth movements (Puce *et al.,* 1998). Monkey neurophysiology has suggested that a majority of multisensory cells in the STP (in particular, STPa) are visual dominant (e.g. Schroeder *et al.*, 2002) and sensitive to motion (Bruce *et al*, 1986). This region has also been proposed to contribute to the processing of socially relevant inputs (Neville *et al.*, 1998) and is considered to be at the origin of multisensory integration in the AV speech domain (Calvert, 2001).

Hence, the diversity of neural populations in the vicinity of STS renders difficult the interpretation of supra-additivity as originating only from multisensory cells. A systematic description of neural dynamics would help to elucidate the nature of neural interactions that underlie the processing of AV stimuli. It was recently argued that long-range synchronization in low-frequency ranges (theta, alpha) and high-frequency local computations (gamma band) need to be considered in percept formation (von Stein & Sarnthein, 2000). In the context of multisensory perception, the spread of activation over sensory-specific and multisensory areas further support complex cortical dynamics.

Amodal versus multisensory processing

A prominent hypothesis in multisensory integration is that amodal representations may be stored in memory only after multisensory associations have been strengthened or 'explicitly learned' (Mesulam, 1998). In using arbitrary AV pairings, it is thus the 'sensory invariance' (global spatio-temporal relationships) or the *multisensory associative* process that is being tested. Arbitrary AV pairings constitute a case of 'on-line' multisensory integration. As discussed in the introduction, this processing is proposed to predominantly involve multisensory-by convergence sites, where AV *sensory invariant redundancy* is *strengthened* while it is assumed that no modulatory effects of *perceptual redundancy* intervene (because no natural categorization of arbitrary pairing is readily available in memory).

In Experiment 9, participants were tested on their auditory judgments of a sound paired with a congruent or incongruent visual display while being recorded with EEG. A dynamic visual display was hypothesized to modulate an otherwise static auditory input on the basis of sensory invariance alone. In particular, the dynamic component of visual input encompasses both *temporal* and *spatial* aspects of stimulation, without pertaining to the localization of inputs *per se*. In contrast with the 'visual capture effect' (i.e. the classic ventriloquism effect observed in spatial mis-localization), it was predicted that the *quality* of the sound would be changed (visual stimuli will always be centered on the screen). This stimulation was designed as an attempt to separate a first level of sensory invariance that may also intervene in the

integration of AV speech. Although arbitrary pairings were used, the range of dynamic modulation corresponds to a biologically significant signal, namely speech (see Methods).

It was further hypothesized that supra-additive effects would be obtained under these conditions; in particular, supra-additive effects in the gamma band were predicted in the 'co-modulated' conditions (contrary to AV speech findings discussed in Chapter II).

5.2 Materials and Methods

Participants

Thirteen participants (5 males, mean age = 20.8 years old) were recruited from the University of Maryland population and took part in this experiment. No participant had diagnosed hearing problems: all had normal or corrected-to-normal vision and were right-handed. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli and Procedure

In the auditory modality, dynamic stimuli consisted of white noise amplitude modulated at 2Hz. In the visual domain, a modulated 'noisy' circle (V_m) was created.

The contour of the circle consisted of a pixelized white line on a black background. A 2 Hz modulation consisted of the circle shrinking down and expanding back to its initial size (a 2Hz modulation was chosen to avoid explicit phase-locking in the theta band).

Congruent AV pairings consisted of co-modulated AV noise (A_mV_m) (where peak amplitude in the auditory domain was matched with the largest radius contour on a frame basis) and congruent static AV noise (A_sV_s). Static stimuli consisted of identical duration (~ 500ms) non-modulated white noise (the same that was used for the modulated condition) and a static circle (the largest radius contour) in the auditory and visual modalities, respectively. Incongruent stimulation consisted of a static audio noise and a modulated visual stimulus (A_sV_m) and modulated auditory noise paired with a static circle (A_mV_s). Modulated audio alone (A_m) and visual alone (V_m) stimuli were also presented to ensure that each participant could detect the modulations in each modality. No training was provided.

In a 2-alternative forced choice task (2AFC), participants were asked to judge whether what they heard (or saw when visual alone was presented) was static or modulated noise while looking at the screen. They were asked to answer as quickly and accurately as possible and not to close their eyes.

Electroencephalographic recordings

EEG recordings were made using a Neuroscan system (Neurosoft Systems, Acquire 4.2b), using 32 Ag/AgCl sintered electrodes mounted on an elastic cap

(Electrocap, 10-20 montage) individually chosen according to head size (distance inion-nasion and circumference). Data were continuously collected in AC mode at a sampling rate of 1kHz. The pre-amplifier gain was 150 and the amplifier gain was set to 1000 for all thirty-two recording channels. Reference electrodes were linked left - right mastoids grounded to AFz. A band-pass filter from 1Hz to 100Hz was applied online. Two electrodes monitored horizontal eye movements (HEOG) and two others recorded vertical eye movements (VEOG) for off-line data artifact reduction and rejection.

EEG data processing

After artifact rejection (extraction of noisy data originating from non-ocular movements) and ocular artifact reduction (linear detrending based upon a blink template), epochs were baseline corrected on the basis of a pre-stimulus interval of 100 ms chosen prior to stimulus onset. A threshold of ± 100µV was used to reject residual artifacts. Trials were sorted by stimulus and response.

Time-frequency analysis

Epoched EEG data were transferred into Matlab using EEGlab (Salk Institute). Matlab routines were created for the remaining analyses. Regions of interests comprising three to five electrodes were defined in order to increase the signal to noise ratio. The five regions of interest (ROI) were: frontal (F, which

includes electrodes FC3, FC4 and FCz), occipital (O, which includes O1, O2, Oz), centro-parietal (CP, which includes CP3, CP4, and CPz), right hemisphere (RH, which includes P8, TP8, T8, FT8, and F8) and left hemisphere (LH, which includes P7, TP7, T7, FT7, and F7). A Morlet wavelet transform was separately applied for each ROI on a single-trial basis and baseline corrected on a 100ms pre-stimulus interval. Wavelet coefficients were then averaged across trials for each stimulus for each participant. Time-frequency spectra were produced for each stimulus condition and each individual.

Statistical analysis

Frequency bands of interest (FBI) were defined as theta (θ, 4-7Hz), alpha1 (α_1, 8-10Hz), alpha 2(α_2, 10-12Hz), alpha (α, 8-12Hz), beta1 (β_1, 13-18 Hz), beta2 (β_2, 18-25Hz) and gamma (γ, 25-55Hz). Wavelet coefficients comprised within the defined frequency bands were averaged by ROI, by stimulus, and by individual. Non-overlapping windows of 25ms were then defined to provide a power estimate for each FBI. The data were then exported for statistical analysis in SPSS (SPSS Inc., Illinois).

5.3 Results

Performance

Figure 5.1.1 reports the average performance for all stimuli. Two groups of participants could be distinguished on the basis of their performance in incongruent conditions. All participants performed accurately in auditory and visual alone conditions and detected the modulation except for 1 participant (male) who was set aside. Hence twelve participants were considered. In incongruent AV presentations, a performance (i.e. auditory-based correct identification) decrease was observed for all participants.

A majority of participants (9 participants, Figure 5.1.1, panel a) showed a small but significant visual bias in the A_sV_m and in the A_mV_s conditions. A one-way analysis of variance with performance as the dependent variable, showed a main effect of stimulus ($F(5, 48) = 460.013$, $p \leq 0.0001$). Paired t tests showed a marginally significant influence of static visual inputs with modulated audio signals (A_mV_s) when compared to A_m alone ($p \leq 0.04$) and co-modulated A_mV_m ($p \leq 0.04$); modulated visual signals significantly biased static audio input (A_sV_m) as compared to static visual signals (A_sV_s) ($p \leq 0.008$).

In contrast, the remaining 3 participants (Figure 5.1.1, panel b) showed a visual bias in the A_sV_m ($p \leq 0.005$ and $p \leq 0.0003$) and in the A_mV_s ($p \leq 0.0007$). One possibility is that in this second group, whenever visual information was presented participants responded to visual inputs rather than auditory inputs. Because of the signal-to-noise ratio differences in comparing 9 subjects to 3 subjects, cross-groups comparison could not be pursued in the EEG data and analysis was focused on the largest sample (9 participants). Three subjects could also not provide a sufficient signal-to-noise ratio in the EEG recordings for robust analysis.

EEG results – Glbal Field Potentials

Figure 5.2 reports the global field potentials (GFP) (i.e. all electrodes) and their distribution observed over the scalp for each stimulus condition.

Panel (a) reports results obtained in the A_m condition showing strongest field potentials at 41ms, 162 ms and 282 ms post-stimulus onset and originating mostly from a central location. In the V_m condition (panel (b)), the GFP was strongest at 61ms and 180ms post-stimulus onset showing a bilateral location over the occipital channels. In congruent dynamics conditions (A_mV_m, panel (c)) the GFP was strongest at 160-180ms post-stimulus onset over both central and occipital recording sites. In incongruent conditions (A_mV_s, panel (d) and (A_sV_m, panel (e)) and static conditions (panel (f)) two strong GFP were observed at ~ 150ms and ~280ms, again over central and occipital channels.

EEG results –Time-frequency analysis

The power in each frequency band of interest (FBI) and for each region of interest (ROI) was analyzed. Non-overlapping time bins of 50 ms were created for statistical analysis.

A four-way repeated measure analysis of variance with factors of FBI (5: theta, alpha, beta1, beta2, and gamma), ROI (5: occipital, centro-parietal, frontal, left hemisphere, and right hemisphere), stimulus condition (6) and time window (11: including pre-stimulus interval (i.e. baseline 100 ms prior to stimulus onset) to 550ms

post-stimulus onset) was performed. A significant main effect of FBI was found ($F(18.262, 12.351) = 1.544$, $p \leq 0.0001$) and an analysis by frequency band was then pursued.

A three-way repeated measures analysis of variance with factors of ROI (5: occipital, centro-parietal, frontal, left hemisphere, and right hemisphere), stimulus condition (6) and time window (11: including pre-stimulus interval (i.e. baseline 100 ms prior to stimulus onset) to 550ms post-stimulus onset) was performed with FBI as the dependent variable.

i. Theta band (4-7Hz)

A main effect of ROI ($F(17.43, 12.808) = 1.601$, $p \leq 0.0001$) and a two-way interaction of ROI with time window ($F(8.383, 22.772) = 2.846$, $p \leq 0.001$) were observed in the theta band. A main effect of stimulus was also found ($F(3.47, 22.311) = 2.789$, $p \leq 0.036$). The power in the theta band was particularly dominant in bimodal conditions regardless of the congruency of stimulation.

Figure 3 reports the comparison of theta band power over time observed in the occipital region for A_m, V_m and A_mV_m. Because static signals were not presented unimodally, a direct comparison between linear sums (A + V) and their respective bimodal presentations (AV) could not be realized.

A repeated measures analysis of variance was performed with factors of ROI (5), condition (2: unimodal sum (A_m+V_m) and recorded values for A_mV_m) and time window (11) and modulated pairs as dependent variable. A significant effect of condition was found (i.e. the bimodal A_mV_m significantly differed from the linear summation of unimodal conditions) ($F(13.039, 8) = 1$, $p \leq 0.007$) and a two-way interaction between ROI and FBI was also found ($F(6.335, 21.624) = 2.703$, $p \leq 0.004$). In addition, a three-way repeated measures analysis of variance was performed with factors ROI (5), condition (2) and time window (11) with FBI as dependent variable. Effects are reported in table 1.

A *post-hoc* comparison of the sum of unimodal presentations (A_m+V_m) with the congruent (A_mV_m) condition over the occipital region showed a significant supra-additive effect in the theta band ($p<0.003$). No supra-additivity of the theta band was observed in other regions of interest.

	ROI	**Conditions**	**Time window**
Theta	F (1.593, 12.746) = 13.605 P ≤ 0 .001***	F (1.000, 8) = 15.499 P ≤ 0.004**	F (1.893, 15.142) = 12.209 p ≤ 0.001***
Alpha	F (1.348, 10.783) = 2.494 p ≤ 0.139	F (1.000, 8) = 7.322 p ≤ 0.027*	F (1.089, 8.709) =6.932 p ≤ 0.026*
Beta1	F (1.07, 8.557) = 3.187 p ≤ 0.108	F (1.000, 8) = 0.42 p ≤ 0.535	F (1.128, 9.021) = 6.034 p ≤ 0.034*
Gamma	F (1.688, 13.507) = 0.751 p ≤ 0.469	F (1.000, 8) = 2.447 p ≤ 0.156	F (2.305, 18.437) = 6.57 p ≤ 0.005***

Table 5.1: Significant supra-additive interactions of ROI, condition and time windows

ii. Alpha (8-12Hz) and lower beta (beta1, 14-18Hz) bands

In the alpha, beta1 and gamma ranges, marginal effects of time window were obtained over all regions of interest ($F(1.096, 8.765) = 7.333$, $p \leq 0.023$, $F(1.114, 8.911) = 6.58$, $p \leq 0.028$, and $F(1.598, 12.785) = 7.031$, $p \leq 0.012$, respectively). Figure 4 compares the alpha and lower beta1 power in the left and right hemispheric regions for A_m, V_m and A_mV_m. Differences in the duration of power increase over time were most readily observable in the right hemisphere when comparing unimodal versus bimodal conditions. In particular, in the alpha and beta1 bands, the effects of time were observed as a shortening of the period of increased power in bimodal conditions.

iii. Supra-additivity in gamma band (25-55Hz)

Significant enhanced gamma responses were observed in all but the left-hemisphere region over time ($F(2.305, 18.437) = 6.57$, $p \leq 0.005$). A *post-hoc* comparison of the sum of unimodal presentations (A_m+V_m) with the congruent (A_mV_m) condition showed a significant supra-additive effect in the gamma band over the right-hemisphere ($p \leq 0.004$), the occipital region ($p \leq 0.02$) and the frontal region ($p \leq 0.00003$). The supra-additive effect was observed in the first 50ms following the auditory onset in the right-hemisphere, followed by the occipital region at 50-150ms and frontal region after 100ms. This effect was sustained in the right-hemisphere and frontal region but more localized in the occipital region (i.e. ‘burst of gamma’).

Figure 5 reports the comparison of gamma power over ROI. No significant effect of stimulus was however obtained.

5.4 Discussion

The data presented here show that a 2Hz modulated arbitrary visual input can affect the perceptual judgment of auditory static noise. This biasing effect was accompanied by specific EEG results in various frequency bands and at different regions over the scalp. First, all AV conditions were characterized by an enhanced theta (4-7 Hz) power over the occipital region locked to the duration of the stimulation. This *theta enhancement was supra-additive for co-modulated AV signals*. Second, alpha (8-12 Hz) and beta1 (12-18Hz) frequency ranges showed a narrower spread of activation over time in AV conditions when compared to audio alone or visual alone conditions. Third, supra-additivity in the gamma band (25-55Hz) was observed for co-modulated signals over occipital, frontal and right-hemispheric regions. The supra-additivity observed in the theta and gamma ranges suggest the involvement of multisensory sites, while alpha and beta1 effects suggest more sensory-specific modulations. Together, these results are interpreted in the context of a global network of multisensory processing.

Biasing effect

Following the notion that the most *precise* modality will be weighed more in the registration of a multisensory event, the 'modality precision hypothesis' (Welch & Warren, 1980) was proposed to account for multisensory biases in discrepant stimulus presentations. This hypothesis has been extended to account for various multisensory biases. In particular, ventriloquism effects are accounted for by invoking the better spatial resolution of the visual system for localization over the auditory system. Conversely, the better temporal resolution of the auditory system accounts for temporal biases in the visual domain (e.g. Recanzone, 2003). However, AV interactions in the spatio-temporal domain show a possible trade-off (e.g. Bertelson & Aschersleben, 2003) and the modality-specific contribution becomes difficult to determine when solely based on the notion of best (spatio-temporal) resolving power in each modality. The present findings further illustrate a case where the spatio-temporal dynamics of visual stimuli can affect the perceptual quality of the auditory percept.

In a study involving pairings of static tones and circles, Giard & Peronnet (1999) found an inter-individual difference in reaction times that correlated with a different pattern of EEG activation over the non-dominant modality -e.g. auditory dominant participants showed stronger activation over occipital areas. Here, two patterns of biasing effects were found in incongruent conditions, suggesting similar inter-individual differences. However, the smaller number of visual dominant participants did not permit a valid comparison of the two samples for the EEG data.

Hence, the reported results essentially relate to the auditory dominant participants. Specific effects over the non-dominant modality (i.e. occipital region) were found in the theta band.

Multisensory mode of processing for arbitrary pairings

i. Theta supra-additivity

A supra-additive theta power in co-modulated AV conditions was found over the occipital region. Additionally, bimodal conditions showed a larger power in the theta band as compared to visual alone stimulation regardless of the congruency of stimulation. This power increase spread over ~100ms to ~400ms post-stimulation and covered much of the stimulation period.

The extent of the supra-additive effect suggests that AV interactions are maintained over the course of the stimulation period. Specifically, the enhanced theta appears time-locked to the stimulus onset and naturally defines a window over which multisensory interactions could take effect. This supra-additive effect was found significant over occipital regions, i.e. in the non-dominant modality of the participants. This result is in agreement with evoked-related potentials findings discussed previously (Giard & Peronnet, 1999).

Multisensory cells display a broad temporal tuning, and supra-additive effects are observed when multisensory inputs are separated in time by as much as 1.5

seconds (e.g. Meredith *et al.*, 1987; see Chapter I, section 1.2). The window of preferred simultaneity has however been suggested to approximate ~200ms (Meredith *et al.*, 1987). Because supra-additivity is being observed, the implication of multisensory neurons should be considered. The *enhanced* theta power suggests that multisensory neurons are part of a network operating in this frequency range. An increased power in a particular frequency band is interpreted as the synchronization of underlying neural populations (da Silva, 1991). The characteristics of the theta band observed here in power enhancement, in time locking and in location over the occipital region suggests a possible synchronization of multisensory sites with visual cortices.

However, this supra-additivity is not observed globally (i.e. it does not spread over areas covering the auditory regions). In a previous study of AV speech, it was argued that the theta power implicates a 'global' network (Chapter IV). In particular, it was found that global theta power was higher in a temporal order judgment task, where the explicit comparison of spatio-temporal properties of the stimuli was needed in contrast to an identification task, where featural processes were expected to engage in finer temporal processing. The supra-additivity observed here is in agreement with the hypothesis that explicitly directing the attention of participants to the spatio-temporal relationship of multisensory stimuli may engage a theta-based process. However, the characteristics of the theta activation for this AV pairing differ with that found in AV speech (essentially in spread of activation).

ii. Gamma band supra-additivity

The arbitrary pairing of the AV stimuli was manipulated under the co-modulation hypothesis (Grant, 2001; Grant & Greenberg, 2001). The dynamics of the stimulation bear a resemblance to naturally occurring biological motion but for *meaningless* events.

The anatomical substrates for biological motion processing are diverse; they essentially include the Lingual Gyrus (Vaina *et al.*, 2001; Servos*et al*., 2002) and the Superior Temporal Sulcus (STS) for 'socially meaningful events' (Neville *et al*, 1995; Puce *et al*, 1999; Calvert *et al*, 200). More importantly, in the context of AV stimulation, the posterior Superior Temporal Gyrus (STG) has shown activation to biological motion in visual conditions alone (Calvert, 1997; Grossman *et al*., 2000; Ludman *et al*, 2000; Vaina *et al*., 2001) and in AV speech, essentially in the right-hemisphere (Calvert *et al*., 2000). Activation of STG in AV speech studies is accompanied by supra-additivity in the STS region. STG is also one of the potential generators of the N1 (~100ms post-auditory onset) auditory-evoked potential (Näätänen & Picton, 1987) and EEG/MEG studies of AV speech (e.g. Sams *et* al, 1991; Colin *et al*, 2001; Möttonen *et al*, 2001; Chapter 3) and non -speech (Giard & Peronnet, 1999; Molholm *et al.*, 2001) suggest that the generators of N1 are modulated by visual inputs.

Together, these results are ambiguous as far as the nature of computations taking place in STS and STG. For instance, if presentation of visual alone stimuli can

elicit the activation of STS and STG, which activation characterizes 'biological motion' processing and which multisensory integration by convergence?

The early effects (observed in AV conditions as gamma supra-additivity) may, thus, originate from two possible processes: (i) an early interaction of the visual motion processing stream with auditory cortices and (ii) multisensory-integration-by-convergence in STS. STG is three-synapses away from the Heschl's gyrus and shows activation in the auditory alone condition as early as ~40ms (Yvert *et al.*, 2001). It is however uncertain how early visual motion could modulate STG. While the gamma frequency band at ~50ms post-auditory stimulation is a classic mid-latency component of the auditory cortex, the supra-additive effect may originate from multisensory modulation.

Additionally, the supra-additivity observed in the gamma band extends to three regions over time: first, in the right-hemisphere (~50ms, temporal) followed by occipital (50-100ms) and frontal (100ms) recording sites. The spread of local computations indicates that multisensory interactions intervene over time and cortical regions and again, suggests an involvement of a global network. In AV stimulation, a report of enhanced gamma band response was previously hypothesized to originate from early sensory-specific inter-connectivity in occipital regions (Bhattacharya *et al.*, 2002).

Overall, it remains unclear whether the local supra-additivity observed here (and in other studies) specifies (i) multisensory feedback, (ii) transient multisensory synchronizations with sensory-specific cortices, (iii) early intersensory-connectivity interactions or even (iv) sensory-specific local integrative mechanisms.

Saliency of stimulation

A right-hemispheric lateralization of the gamma supra-additivity is in agreement with an earlier fMRI study of multisensory saliency detection (Downar *et al.*, 2000). Frontal activation has been observed in prior studies of multisensory integration of arbitrary pairings (Giard & Peronnet, 1999; Gonzalo *et al*, 2000; Fort *et al.*, 2002) and visually induced auditory imagery (Hoshiyama *et al*, 2001).

The frontal activation suggests that perceptual tagging may involve an associative frontal network (Downar *et al*., 2000; Fuster *et al*, 2000). As was pointed out by Downar *et al.*(2000), it is in the change of sensory inputs that the saliency of stimulation emerges. In the co-modulated AV pairing, this saliency is naturally strengthened by the spatio-temporal dynamics (a strengthening possibly mediated by multisensory convergence). In non co-modulated inputs, unimodal conditions were however not presented and the sub-additivity hypothesis could not be tested.

In the context of distributed coding for multisensory percept formation, multisensory cells may participate in the weighing of sensory information (see Chapter I on multisensory cells as 'intersensory relays'). The difference of neural dynamics for arbitrary pairings as compared to results of AV speech integration could pertain to the need for the system to extract a maximal amount of information from both sensory modalities. For instance, modulation in the visual domain carries more specific content than does auditory static noise (in the A_sV_m or in the A_mV_s).

If the spread of supra-additive effects over cortical areas indicates multisensory networking (whether mediated by multisensory-by convergence sites or intersensory connectivity), what are the sensory-specific dynamics?

Narrowing the sensory-specific temporal windows of integration

In the alpha (8-12Hz) and beta1 (14-18Hz) frequency bands, the pattern of activation did not show a trend for supra- or sub-additivity in amplitude (i.e. the overall power remained similar in unimodal and bimodal conditions). Rather, it is in the width of activation time that a difference was observed. Specifically, bimodal conditions were associated with a narrowing of the window of activation as compared to unimodal conditions in the same regions and this effect was essentially observed in the right hemisphere.

The alpha frequency band is primarily associated with early modality specific processing (e.g. Dinse *et al.*, 1997; Başar, 1998) while beta1 has been recently hypothesized to be involved in sensory-specific local memory processes (Tallon-Baudry, 2001). Both frequency ranges are modulated by or involved in attention (e.g. von Stein & Sarnthein, 2000; Suffczinski *et al.*, 2001).

These frequency ranges have been less studied than the gamma range (e.g. Başar *et al*, 2000). The present observations suggest that sensory -specific processing may be preserved for perceptual completion, yet also indicate a modulation in multisensory context. It is here speculated that the trend for a narrowing of sensory-specific alpha and beta1 relates to the primacy of a multisensory mode of processing

as opposed to the unisensory mode. Whereas multisensory effects described in the theta and gamma ranges suggest a global multisensory network of activation, the effects observed in the alpha and beta 1 show a functionally different type of modulation: the duration of activation suggests a shortening of the sensory-specific processing time.

Recent AV speech findings, where latency shifts were observed in the auditory evoked potentials (Chapter II) may also involve multisensory attentional effects. Note, however, that these latency shifts were systematic –i.e. a function of informational content in the visual domain. It was noted in Experiment 5 (Chapter III), that visual attention may also impact the degree of latency shift in incongruent conditions. Further experiments may help to disambiguate the contribution of multisensory processing (redundancy) and sensory-specific informational content.

A second possible interpretation for this shortening of low-frequency power is inline with the possible involvement of mechanisms operating on low-frequency ranges in sensorimotor integration (e.g. Dinse *et al*, 1997; Ba şar *et al*, 2001). Multisensory integration is often accompanied by reaction time facilitation (e.g. Hershenson, 1962; also cf. Appendix A) and in the current study, participants were asked to press a button. It is thus possible that sensorimotor integration is realized faster via shortening of sensorimotor integration times.

In summary, this study adds to the evidence that multisensory integration encompasses many perceptual domains by showing that changes in perceptual quality can be affected by dynamic stimulation. Local supra-additive effects in the theta and gamma ranges suggest a sparse network mediated by multisensory interactions. The right-hemispheric lateralization of AV dynamic processing is in agreement with prior studies of saliency and intersensory effects. Shortening of the power increase in the alpha and beta1 frequency ranges is hypothesized to reflect multisensory attentional modulation and/or possibly shortening of the sensorimotor integration time period.

FIGURE 5.1

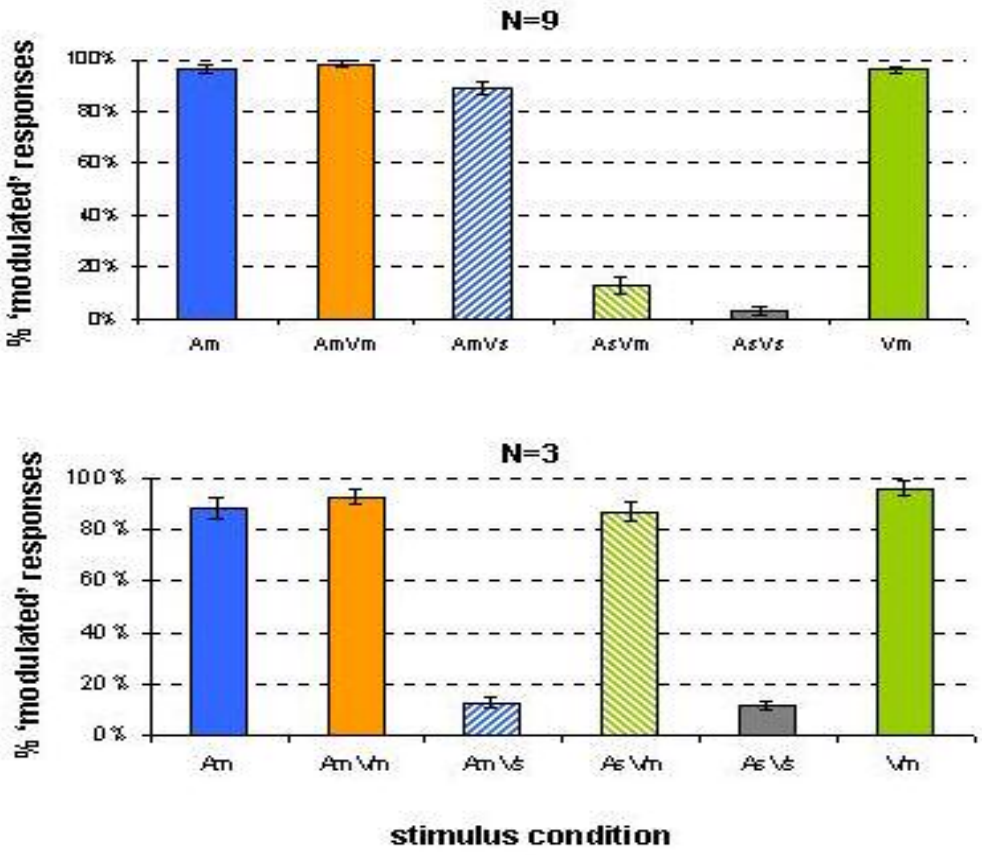


Figure 5.2: Percentage of modulated percepts as a function of stimulation

Both groups of participants (N=9 and N=3) performed accurately in unimodal conditions (Am in blue and Vm in green) and congruent bimodal conditions (AmVm in yellow and AsVs in grey). In incongruent conditions, a majority of participants (N=9) showed a significant but small increase of their 'modulated' auditory judgment to the presentation of unmodulated audio input in presence of modulated visual inputs. In AmVs condition, a significant effect was also observed. The opposite biasing pattern was observed in a minority of participants (N=3).

FIGURE 5.2

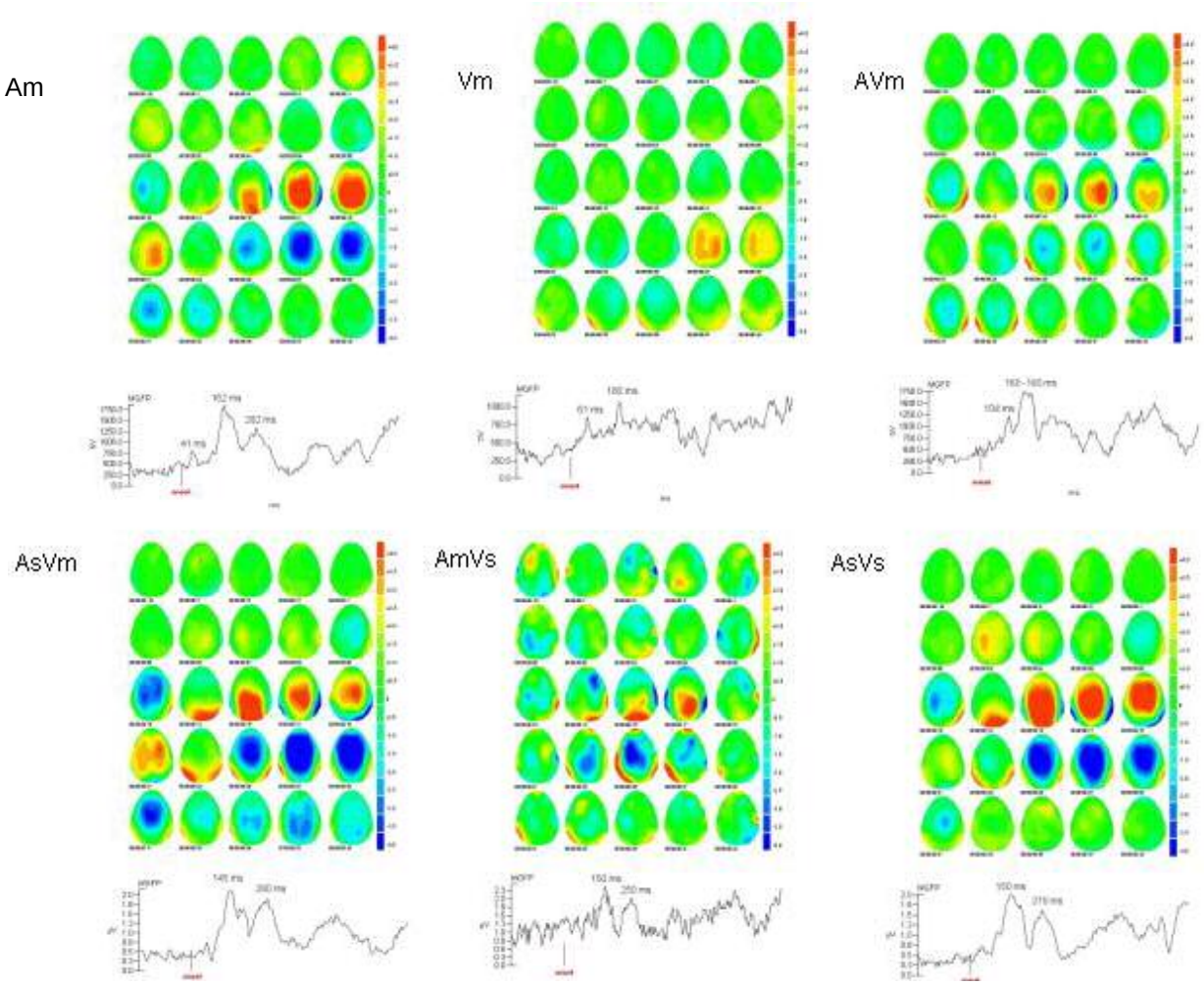


Figure 5.2: Global field power (GFP) and scalp distribution for all stimuli conditions

Red corresponds to positive GFP, blue to negative GFP. See EEG Results in text for description.

FIGURE 5.3

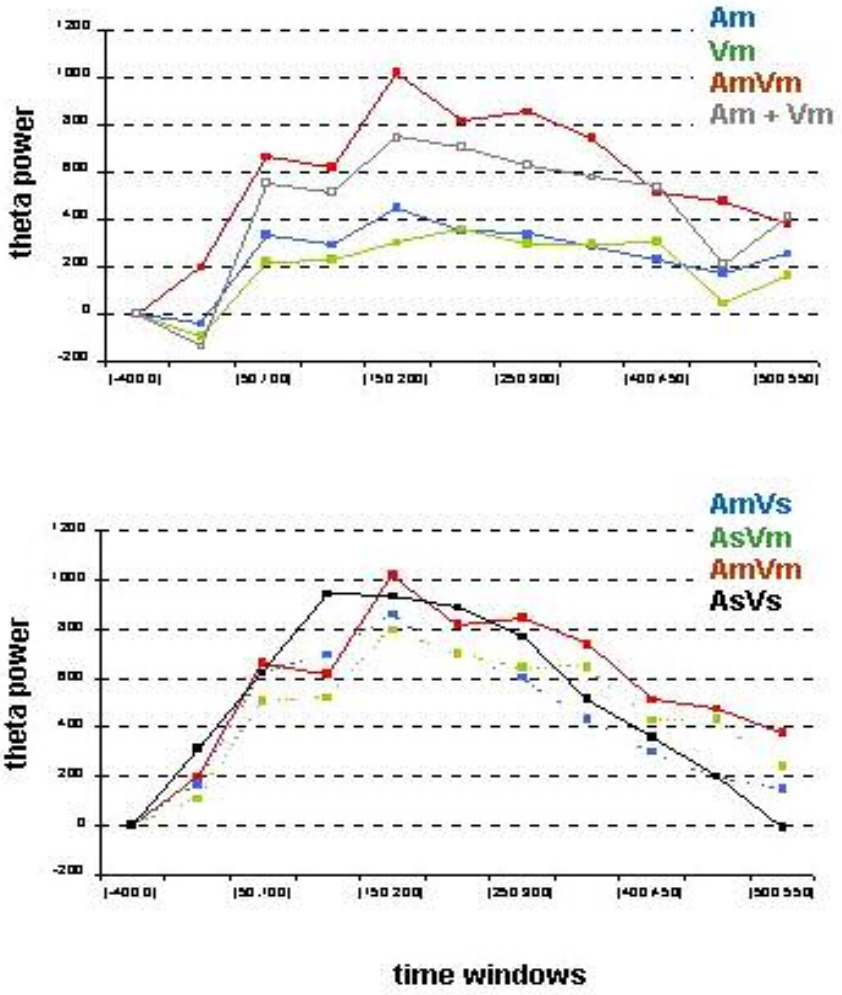


Figure 5.3:3 Occipital theta power as function of time in all conditions

In co-modulated condition (red, upper panel) a supra-additive effect in the theta power (4-7Hz) was observed spreading over the course of the stimulation (~500ms). The power increase is observed equally in incongruent conditions.

FIGURE 5.4

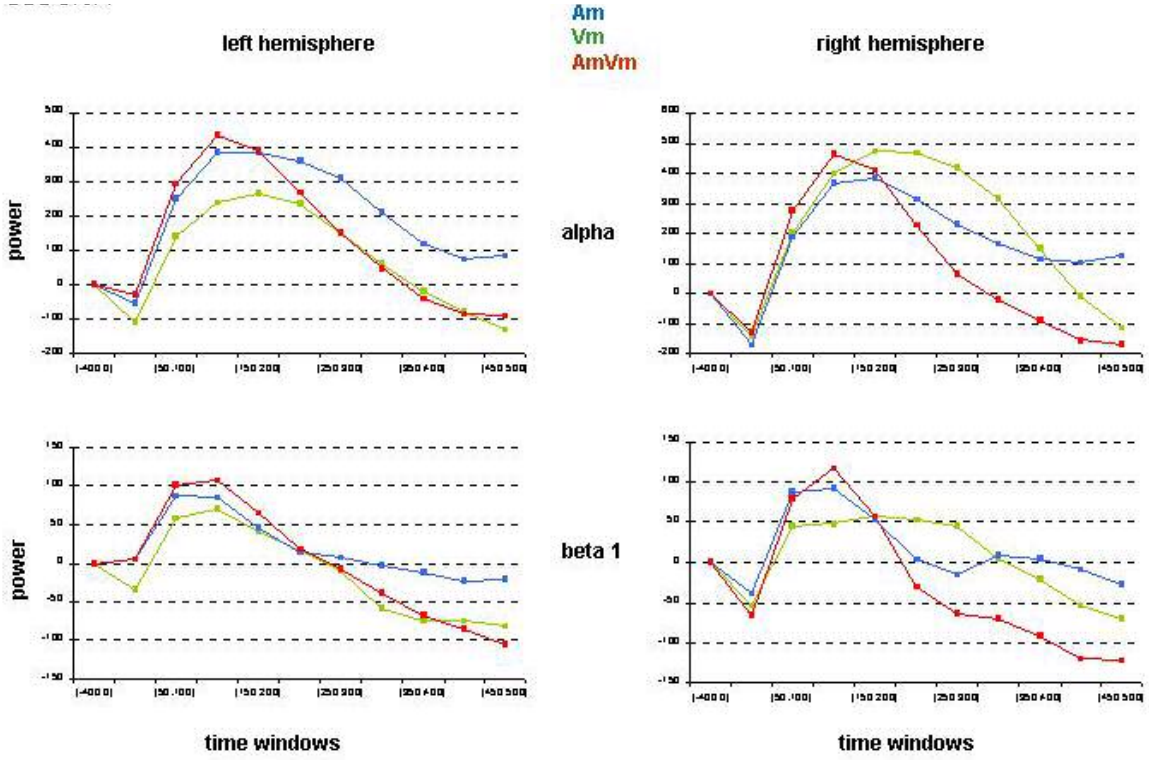


Figure 5.4: Hemispheric alpha and beta1 power as a function of time in A, V and congruent AV modulated stimuli.

The power increase in the alpha and in the beta1 band is comparable in unimodal (audio, blue and visual, green) or co-modulated (red) conditions and over both hemispheres. However, in the right hemisphere (right panel) the temporal spread of activation in the co-modulated AV condition is shorter than in unimodal conditions.

FIGURE 5.5

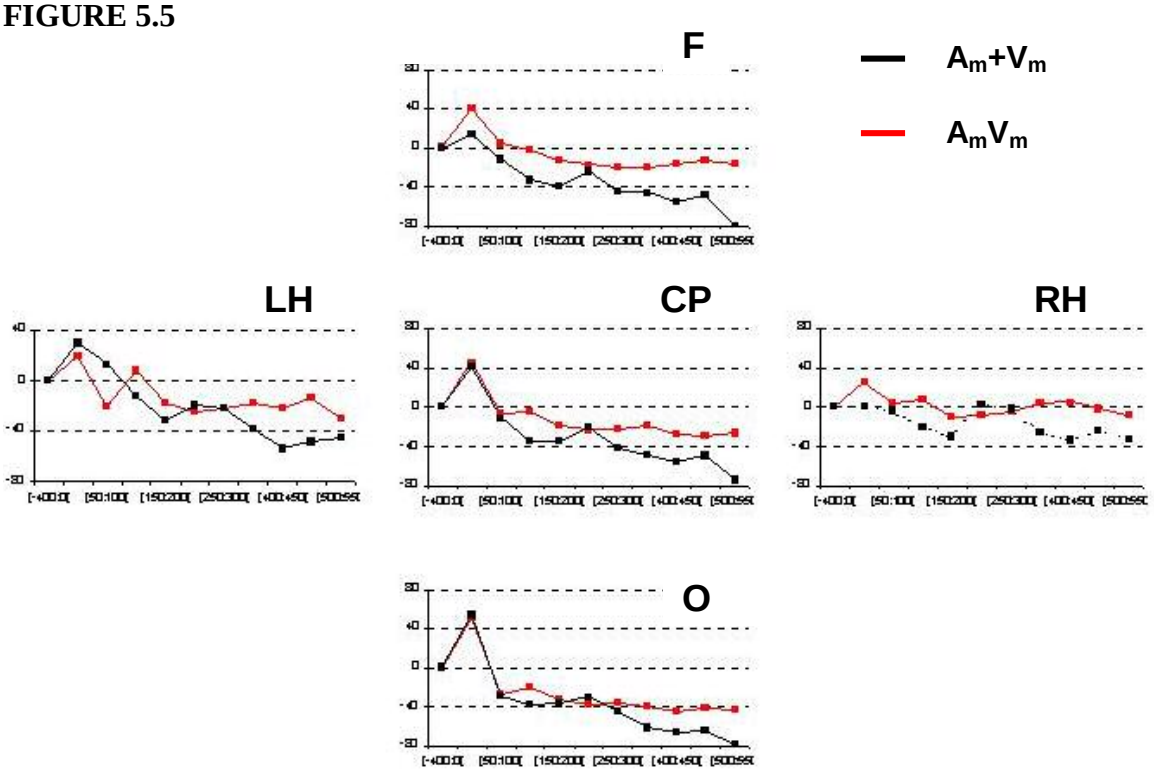


Figure 5.5: Gamma power supra-additivity in five regions of interest as a function of time.

Significant supra-additivity in the gamma band power (25-55Hz) was obtained in the right hemisphere (RH) starting at ~50 ms, in the occipital (O) region at ~100ms and in the frontal (F) region at ~100-150 ms.

Chapter 6: 'Analysis-by-Synthesis' in auditory-visual speech integration

Each of the three levels of description (computation, representation & algorithm and implementation) will have its place in the eventual understanding of perceptual information processing, and of course they are logically and causally related.

David Marr

[…] because of the Bayes' rule, the meaning of a spike train must depend on the ensemble from which sensory stimuli are drawn – we interpret what we "hear" from the neuron in light of what we expect. This is a mathematical fact that we can choose not to emphasize, but it will not go away.

Fred Rieke- David Warland

Rob de Ruyter van Steveninck - William bialek

Science is always wrong. It never solves a problem without creating ten more.

George Bernard Shaw

In the previous chapters, I have identified a few critical features of auditory-visual speech integration. Here, I turn to three major theories of speech that make explicit predictions about AV speech perception. First, I will briefly outline the theories and what issues and studies motivated their formulation. Next, I summarize the main findings from chapters II to V and subsequently analyze which of the theories can deal effectively with the data shown here. In particular, I will argue that an analysis-by-synthesis style theory provides the best basis to integrate the cognitive, computational and neural aspects of auditory-visual integration of speech.

6.1 Major speech theories and auditory-visual speech integration

There exist many theories dealing with speech, most of which follow the information-theoretic approach described in the Introductory Chapter (Chapter I, section 1.3 and Appendix B). Here, I focus on three models of speech perception that can provide explicit predictions about AV speech integration.

The Motor Theory of Speech Perception (MST)

i. Acoustic invariance

A major issue in models of speech perception was pointed out early on by Liberman *et al.*(1967) who observed that auditory speech was not invariant. More precisely, 'while the language is segmented at the phoneme level, the acoustic signal

is not' (Liberman *et al*., 1967). For instance, the directionality of formant transitions (an acoustic cue specifying place-of-articulation) is influenced by preceding and following vowels and more generally by its immediate 'context' (Liberman *et al*., 1967). Consequently, a speech segment extracted out of a longer speech stream is not reliable. Conversely, the same consonant in a speech stream can be signified by two different acoustic patterns (e.g. cue-trading relations). The contextual effects in auditory speech originate from the co-articulation in the speech production; the rapid flow of information in speech production (as many as 30 phonemes per second in a rapid speech rate (Liberman *et al.,* 1967)) creates the classic segmentation problem in the auditory domain.

Ultimately, invariant representations must be derived from varying information through a restructuring of speech inputs. This restructuring must take into account acoustic signals of varying dynamics (e.g. duration) in the frequency spectrum. For instance, transient burst of noise marking the presence of stop-consonants (e.g. [pa]) can be distinguished from the longer noise of fricatives (e.g. [sa]) while at the same time, each of these cues contribute to the same phonetic representation.

The non-invariance of acoustic inputs is crucial in understanding the limitations of current speech models. In particular, exclusive 'bottom-up' approaches necessitate that speech inputs be invariant i.e. that a one-to-one mapping exists

between acoustic signals and speech representations. Attempts to find invariant acoustic cues in speech have thus far remained ineffective (e.g. Lisker, 1985).

ii. 'The perceiver is also a speaker' (Liberman et al., 1967)

The claim that 'production and perception are two aspects of the same process' lead to the Motor Theory of Speech Perception (MST). The crucial aspect of the MST is that auditory (and presumably visual speech) inputs are 'automatically' encoded into the 'intended motor gestures' of speech production, making speech a special mode of information processing (Liberman, 1996), i.e. a subsystem or domain of its own.

From this very first description, the appealing feature of the MST was its close relationship to physiology. The original consideration of articulatory gestures as constraints for auditory speech processing made it not only plausible but also straightforwardly testable empirically. However, movements of the articulators did not bear a one-to-one mapping with speech utterances and the MST was back to being confronted with the original invariance problem observed in the acoustic speech pattern. To circumvent this issue, the MST was revised and the initial 'physical articulatory units' underlying the proposed abstract representation of speech were replaced with the 'intended articulatory gestures' (Liberman & Mattingly, 1985).

Speech as a special mode of processing engages a specific metric, namely the 'speech code' (Liberman, 1996). Speech representations are discrete and abstract, and originate from a natural propensity of the speech system in detecting acoustic /

articulatory invariance. By extension, the perception of surface articulatory gestures in visual speech (or lip-reading) should follow a similar invariant mapping process. In visual speech, the speech unit is the 'viseme', which is essentially categorized on the basis of place-of-articulation. From the MST viewpoint, auditory-visual speech integration should thus be realized in the speech mode.

The Fuzzy Logical Model of Perception (FLMP)

The dominant model in AV speech integration is the Fuzzy Logical Model of Perception (FLMP) (Massaro, 1987, 1998). The FLMP argues that the evaluation of auditory and visual speech inputs is independent - with the strong implication that AV speech integration is not obligatory.

Detection, evaluation and decision are general stages of speech processing from acoustic/phonetic features to percept formation (see Appendix B). In the initial FLMP, the detection and the evaluation stages were separate (Massaro, 1967) but eventually merged into a single 'evaluation' process (Massaro, 1996). In this stage, the speech signals are evaluated against prototypes in memory store and assigned a 'fuzzy truth value' which represents how well the input matches a prototype (measured from '0' (does not match at all) to '1' (exactly matches the prototype'). The prototypes are here defined as the units of speech or 'categories' and stem as an ensemble of 'features' and their 'conjunctions' (Massaro, 1987). Specifically, the prototypical feature represents the ideal value that an exemplar of the prototype would hold, i.e. '1' in fuzzy logic and thus intrinsically indicates the probability of a

feature to be present in the input. This same process holds true for visual speech processing.

At the integration stage for AV speech, the 0 to 1 mapping in each sensory modality permits the use of Bayesian conditional probabilities. At this stage, computations take the following form: what is the probability of the AV input being a /ba/ given the probability 0.6 of being a bilabial in the auditory domain and 0.7 in the visual? And so on, for each feature. The 'best' output is then selected based upon a goodness-of-fit determined by prior experimental datasets – through a maximum likelihood procedure. The independence of sensory modality is thus necessary to allow the combination of two estimates of POA and a compromise at the decision stage is eventually reached through adjustments of the model with behavioral data. A major criticism pertains to the fitting adjustments of the model outputs which render the FLMP too efficient (e.g. Schwartz, 2003) or unfit to the purpose (Grant, 2003?).

The nature of 'features' in the FLMP is unclear. While phonemes are clearly considered a basic unit of speech in the latest version of the model (Massaro, 1998), the phonetic analysis is not explicitly posited (partly because speech is only but one instance of the perceptual domain targeted in the original FLMP). As was mentioned earlier, 'phonetic' features can take up continuous values in the FLMP - aside maybe for formant transitions (Massaro, 1998). It follows that phonological categorization is replaced by a syllabic-like stage (and word structuring) as constrained by the more classic phonological rules: while the existence of 'phonemes' is not stated, the actual categorization rules are preserved.

‘Analysis-by synthesis’ (AS)

The ‘analysis-by synthesis’ model proposed by Halle & Stevens (1962) shares the major premise found in the MST that speech perception and speech production use common speech representations (Liberman *et al.*, 1967).

The AS, like the MST, uses abstract representations. In the MST, the restructuring of the acoustic pattern into a speech code is *passive* because the speech system inherently recognizes and analyzes speech inputs in a ‘speech mode’. In the ‘analysis-by synthesis’ model, the comparison between the internalized rules of the speech production system (articulatory based) and the perceptual stream of speech processing is an *active* process. In the AS, acoustic inputs are matched against internal abstract representations via a ‘comparator’ (Stevens, 1960); the templates, whether implicit in the MST or explicit in the AS, are motor-based and serve as metric for the discrete units of speech perception.

Specifically, AS enables a plastic and dynamic system in which the acoustic inputs follow a classic acoustic feature analysis, whose products are then compared on-line with internal speech representations. This plasticity is appealing in light of the lack of acoustic invariance mentioned earlier, and provides a testable compromise with the MST. In effect, because the coding of acoustic inputs into speech is an active process, invariance can be compensated for by the existence of trading cues, which can be matched against internal rules of the speech system. The output of this comparative system provides residual errors, which enable an active correction of the

percept (i.e. recalibrating so as to match the best fitting value) or of the production (for instance in the case of speech acquisition).

In summary, the MST and the AS share similar *discrete* speech representations, while in the FLMP prototypes are continuous. The MST and the AS mainly differ in the interfacing of acoustic-to-speech modes of processing; while the former suggests a *passive* process, the latter is explicitly *active*. I will now summarize the new findings that will serve as the basis for a more thorough discussion of the models' predictions for AV speech perception.

6.2 Summary of findings

In Chapter II, Experiments 1 and 2 characterized the temporal window of integration (TWI) for auditory-visual speech perception. Specifically, two features of the TWI were (i) a ~250ms width within which identification and simultaneity ratings of the AV speech inputs are stable and (ii) an asymmetrical profile such that preceding visual inputs are better tolerated than leading auditory inputs. This TWI was observed for voiced and voiceless, and congruent and incongruent pairs of AV speech syllables.

In Chapter III, Experiments 3 and 4 showed that visual speech inputs systematically modulated auditory-specific cortical signals. First, a temporal

facilitation of the auditory evoked potentials in AV speech condition was observed as a function of the ease of categorization in the visual domain. This temporal facilitation was robust as early as 100 ms post-stimulus onset and reached a value of ~30ms for salient visual stimuli at ~200ms post-stimulation. Second, a ~250ms amplitude reduction affected the auditory N1/P2 complex regardless of the visual speech stimuli. These observations were made in both congruent and incongruent speech. Furthermore, Experiment 5 suggested that directing attention to the visual inputs might affect the weighing of visual information in the AV speech integration process. In particular, this attentional drive appeared to affect the latency of auditory evoked-potentials but not their amplitude, further suggesting that two functionally independent processes condition AV speech integration. Experiment 6 provided further controls for the results of Experiment 3 and 4.

In Chapter IV (Experiments 7 and 8), the cortical dynamics underlying the processing of AV speech stimuli in two different cognitive contexts (identification and temporal order judgment tasks) were compared. A time-frequency analysis showed significant differences in the theta (4-7Hz) and gamma (25-55Hz) power. Precisely, the identification task was associated with a higher gamma level than the temporal order judgment task. Conversely, the temporal order judgment task was associated with a higher theta power when compared to the identification task. These results suggested that cortical dynamics underlying the processing of identical speech inputs could vary depending upon the kind of information to be extracted as a function of the task demands. In particular, a lateralization of the gamma-theta ratio

was observed as a function of AV speech inputs desynchronization and the TOJ task was associated with a right-hemispheric dominance.

In Chapter V, Experiment 9 tested the AV co-modulation hypothesis and showed a new biasing effect, where visual dynamics affected the perceptual *quality* of auditory inputs. The cortical underpinnings of this effect showed complex dynamics in major frequency bands. These effects included supra-additivity in the theta (4-7Hz) and gamma (25-55Hz) frequency ranges. Modulations of modality-specific neural dynamics were further observed in the alpha band (8-12Hz) as a shortening of power increase duration. A similar effect was observed in the lower beta frequency range (beta1, 14-18Hz). Despite a similar global dynamic features of the AV stimuli that were used in this experiment, quite different cortical dynamics were observed as compared to those of Chapter III and IV for AV speech. To further understand the interplay of cross-modal stimuli attributes, a parametric study using MEG methodology is currently being undertaken. In particular, the initial MEG study aims to characterize (dynamically and anatomically) the cortical dynamics of AV events in static, co-modulated and speech conditions.

6.3 Analysis-by-synthesis in AV speech integration

A cognitive neurosciences viewpoint

The cognitive neurosciences investigate the nature of mental events in relation to their underlying neural principles. A fundamental function of the nervous system is the transformation of a continuous (analog) flow of information into a discrete (and possibly symbolic) mode of representation -i.e. the "neural code" whose dynamics carry internal representations, the "mental events".

i. Constraints on representations

Two physical constraints in this transformation process are very general features of the nervous system, (i) its architecture (anatomical connectivity) and (ii) its dynamics (neural activity). Both components have evolved so as to specialize into separate subsystems that extract information from (sensory - e.g. auditory, visual systems) and plan upon (motor system) arrays of energy. Under the assumption that there exists one neural 'language', internal representations are to some extent *amodal* in nature. Specifically, an internal representation is only specified by the physical continuity of the signal carriers, i.e. the neural connectivity within which voltage changes are transmitted. The specificity of an internal representation is thus contingent on (i) the availability of encoded information within and across specialized subsystems as defined by the neural architecture and (ii) on the dynamical limits of the network as defined by the amount of information that can be encoded in the neural signal. In this context, the study of multisensory fusion provides a (psychophysically and neurophysiologically quantifiable) means to investigate the neural mechanisms by which internal representations lose one level of 'physical specificity' and generate abstraction.

ii. Levels of description

A fundamental contribution to Cognitive Neurosciences was made by Marr (1982), who clearly distinguished levels of (i) representation, (ii) computation and (iii) implementation. Specifically,

"A *representation* is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. And I shall call the result of using a representation to describe a given entity a *description* of the entity in that representation."

Additionally, computations are taken as the rules used in this representational system and they stand *independent* from their *implementation* –i.e. the computational rules are not defined according to the 'physical' system that realizes them.

In proposing a *forward* or *analysis-by-synthesis* model for AV speech integration, I here consider that it offers the most complete representational system from a brain's viewpoint in that it provides (i) the *descriptions* (the core of the original Halle & Stevens proposal (1962)) and (ii) the clearest prediction of *implementation* levels, while allowing for flexible (iii) computations –and Bayesian probabilities are here posited as a possible computational rule.

Accessing the (abstract) speech code

i. Multisensory access to the 'speech code'

In monkeys and nonhuman primates, vocalizations are accompanied by facial expressive gestures, which not only convey social information from their conspecifics but are also a function of environmental context (Morris, 1967; Byrne, 1995). From an evolutionary standpoint, proximal communication in primate societies naturally engages auditory-visual interactions. While detailed observations in primatology and ethology have been made, it is not until recently that vision in the communication system of primates has started to be investigated neurophysiologically (e.g. Barraclough *et al*, 2003). A recent behavioral paradigm has already shown that visual displays can affect the processing of vocalizations in primates (Ghazanfar & Logothetis, 2003).

In humans, visual speech continues to play an important role in social interactions (deGelder *et al*., 1999) but it also interfaces with the human-specific language system at various degrees of language processing, i.e. from the detection (Grant *et* al., 2000) to the intelligibility of auditory speech (MacLeod & Summerfield, 1990; Grant & Seitz, 1998). Additionally, speech representations are also influenced by tactile inputs (Blamey *et al.*, 1989). How does this 'sensory plasticity' or multisensory access to speech representations fit in the context of a supramodal speech code proposed by Liberman (1996)?

ii. (Multi-) sensorimotor interactions in speech

During development, the acquisition of speech production could undergo an imitative stage from visual speech perception to speech production. In principle, the imitative stage would permit the child to learn how to articulate a speech sound by explicitly reproducing the caretakers' and peers' facial gestures. However, two types of evidence suggest that imitation does not operate on a blank-slate system and rather that internal motor representations (or templates) for speech are available very early on.

First, the gestural repertoire is already rich only three-weeks after birth, suggesting an innate ability for the articulation of elementary speech sounds (Meltzoff & Moore, 1997). Second, auditory inputs alone suffice for the infants to reproduce accurately simple speech sounds and enable the recognition of visual speech inputs matching utterances that have only been heard (Kuhl & Meltzoff, 1982, 1984, 1996). Furthermore, during the acquisition of speech, the infant does not see his own gestures. Consequently, infants can only correct their own speech production via auditory feedback or via matching a peer's gestures (provided visually) to their own production, i.e. via proprioception (Meltzoff, 1999). Hence, early in development, the interfacing of auditory, visual, and tactile inputs with motor commands used in speech production seemingly provide an 'internal loop' and this mechanism is readily observable in infancy for speech acquisition.

Crucially, this mechanism suggests a highly active and dynamic system where, in the course of development, a child fine-tunes his gestural repertoire while

losing the natural plasticity of speech representations. This active processing of information is explicitly predicted in the AS model (Stevens, 1960) which provides an interface ('comparator') between the perception and production of speech.

Auditory speech and visual speech

Auditory-visual speech provides a unique opportunity to study multisensory processing in a well characterized representational domain and to build upon a rich theoretical and empirical framework elaborated in linguistic research in general (Chomsky, 2000) and in speech research, in particular (e.g. Chomsky & Halle, 1968; Liberman, 1996). An obvious characteristic of speech (whether auditory, visual or auditory-visual) is that it is a kind of spectro-temporally complex event that unfolds over time. In classic models of speech processing including the MST and the FLMP, processing stages are hierarchically organized such that the speech/acoustic inputs follow a serial processing from the smallest speech features (i.e. phonetic) to larger speech unit formation (e.g. phonemes and syllables). Given the critical role of temporal information in speech (Rosen, 1992; Greenberg, 1996), their lack of account in speech models is surprising (see also Chapter I, section 1.3).

i. Auditory speech - Multi-temporal resolutions in the auditory system

Auditory speech has recently benefited from extensive anatomical (e.g. Celesia, 1976; Hackett *et al*, 1998, 1999), neurophysiological (e.g. Steinschneider *et*

al., 1990, 1994, 1995a, 1995b; Rauschecker *et al*, 1995, 1998; Ehret, 1997; Heil, 1997a, 1997b) and functional neuroimaging studies (e.g. Belin *et al*, 2000; Binder, 2000), all of which have permitted detailed hypotheses on the neural implementation of speech perception (e.g. Hickock & Poeppel, 2000, 2004; Poeppel & Marantz, 2000; Poeppel, 2003; Scott and Johnsrude, 2003, Poeppel & Hickock, 2004).

The growing body of neurophysiological evidence for temporal windows of integration in the auditory system (Näätänen, 1992; Yabe *et al*, 1997; Winkler *et al*, 1998; Shinozaki *et al*, 2003) highlights the non -linear nature of auditory processing. In the context of parallel streams of information processing from periphery to cortex (e.g. Steinschneider *et al*, 1994, 1995), temporal windows of integration further confer the auditory system multiple-resolution yet simultaneous analytic streams (e.g. Viemeister & Wakefield, 1991). From a speech standpoint, the recent Asymmetric Sampling in Time hypothesis proposes that simultaneous analytical streams are functionally differentiated on the basis of their respective temporal resolutions (Poeppel, 2003). In particular, fine-grained temporal resolution can appropriately serve phonetic-based computations (in the order tens of tens milliseconds, ~25 ms), while longer temporal windows (in the order of hundreds of milliseconds, ~250 ms) may serve a syllabic-based representation.

ii. Acoustic versus speech mode in cortex: architectural dynamics

Empirically, the separation of auditory versus phonetic modes of processing is a difficult endeavor. A growing body of evidence for phonetic processing and

phonological categorization of auditory speech have been shown (Maiste *et al.*, 1995; Steinschneider *et al*, 1995; Simos *et al*, 1997; Liégeois *et al*, 1999; Sharma & Dorman, 1999; Philips *et al*, 2000). Recent findings in auditory neurosciences have suggested the existence of a voice-specific pathway in the auditory system (Belin *et al.*, 2000), which may support the speech-specific pathway predicted by the MST. More precisely, the productions of the vocal tract would be channeled early on into a speech specific neural analysis as implementation of the supramodal 'speech mode'. The FLMP implicitly assumes an amodal metric (Massaro, 1987, 1998) but does not provide an explanation as to the nature of this non-sensory specificity. From that standpoint, the FLMP is highly unspecific.

From an anatomical standpoint, the voice-selective areas in the auditory cortices (Belin *et al*., 2000) may have evolved from the classically described vocalization specific areas in monkeys (e.g. Rauscheker *et al*, 1995). However, this anatomical differentiation may be paired with an important functional step in evolution; the dynamics of the neural processing for voice-specific areas may undergo specific computations. For instance, a recent study shows that for equivalent stimulation, a voiced utterance produced a stronger response than a musical instrument production at the same cortical source (Gunji *et al*, 2003). Here, the notion of ecological validity or saliency of stimulation for the brain system is highly significant.

iii. The inherent lack of one-to-one-mapping from inputs to neural representation

Recent investigations in auditory computational neuroscience show that the peripheral auditory system in mammals could efficiently encode a broad category of natural acoustic signals by using a time-frequency representation (Lewicki, 2002). In this study, the characteristics of the auditory filters depend upon the statistical nature of sounds, i.e. the *coding scheme shows plasticity as a function of acoustic inputs.* The intrinsic dynamics properties of neurons would then allow for multiple modes of acoustic processing (trade-offs in the time and frequency domain), which naturally partition the time-frequency space into sub-regions. Importantly, such a time-frequency representation strategy is inline with the description of temporal windows of integration described earlier, and these results support the notion that *the saliency of signals can condition different modes of information extraction in neural systems.*

However, the (most accurate) statistical description of inputs to the system does not fully predict its neural dynamics. For instance, in their early attempt to formalize neural activity, McCulloch & Pitts (1943, 2000) wrote,

> "At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation."

In Chapter IV, an example of internal determinism was described, where the state-dependent activation of neural populations (or the 'neural context') may *determine* which preferred temporal resolution for information extraction. As a fundamental consequence of the intrinsic determinism of neural systems, the description of a representational system needs to consider (i) the statistical nature of the inputs in

relation to the type of information extraction they undergo and (ii) the state of the system upon presentation of the input.

In particular, a passive and unidirectional (e.g. 'bottom-up') description of information processing as it currently stands in models of auditory speech processing is incomplete. The general pattern of 'top-down' and 'bottom-up' regulations are but an intrinsic dynamic property of the nervous system when not considered a 'blank-slate', and when put in the context of a dynamic sampling of information. As they are currently described, the FLMP and the MST fit a *feed-forward* implementation (and possibly feedback at later stages): as the acoustic inputs unfold in time they undergo a hierarchical analysis (whether as continuous or discrete units). In contrast, the AS is *forward* in nature: the comparative system is a built-in constraint and, as such, constrains locally the analytical stream of information processing. In particular, these intrinsic dynamics confer the speech system with predictive capabilities, where any inputs at instant (t+1) will occur based upon (thus *depending on*) the outputs of the state of the comparative loop at instant t.

iv. Visual speech

'Visual speech,' i.e. how visual signals alone provide speech information, is at an earlier stage of investigation, and very few studies have focused on the neural bases of visual speech alone (Calvert, 1997; Calvert & Campbell, 2003).

The neuropsychology of lipreading has been described (Campbell, 1986, 1989, 1992) but often so in the context of AV speech (e.g. Summerfield, 1992, Massaro, 1996). For instance, the exact nature of visual speech information is unclear.

Visual speech is a particular form of biological motion, which readily engages face specific analysis yet remains functionally independent (Campbell, 1992). The extraction of form and motion may both contribute to its processing (Campbell, 1992; Campbell & Massaro, 1997), yet even kinematics appear a sufficient source of information to maintain a high rate of AV fusion (e.g. Rosenblum & Saldaña, 1996). For instance, in AV speech, the contribution of foveal information (i.e. explicit lip-reading with focus on the mouth area) versus extra-foveal information (e.g. global kinematics) has not been thoroughly investigated, leaving uncertain which visual pathway(s) may actually be predominantly engaged.

Two recent eye-tracking studies have started addressing this question and results have shown (i) that eye movements were highly variable and (ii) that a gaze fixation at 10 to 20 degrees away from the mouth was sufficient for AV speech integration (Vaitikiosis-Bateson *et al*, 1998; Paré *et al*, 2003). These results suggest that the sampling of information from the face in movement is a highly *active* process. It is noteworthy that changes of gaze direction can be crucial for the extraction of auditory information according to a recent study showing that the tuning properties of primary auditory neurons are a function of gaze direction (Werner-Reiss *et al.*, 2003). New insights on the neural bases of visual speech may be provided by studies of 'biological motion' (e.g. Grossman *et al*, 2000; Vaina *et al*, 2001; Servos *et al.*, 2002). In particular, the finding of mouth-movement specific cells in the temporal cortex may provide an interesting starting point for understanding the binding of AV speech information (Desimone & Gross, 1979; Puce *et al*, 1998; Karnath, 2001).

Auditory-visual speech

i. AV speech integration as an active analytic process

What sets the AS apart from other speech models is the inclusion of a 'comparator' module. This comparator operates in the 'speech mode' following internalized phonological rules and operates on discretized speech units (Halle & Stevens, 1962; Stevens, 1960).

Bearing in mind that inherent dynamics of neural systems act as constraints on the speech system, the AS instantiates a case where these dynamics show non-linearity and where the ongoing comparative mechanism (or 'switch') serves as a predictive device. In effect, and in relation with the notion of temporal windows of integration discussed earlier, the temporal overlapping in parallel processing streams may induce such non-linearity in a neural system; this non-linearity can serve as a direct implementation of the comparative mechanism (i.e. this implementation is not 'structural' or anatomical but 'functional' or dynamical, cf. Chapter IV).

The notion of 'comparator' is in fact a major theme between the FLMP and the AS, but it is *described at two different levels*. Both models argue for the existence of speech prototypes but at various levels of representation. The former is based on 'fuzzy' continuous prototypes and the latter uses categorical 'phonetic-segments'. Yet fundamentally, they postulate the existence of a set of possibilities and consequently, both necessitate comparative/integrative rules between the internal representations in memory and the acoustic inputs. Although the FLMP is presented as an 'algorithm' or a 'computational' model, it is in fact a sum of computational rules if one considers

within Marr's framework. As such the FLMP provides a formal, abstract and optimized account of AV speech integration that does not bear -nor pretend to have- any biological implementation (Massaro, 1998).

ii. Internalized perceptual categories

The psychophysical evidence for categorical perception of speech follows a long history of controversy ranging from continuous mode to categorical, with variants depending on the stage at which information loss permits categorization (e.g. for review of the arguments, Paap, 1975). Regardless of the robustness of categorical perception in speech as it is currently measured psychophysically, the general assumption that prevails in most models of perception is that *internal categories* of the world must exist in memory as prototypes of more (phonetic segments) or less specific (FLMP features) nature. Additionally in speech, these prototypes owe to be very flexible (i.e. influenced by experience) or the comparative process between inputs and prototypes owe to be very plastic so as to allow for new categorizations in the acquisition of a foreign language for instance. Again, this plasticity can only be achieved within a dynamic implementation of speech encoding.

iii. Continuous versus categorical representation: a fuzzy twist in level of description?

In the auditory domain, if the parsing problem is 'solved' by the inherent dynamic properties of the auditory system through temporal windows of integration,

the resulting product (auditory object) is intrinsically defined within that temporal resolution. From a computational standpoint, how the system assigns a speech value to the inputs remains unsolved.

In the FLMP, the comparison is rendered continuous -from a theoretical or psychological stance- and unless this continuity is assigned to the discretized acoustic inputs, the model cannot be straightforwardly implemented in the neural domain. Here, I would like to stress that in the FLMP, (i) the apparent 'amodal' nature of the features, (ii) the close resemblance of feature space with phonetic space and (ii) the preservation of phonological rules are puzzling with Massaro's interpretation of continuous representations or continuous perception (e.g. Massaro, 1987).

Categorical perception was reassessed in this model by showing that speech inputs can be perceived on a continuum if categories along this continuum are made available to the listener; these results are interpreted as evidence for the assignment of continuous values to features in the FLMP (Massaro, 1996). However, this approach is misleading. Because of its formalism (i.e. the use of Bayesian probabilities), the FLMP is not incompatible with the discretization of inputs (e.g. as posited at the level of description (algorithm) in the AS). The goodness-of-fit in the matching of speech inputs with prototypes stands as a level of computation that is *independent* from the representational status of speech elements. Specifically, if the computational rules that permit the *description* of the representations can be conceivably continuous (and in fact could be implemented neurally), this does not entail that the *representations* themselves need to be continuous.

For instance in the AS, the auditory inputs (after a 'preliminary analysis' resulting in a spectral characterization of the inputs) are matched against the internal articulatory rules that would be used to produce the utterance (Halle & Stevens, 1962). Such rules, from the speech production side, can take upon continuous values. The rationale for a continuous parameterization at this stage is as follows: the set of commands in speech production change as a function of time but "a given articulatory configuration may not be reached before the motion toward the next must be initiated". Even though the rule uses a continuous evaluation of the parameters, the units of speech remain discrete and articulatory based. By analogy with the overlap of articulatory commands, the auditory speech inputs contain the traces of preceding and following context (co-articulation effects). Hence, the continuous assignment of values need not bear a one-to-one relationship with the original input signals; again, overlapping streams of information extraction (via temporal window of integration) may enable this process.

The point here is that the phonological rules exploited by the FLMP are not incompatible at this stage of the AS. From a computational standpoint, the use of Bayesian conditional probabilities at the level of comparison between auditory speech inputs and internalized articulatory commands of the system may enable a constrained implementation than is not currently offered by the FLMP.

In AV speech, the 'predictive value' taken by the visually-initiated abstract speech representation in the visual domain would correspond to the *initial state of the comparator* upon arrival of the audio inputs. Hence, visual speech provides the

context in which auditory inputs are analyzed and the degree to which the abstract representation is specified prior to auditory inputs condition the evaluation stage. The results described in chapter III suggest that visual speech inputs may drive this comparative system, for example with POA information. In particular, the natural precedence of visual speech over auditory speech, and its saliency serve as fundamental parameters in the proposed model.

iv. An example of AV speech cue in the integration of place of articulation information

From an information-theoretic perspective, an opposite trend is observable in the auditory and visual processing of place-of-articulation: auditory place-of-articulation (POA) is most susceptible to noise perturbations while the visual is least (Summerfield, 1987). This potential bimodal speech cue trading is desirable at a pre-phonetic stage of AV integration (e.g. Green, HBE II).

In an acoustic/speech selective adaptation paradigm, the integration of AV speech was suggested to occur very early in the speech process, i.e. prior to phonetic evaluation and to evolve in the 'acoustic' rather than in the 'speech mode' (Roberts & Summerfield, 1981). Green & Norrix (1997) showed that the dynamics contained in the formant transitions were crucial for the McGurk effect. Three acoustic cues for place-of-articulation were manipulated (the release burst, the aspiration and the voiced formant transitions) were shown to influence differentially the magnitude of the McGurk effect. The lack of specific invariance for POA in the auditory domain suggests that each type of manipulated cue provides various levels of auditory

saliency for the evaluation of POA representation; visual speech shows similar degree of saliency for POA (e.g. clear bilabials versus ambiguous velar).

Hence, in the integration of AV speech, the channel providing the most reliable information may also be most influential for the AV representation of POA. However, the stage at which the acoustics and visual inputs integrate is unclear.

The desynchronization of inputs pertaining to POA in stop consonants results in a similar integrative time constant whether in the auditory or in the AV domains. Repp & Bentin (1984) looked at the effect of input desynchronization in auditory fusion by using the 'duplex perception' paradigm (e.g. Mann & Liberman, 1983) where an F3 formant transition ('chirp' like sounds in isolation) is separated from the base (resulting in a speech percept when paired with the F3 formant transition). Spectro-temporal fusion tolerated surprisingly well desynchronizations of as much as ~150ms – i.e. performance did not significantly differ within that range. Furthermore, this temporal integration window in diotic conditions, showed an asymmetric profile such that the precedence of the F3 formant was better tolerated than that of the base (for as much as ~70-100 ms). In AV speech integration, a similar window of temporal integration -in both length and asymmetry- was observed, such that visual leads favored AV speech integration (Munhall, 1996; vvw *et al*, 2001; Grant *et al*.).

The integration of information pertaining to POA seem to operate on a ~150-200ms time-scale regardless of the input modality which suggests the existence of a functionally independent module for POA processing. This hypothesis would be most advantageous for the implementation of a comparative or predictive mechanism

interfacing audio and visual channels. However, recent findings using frequency slitting of speech inputs (sentences) -as opposed to synthetic formant versus base of syllables- have shown a better temporal resolution under diotic presentation (Grant *et al.*, (submitted)). In light of recent modeling findings (Lewicki, 2002), the auditory processing modes in the Repp & Bentin (1984) and Grant *et al.*(submitted) may differ. Studies of duplex perception (synthetic parameters) specifically isolate a speech feature from the base, thus 'forcing' a particular mode of acoustic processing more similar to a time-frequency 'tiling' (observed for vocalizations and speech processing); in the contrary, the spectral slitting of natural speech provides information best handled by the generalist dynamics of acoustic processing (see frequency space representation for Fourier Transform versus Wavelet representation in Fig. 6, Lewicki (2002)). This difference pertains to the time-frequency trade-off. It is also noteworthy that in the duplex perception paradigm a syllabic scale stimulus was used whereas in the spectral slitting experiment sentences-long segments were tested that may allow for a reduction of uncertainty on a longer time scale, allowing for a better temporal resolution locally.

In AV speech, the movements of the lip area and the amplitude of the acoustic spectrum domain might correlate (Summerfield, 1992). Recently, Grant & Greenberg (2001) provided evidence for the existence of spectral AV co-modulation in the F2-F3 formant frequency region. In an AV speech detection paradigm, a higher bimodal coherence masking protection amplitude was also observed in the F2 region (Grant,). These results are in line with a pre-phonetic integration of AV speech, but remain

ambiguous as far as the nature of representation (i.e. again acoustic versus phonetic mode). Figure 6.1 provides a schematic version of the processing stages of the proposed AS model of AV speech integration at different levels of representations.

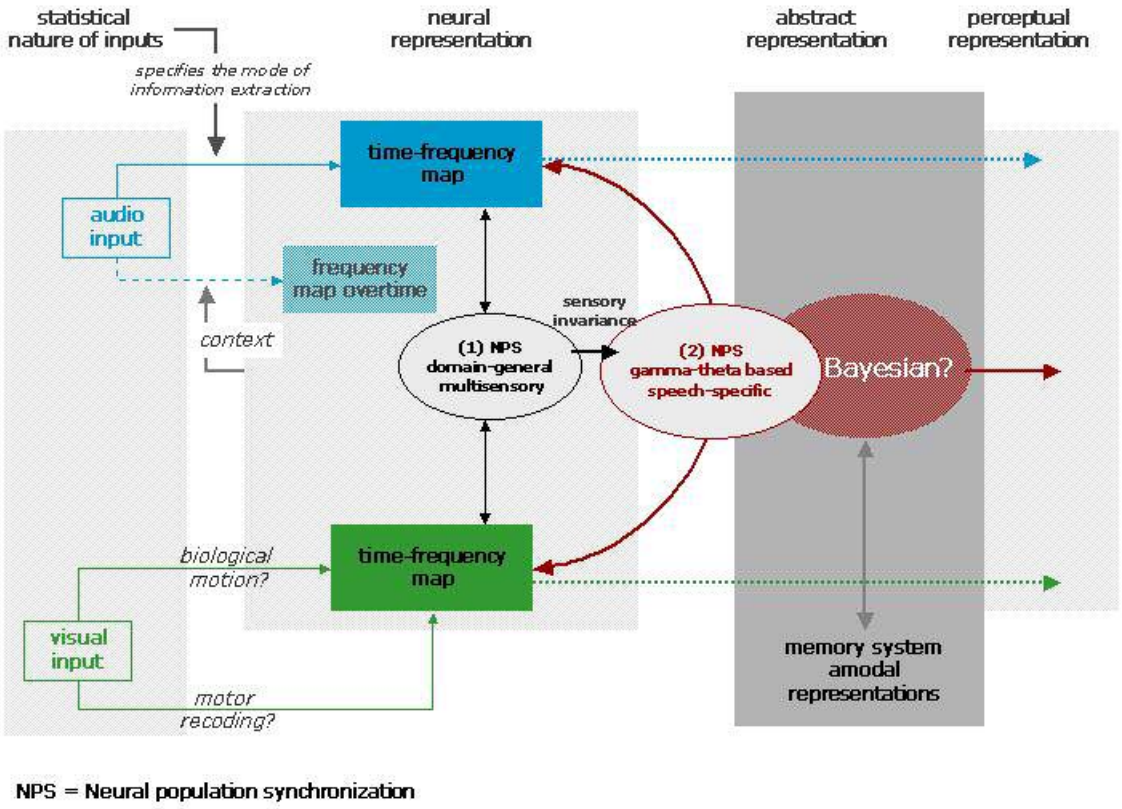


Figure 6.1: Representational system of AV POA integration in an Analysis-by-Synthesis Framework

Multiple modes of information extraction in neural systems were discussed earlier that depend on (i) the statistical nature of stimulation and (ii) the neural context. Specifically, a time-frequency map of the acoustic pattern may be readily available in the speech system that provides different temporal resolutions of the

same acoustic pattern simultaneously. Evidence was provided throughout the manuscript supporting the notion of multiple-temporal resolutions, specifically on time-scales predicted by the AST –i.e. phonetic and syllabic scales- (Poeppel, 2003). Evidence for the AV co-modulation hypothesis, show that higher local amplitudes of the co-modulation spectrum in the F2-F3 frequency space, a locus particularly relevant for the coding of 'articulatory gestures'. If a time-frequency mapping (i.e. a mapping that preserves the dynamics) of the acoustic pattern derives from this 'specialist (speech) mode' of processing is available to the auditory system, the comparator described in the AS model essentially permits an explicit comparison from time-frequency representation of POA brought about in the auditory domain to the abstract 'intended gestures' initiated in the visual domain.

The co-modulation spectrum representation may also benefit other features of AV speech integration. For instance, while visual speech does not provide explicit voicing information (e.g. Appendix B), changes in the rate of visual speech (i.e. speeding up or slowing down the frame rates of a movie) influences the perception of auditory voicing (Munhall, Green..). In the hypothesis that local higher co-modulations are computed neurally (neural population phase synchronizations/cross-correlations), the displacement in time of one the auditory or visual spectral property can bias the other. Importantly, the co-modulation of AV speech inputs may serve as a local or featural temporal tagging.

Predictive function of cortex

One of the recent notions in the understanding of brain systems is that the brain is a *predictive* device that operates on *symbolic* representations. From this standpoint, the MST and the AS models clearly posit a symbolic system of representation for instance, at the phonetic level. The 'analysis-by synthesis' model is however more appealing because it bears a neurophysiological reality. I will first illustrate this argument in the context of the classic Mismatch Negativity paradigm and subsequently extend this notion to AV speech processing.

Starting from the classic notion that the brain system is tuned to detect changes in the environment, the Mismatch Negativity (MMN) paradigm (Näätänen, 1995) was developed in electrophysiology to test the discriminative capabilities of the auditory system. The premise is as follows: during an (EEG or MEG) experiment, an auditory object A is presented more often ('standard', 85% of the time) than an auditory object B ('deviant', 15% of the time). The assumption is that the auditory system keeps in memory storage a 'template' of A. If the 'deviant' B can be neurally discriminated from the experimentally-induced 'standard' A, B elicits a MMN –i.e. the recorded brain wave for B shows a larger negativity as compared to that of A. If the deviant B does not elicit a MMN, this signifies that the physical disparity between the deviant and the standard (in memory store) cannot be neurally detected.

This paradigm has been extended to the phonemic level of representation. Specifically, for equal variations in the acoustic domain, stimuli that lead to phonemic

representation in the natural language of the participant elicit an MMN, while they do not if these categories are not part of the natural language or when the language is unknown by the participant (e.g. Philips *et al*., 2000). In the AV speech domain, the presentation of a face articulating a sound that does not match the auditory stimulus (e.g. McGurk) also elicits a MMN (Sams, 1991; Colin *et al*; Möttönen *et al.*, 200). These results suggest that the MMN encompasses the representational level at which AV speech interaction occurs -phonemic stage of representation- and these results are in line with reports of early AV interactions in Chapter III and IV.

What is the computational significance of the MMN in the context of the predictive function of neural systems? Precisely, why would the auditory system *automatically* (or 'pre-attentively') compare two inputs under this experimental condition? It is understood that in this paradigm a representation (object A) is held in memory store but this description does not provide a clue as to *why* when a different input is presented, the system would compare what is in store with what arrives. Unless, one assumes that what is in store actually stands as a 'prediction' of what comes in. In a forward model such as the analysis-by-synthesis of AV speech presented here, the architecture and the connectivity act as constraints on the neural computations. These 'on-line' constraints can be seen as recurrent circuits where incoming inputs are (locally) matched against preceding states of the system along the (global) hierarchy (again keeping in mind the inherent time-frequency processing). This functional architecture is a fundamental assumption for the implementation of an analysis-by synthesis model in AV (A) speech.

In this context of 'active analysis' of sensory inputs, the MMN obtained when a deviant stimulus is matched against the standard stimulus in sensory store may reflect the residual error of a comparative mechanism; this residual error enables to recalibrate the predicted (standard) representation induced in standard repetitions. If this predictive capability is characterized at the neural population level (or 'macro-level' as measured by functional brain imaging techniques), it only extends the argument that neural dynamics at smaller scales (for instance, lateral inhibition) and at systemic scales (for instance, top-down regulations) are a natural property of the computational cortex. If such were the case, these predictive mechanisms are observable as early as ~100ms post-auditory onset at the systemic level.

In this view, the specificity of AV (A) speech originates from the representational nature of the inputs; for instance, the very dynamics onto which they evolve that include a ~25ms and ~250ms resolutions (Poeppel, 2003). While predictive abilities of the neural system are non specific locally in that they are viewed as a general scheme of processing in forward modeling, the speech specificity originates from the interface of perceptual and production systems, i.e. in the discrete units of speech. Specifically, in the assumption that the auditory system is able to modulate the extraction of information as a function of inputs (Lewicki, 2001) and that an analogous process is available in vision (e.g. via face-biological motion processing), it is the speech system that is now handling computation as determined by the statistical nature of the inputs *and* the internal constraints posited in this system.

Appendices

Appendix A: Multisensory effects

This section is a non-exhaustive presentation of classic phenomena observed in multisensory perception.

Visual Capture in Space

'Visual capture', a term first introduced by Tastevin (1937) and later borrowed by Hay *et al.* (1965), designates the biasing effect of vision in auditory and somatosensory modalities (and proprioception) spatial localization. The effect originally characterized the biasing of proprioception in a prism-exposure paradigm, where participants were asked to locate (before and after prism-exposure) a visual target by pointing a finger of one hand with respect to their other hand position, which remained hidden at a stable location.

Biasing effects of vision over audition were first noted by Stratton (1897) and Young (1928), and later on systematically studied (e.g. Thomas, 1941) and eventually referred to as the "ventriloquism effect" (Howard & Templeton, 1966). This effect has remained under extensive study from behavioral to neurophysiological approaches, as it is stands as one of the most general effect in multisensory perception and may influence various domains of perceptual and sensorimotor integration (e.g.

Regan & Spekreijse, 1977; Macaluso *et al*, 2000; Harrington & Peck, 1998; Macaluso *et al*, 2002).

More recently with the renewal of research on multisensory integration ensuing the work by Stein & Meredith (1993) -specifically the 'spatio-temporal coincidence principle- spatial biases have been complemented with their natural temporal counterpart (e.g. Radeau & Bertelson, 1987; Soto-Faraco *et al*, 2002; Slutsky *et al*, 2001). Audition, this time, emerges as the 'biasing agent.

Auditory Driving (capture) in Time

The evidence for biasing effects of audition over vision is sparser but it was early on expected that auditory biasing of vision would emerge in the time domain (e.g. Shipley, 1964). In particular, it was noted in the study of visual capture that periodic and incongruent continuous auditory signals were less biased by visual spatial location than in other conditions (cf. Radeau & Bertelson, 1987).

Since then, more studies have started to look at the temporal perception of multisensory events using temporal rate judgment (e.g. Welch *et al*, 1986; Recanzone, 2003), duration and rhythm (e.g. Lewkowicz, 1999) and have further shown new perceptual illusions (Shams *et al*, 2000, 2002, 'auditory -visual flash-lag illusion', 2003). Auditory inputs can furthermore modulate the intensity of a visual percept (Radeau, 1985; Stein *et al*, 1996; Frassinetti *et al*, 2002) and 'capture' the direction of visual motion (Sekuler *et al*, 1997; Meyer & Wuerger, 2001).

In light of these spatio-temporal auditory-visual biasing effects, it has recently been proposed that ‘moving spatio-temporal windows of integration’ be considered in the testing of multisensory perception (Spence & Squire, 2003).

Speech

The classic McGurk effects (McGurk & McDonald, 1976, MacDonald & McGurk, 1978) have been extensively studied in the speech domain. The following table is adapted from the original reports and includes ‘fusion’ and ‘combination ‘types of results:

Input modality / stop-consonant	**Auditory (A)**	**Visual (V)** lip movements	**Percept (AV)**
Voiced	ba-ba (bilabial)	ga-ga (velar)	da-da (alveolar)
	ga-ga (velar)	ba-ba (bilabial)	gabga, bagba, baga,gaba
Voiceless	pa-pa (bilabial)	ka-ka (velar)	ta-ta (alveolar)
	ka-ka (velar)	pa-pa (bilabial)	kapka , pakpa, paka, kapa

Table A-1: Adapted from McGurk & MacDonald (1976)

Influence of lipreading can also be observed in the shifts of categorical perception of place-of-articulation (See Appendix C). Auditory-visual speech interactions are otherwise detailed throughout the manuscript.

Chronometry and Reaction Time (RT) Facilitation

Multisensory effects are accompanied with faster reaction times to the presentation of multisensory events as compared to the fastest unisensory presentation. This effect is classically referred to as 'reaction time facilitation' (Hershenson, 1962; Bernstein, 1970; Dougherty *et al.*, 1971; Gielen *et al*, 1983; Schröger, 1998) .

A RACE model would predict that in multisensory conditions, the fastest modality -i.e. the modality that will provide information first - will drive the reaction time. The RACE model does not account for such facilitation and violation of the RACE prediction serves as an empirical measure of multisensory effects.

Appendix B: Models of auditory-visual speech integration

Models of auditory-visual speech integration are classified according to the critical issue of *when* sensory-specific information combines. Two strategies have been to (i) define the nature of sensory-specific representation arriving at the integration stage -*early* vs. *late* models- and (ii) establish whether sensory-specific information is evaluated prior to the integration stage in the speech system -*dependent* vs. *independent* models. Both approaches emphasize the timing of sensory-specific information upon arrival at the integration stage.

The schematization below is a fairly consensual interpretation of processing stages in auditory speech perception (adapted from Cutting and Pisoni, 1978).

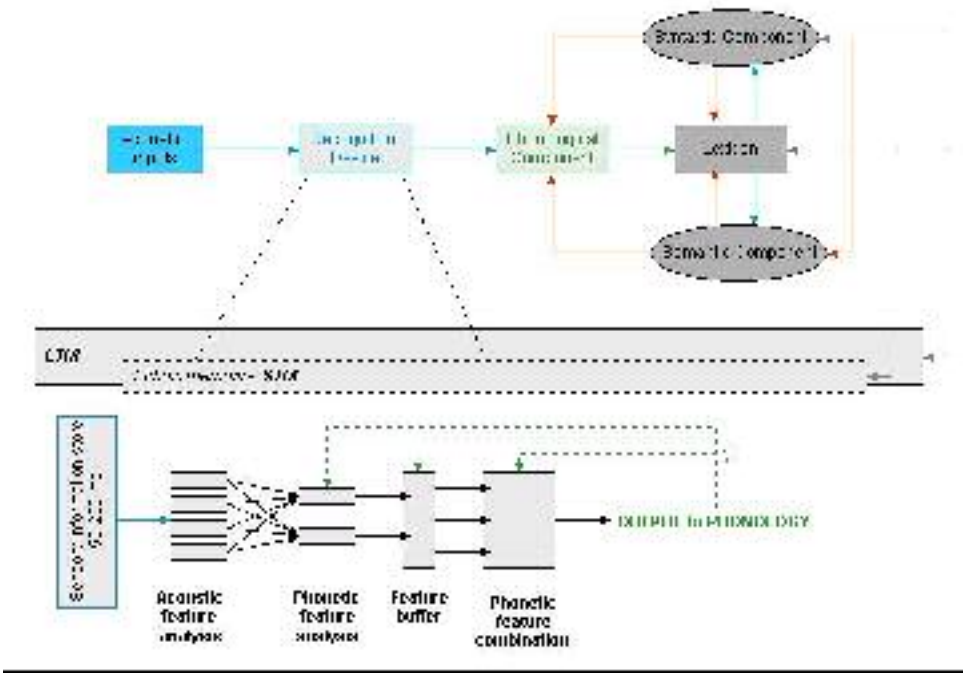


Figure B.1: Information-theoretic diagram of auditory speech processing

The question elicited by this general information-theoretic approach is whether phonological or articulatory constraints can interfere with speech-specific features at various stages of their processing i.e. (i) the early phonetic, (ii) the sensory memory and/or (iii) the phonetic feature combination stages. This question is precisely at the core of AV speech integration, because the visual inputs provide the speech system with articulatory-based information.

Auditory-visual speech models consider that an early stage of integration is pre-phonetic i.e. visual speech information could interact at either or some of the three phonetic feature representational stages. A late stage of integration, also defined as 'post-phonetic', would take place after acoustic and visual information have been categorized in their phonetic and visemic forms, respectively. Similarly, the issue of dependency assumes that sensory-specific information occurs either (i) at the featural level -dependent processing- or after sensory-specific categorization – independent processing and thus prior to or after sensory-specific categorization, respectively.

Direct identification model

In a Direct Identification model, AV speech integration (i.e. the evaluation of AV speech inputs) coincides with the decision stage. This type of model implies that sensory-specific information is in a common readable form at the integration stage but also simplifies the amount of processing that needs to be achieved from the sensory-specific channels, and could be considered an early model of integration. For

instance, the PRE labeling model proposed by Braida (1991) assumes a common representational metric.

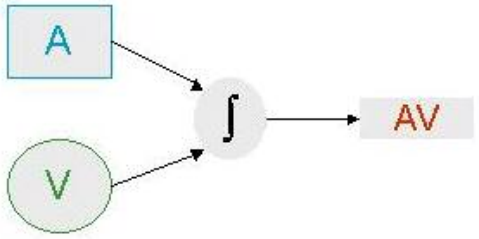


Figure B.2: Direct Identification in AV speech

Dominant recoding model

Following the dominance of the auditory system in processing speech inputs, this type of model argues that visual information be recoded in an auditory form (the dominant form) prior to being integrated with the auditory information.

The advantage is that both modalities share a common metric system prior to acceding the integration stage. This is not however the only model that may solve this issue.

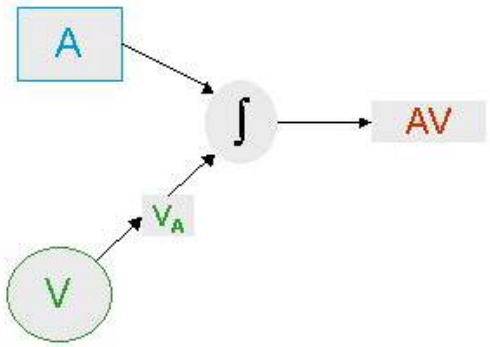


Figure B.3: Dominant recoding model in AV speech

Motor recoding model

In the Motor Theory of Speech Perception proposed by Liberman et al. (1965, 1996), the underlying articulatory gestures of speech are the metric by which auditory speech is being processed. By extension, visual speech may also follow the same encoding procedure and upon arrival the integration stage, auditory and visual information are in a motoric or amodal form of encoding. Note that this recoding stage is not specific to AV speech integration and would occur naturally regardless of the number of input modalities.

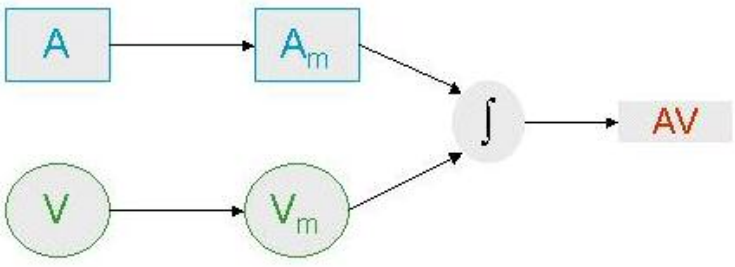


Figure B.4: Motor recoding in AV speech

Separate identification model

This type of model is essentially based on the Fuzzy-Logical Model of Perception proposed by Massaro (1987). The initial proposal was that auditory and visual speech inputs were independently evaluated prior to being integrated thus accounting for a late integration model.

More recently, a second locus of interaction - evaluation stage- has been suggested to occur prior to integration –now a decision stage- (Massaro, 1996), but it is unclear how this dichotomy fits with the early model.

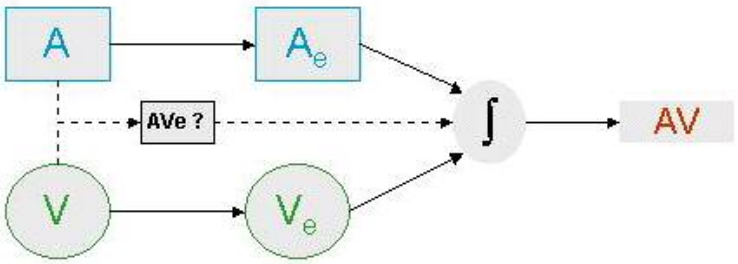


Figure B.5: Fuzzy-Logical Model of AV speech (adapted from Massaro (1998))

Appendix C: Categorical boundaries in auditory-visual speech

The following three experiments were based upon three classical findings of the auditory and auditory-visual speech literature.

First, the phonetic reality of speech perception is most graspable through early studies of categorical perception. The extraction by the speech system of phonetic features such as voice onset time and formant transitions results in the categorical perception of complex acoustic patterns. Extraction of these features from the rest of the speech spectrum results in a continuous acoustic percept.

Second, speechreading essentially provides information on the place of articulation (i.e. through the visual analysis of facial surface articulatory movements) and leads to partial or under-specified phonological categorization.

Third, the McGurk effect (detailed in Appendix A and Definitions) shows that dubbing an auditory bilabial onto a visual velar results in an alveolar auditory-visual percept.

Taken together these results predict (i) that a given VOT categorical boundary in auditory alone condition will not be influenced by the addition of discrepant visual speech input located at either end of the continuum whereas (ii) a given POA categorical boundary in A alone condition will shift in the direction of a discrepant visual speech input i.e. biased towards the visually specified POA. The following experiments were conducted to corroborate this hypothesis.

Voice-onset-time categorical boundary is insensitive to incongruent visual speech

A synthetic voice-onset time continuum was created (Klatt synthesizer) where short VOT values clearly elicited the percept [ba] and longest VOT elicited the percept [pa], for identical formant transition (i.e. directionality). Each auditory token was dubbed onto a natural face articulating [pa], each auditory onset was aligned with the original visual speech token[7].

Each token of the continuum was pseudo-randomly presented in auditory (A) alone and auditory-visual (AV) conditions. Figure C.1 shows three of spectrograms on this continuum. Ten native speakers of English participated in this experiment. They performed a 2 alternative forced-choice task (choices were [ba] or [pa]) and answered by pressing a button.

[7] Synthetic tokens were produced in a male pitch range (i.e. ~125Hz and lower than characteristic female range ~250Hz, (Pickett, 1999)) while visual displays were that of a female face. In spite of the gender disparity, the perceptual robustness of the phenomenon and the magnitude of the observed effects are quite remarkable. These results are in agreement with prior findings, suggesting that auditory-visual speech integration and more specifically, the McGurk effect are not pitch-based / gender identification (ref).

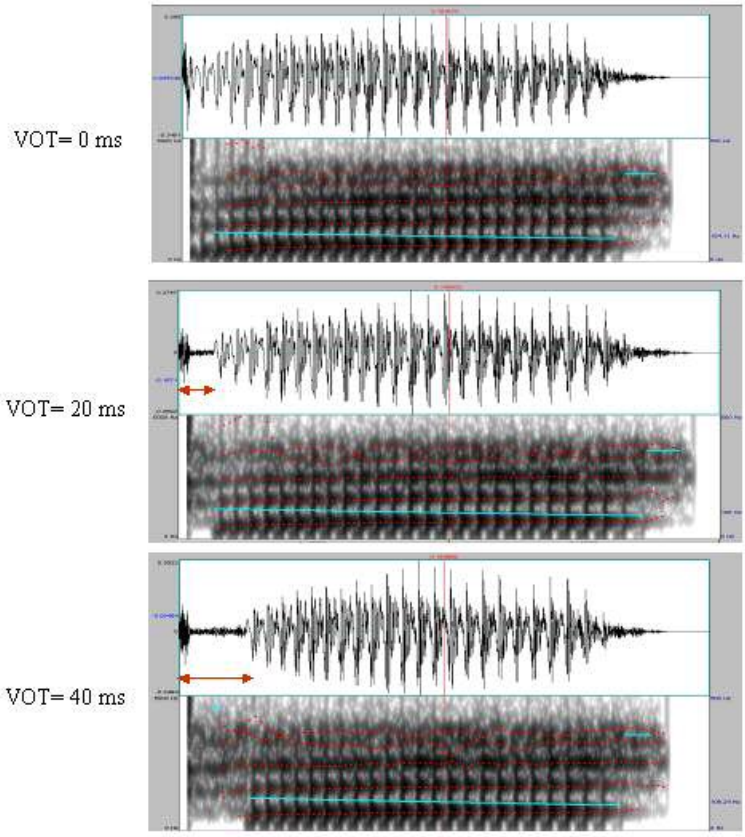


Figure C.6: Bilabial voice-onset-time continuum spectrograms for VOT =0, 20 and 40 ms

Figure C.2 reports the grand average percentage (n=10) of responses [pa] as a function of VOT (ms) to the presentation of an A alone (blue) and AV (green) voicing continuum. The addition of visual speech information did not affect the categorical boundary that was obtained in the A alone condition -i.e. responses [ba] and [pa] were given by chance for a VOT of ~30-40ms regardless of the input modality.

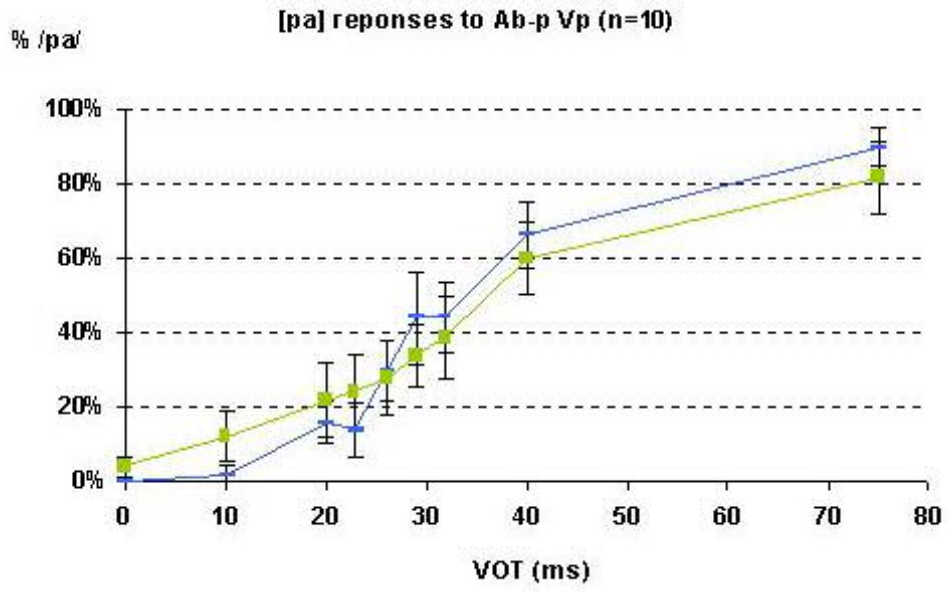


Figure C.2: Categorical perception with an auditory-visual speech voice-onset time in a bilabial /b/– /p/ continuum

Place-of-articulation categorical boundaries are biased towards visual speech

In the following 2 experiments, a synthetic place-of-articulation was used where upwards formant transitions (F2-F3 regions) elicited the percept [ba] or [da] and downwards formant transitions elicited the percept [da] and [ta], for the voiced (VOT less than ~30ms) and voiceless (VOT longer than ~30ms) continua, respectively. As in the VOT experiment, auditory tokens were dubbed onto a natural face articulating [ga] (voiced condition) or [ka] (voiceless condition) and each auditory onset was aligned with the original visual speech token.

Tokens were pseudo-randomly presented in auditory (A) alone and auditory-visual (AV) conditions. Twenty-six native speakers of English participated in these experiments. Participants performed a 2 AFC task by pressing buttons (choices were [ba] or [da] in the voiced condition and [pa] or [ta] in the voiceless condition).

Figure C.3 shows a higher rate of /d/ responses in AV (green) conditions as compared to A alone condition (blue). In AV condition, the transition point (50%) or perceptual shift between /b/-/d/is reached earlier than in A alone.

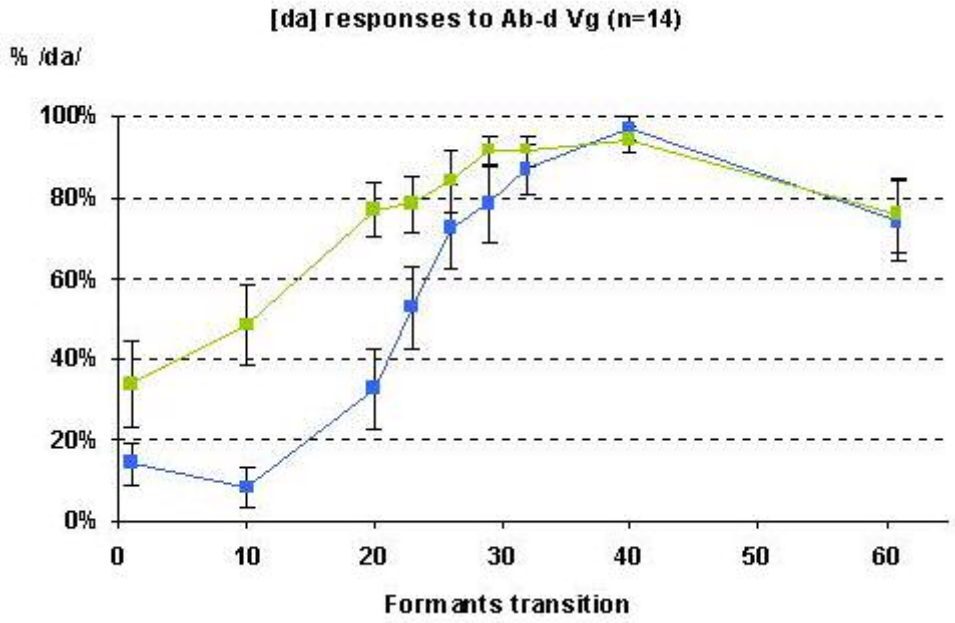


Figure C.3: Shift of categorical in AV perception of a voiced continuum /b/-/d/ dubbed onto an incongruent visual /g/

Similarly, in the Figure C.4 shows a higher rate of /t/ responses in AV (green) conditions as compared to A alone condition (blue), showing that the McGurk effect (McGurk & McDonald, 1976) took place when audio [pa] is dubbed onto a visual [ka] and this effect also affects the perception of a phonetic continuum.

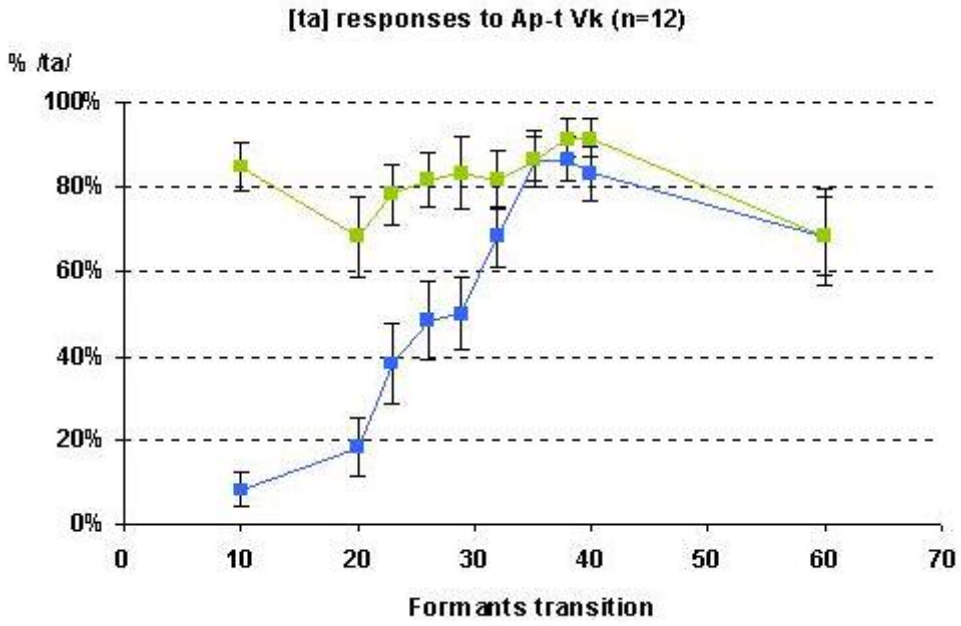


Figure C.4: Shift of categorical boundary in AV perception of a voiceless continuum /p/-/t/ dubbed onto an incongruent visual /k/

Acknowledgments

I would like to thank Colin Philips for providing me with the set of synthetic tokens.

Appendix D: Cortical dynamics and neural source characterization of auditory-visual interactions

(Experiment 10 in progress)

In this study, the working hypotheses follow from Experiment 9. Specifically, in Chapter III we showed that a latency shift and an amplitude reduction of early auditory evoked-potentials characterized AV speech presentation as compared to audio alone presentation. Control experiments also provided in Chapter III supported the idea that the nature and the saliency of visual speech information preceding the audio onset determine patterns of auditory evoked responses in both the latency and the amplitude domain. These results were interpreted as evidence for an 'analysis-by-synthesis' model of AV speech perception that mainly considered the ambiguity of visual speech information and the perceptual redundancy of auditory and visual stimuli.

Note that no prediction was made as to the extent of auditory-visual sensory invariance for two reasons. First, it was assumed that in AV natural condition, AV stimulation provides what is considered a 'natural sensory invariance'. Secondly, in AV speech, no obvious supra-additivity or enhanced activation was observed in any region of the scalp within the ~250ms following auditory onset. The surprising lack of supra-additivity in AV speech contrasts with general findings of multisensory integration (including that reported in Experiment 9).

Here, I want to contrast various parameters, which I consider crucial for multisensory perceptual formation. In particular in the dissociation between meaningless events (Experiment 9) and meaningful events (Chapter III and IV). In light of the difficulty to interpret various effects observed in Experiment 9, the description of cortical dynamics for increasingly spectro-temporally complex AV stimulation may help disambiguate (i) the involvement of multisensory-integration-by-convergence with (ii) perceptual domain specifics computations (such as those observed in AV speech). In particular, in this experiment, the following working hypotheses are:

1) Static meaningless events should lead to classic supra-additive effects and either implicate a multisensory source (e.g. STS) or direct inter-sensory connectivity
2) Co-modulated meaningless events should lead to supra-additive effects and possibly involve the biological motion pathway. Supra-additivity in sensory-specific areas is predicted in both theta and gamma bands (replicating Experiment 9 study) and STS supra-additivity is hypothesized in these frequency ranges.
3) In AV speech, a shortening of sustained gamma band activation is predicted over the auditory cortices (STG). Sources of activation are predicted to include a large network in both frontal and motor cortices (for both visual and AV speech conditions, in agreement with fMRI studies).

In this experiment, no response was required from the participants, who were simply told to attend the stimuli. This paradigm should help disambiguate for instance the effects observed in the alpha and beta1 frequency ranges of Experiment 8.

Materials and Methods

Participants

Seven participants were recruited from the University of Maryland population took part in this experiment. No participant had diagnosed hearing problems, all had normal or corrected-to-normal vision and were right-handed. The study was carried out with the approval of the University of Maryland Institutional Review Board.

Stimuli and Procedure

Three sets of auditory-visual stimulation were used. Auditory stimulations consisted of a 800 Hz pure tone (5ms rise-fall ramping) in the static condition, a band-pass filtered white noise (500Hz-1200Hz; 5ms rise-fall ramping) and an natural audio [pa] recorded from a female voice (same stimulus used in chapter II).

Visual stimulations were derived from the natural female display articulating [pa]. For the static condition, a neutral face was chosen and the mouth area was extracted from the original movie. One frame was coarsely filtered (using a mosaic filter setting 16?, Software) so as to render the stimulus non-identifiable as a mouth.

This stimulus constitutes the static visual stimulus. To construct the dynamic visual stimulus, the mouth and nose areas were extracted and the same filtering procedure as in the static condition was used on every frames. This procedure resulted in a smaller area movie, which preserved the global dynamics of the mouth/nose areas while rendering the stimulus unidentifiable as a face stimulus. Finally the third visual stimulus consisted of the original woman's face articulating the syllable [pa].

Nine conditions were thus obtained that are audio alone 'static' (As), 'dynamic' (Ad), and 'speech [pa]' (Ap), video alone (Vs), 'dynamic' (Vd), and 'speech [pa]' (Vp), and audio-visual conditions.

Magnetoencephalographic recordings

Participants' head shape was digitized using a Polhemus system prior to taking part in the MEG recording session. This digitization procedure permits to specify the parameters of the head-shape model (spherical) for source reconstruction. MEG recordings were made using a whole-head SQUID system (MEG160, Kyoto Institute of Technology) consisting of 160 recording channels.

Data were recorded by blocks of ~8 min (pseudo-randomized presentation of 20 presentations per stimulus and per block). Between 6 to 10 blocks were acquired per participant. Data were band-pass filtered 1-250Hz online with an output gain setting of and input gain setting of. The acquisition sampling rate was set to 4kHz.

Glossary

Categorical perception:

discretized perceptual output resulting from the presentation of a continuously varied feature. Categorical perception is characterized by a step-like shift in perception accompanied by an increased reaction time at the location of the perceptual shift. Within categories variation of stimulation result cannot be discriminated.

Phoneme: perceptual stage of representation resulting from the synthesis of speech features. Complete phonemic representation is achieved in the phonological store.

Place of articulation (POA):

is defined by the place of the articulators in the vocal tract during speech production. The frontal vs. back placement of the articulators impact the directionality of formant transitions. For instance, 'bilabials' ([ba], [pa]) are characterized by frontal occlusion of the upper and lower lips as opposed to 'velars' ([ga], [ka]) where occlusion originates from the velum and the tongue.

Stop consonant:

consonant characterized by a short burst of energy (high turbulences in the vocal tract) at their onset. Examples of stop-consonants by place of articulation feature are:

Bilabials: [ba] voiced (i.e. the vocal folds pulsing continues some time during closure) and [pa], voiceless

Velar: [ga] (voiced) and [ka] (voiceless)

Alveolar: [da] (voiced) and [ta] (voiceless)

Sub –additivity: see supra-additivity

Supra (sub)-additivity:

characterizes a typical response pattern of multisensory neurons. The response of a multisensory neuron to the presentation of co-occurent multisensory events is said 'supra-additive' when it is *larger*(i.e. longer duration and higher spiking rate) than that predicted from the summation of outputs to the presentation of identical unisensory stimuli (i.e. stimuli presented separately0. By contrast, sub-additivity characterizes a depressed output in response to non spatio-temporal coincident multisensory inputs.

Viseme: perceptual stage of representation in the visual domain (i.e. speechreading) analogous to phonemic representation. Visemic representation is articulatory-based and do not result in a complete phonological categorization –i.e. visemes are under-specified in their voice-onset time and possibly manner.

Voiced: speech feature produced by the short continuation of vocal folds pulsing during closure.

Voiceless: speech feature where the pulsing of vocal folds is absent during closure.

Voice-Onset Time (VOT):

time that elapses from the release of the occlusion of the vocal tract to the beginning of the vocal folds vibration (voicing).

Bibliography

Ackermann, H., W. Lutzenberger, et al. (1999). "Hemispheric lateralization of the neural encoding of temporal speech features: a whole-head magnetoencephalography study." Cognitive Brain Research **7**: 511-518.

Alain, C., S. R. Arnott, et al. (2001). ""What" and "where" in the human auditory system." Proceedings of the National Academy of Science **98**(21): 12301-12306.

Allison, T., A. Puce, et al. (1999). "Electrophysiological studies of human face perception. I: potentials generated in occipitotemporal cortex by face and non-face stimuli." Cerebral Cortex **9**: 415-430.

Arai, T. and S. Greenberg (1998) Speech intelligibility in the presence of cross-channel spectral asynchrony, IEEE International Conference on Acoustics, Speech and Signal Processing, Seatle, pp. 933-936.

Aschersleben, G. and P. Bertelson (2003). "Temporal ventriloquism: crossmodal interaction on the time dimension: 2. Evidence from sensorimotor integration." International Journal of Psychophysiology **50**: 157-163.

Badgaiyan, R. D., D. L. Schacter, et al. (1999). "Auditory priming within and across modalities: evidence from positron emission tomography." Journal of Cognitive Neuroscience **11**(4): 337-348.

Bagdaiyan, R. D., D. L. Schacter, et al. (2001). "Priming within and across modalities: exploring the nature of rCBF increases and decreases." NeuroImage **13**(272-282).

Bahrick, L. E. (1992). "Infant's perceptual differentiation of amodal and modaliy-specific audio-visual realations." Journal of Experimental Child Psychology **53**: 180-199.

Barlow, H. B. (1972). "Single units and sensation: a neuron doctrine for perceptual psychology." Perception **1972**(1): 371-394.

Basar, E. (1998). Brain Function and Oscillations. I. Brain Oscillations:Principles and Approaches. Berlin, Heidelberg, Springer.

Basar, E. (1999). Brain Function and Oscillations. II. Integrative Brain Func-tion.

Neurophysiology and Cognitive Processes. Berlin,Heidelberg, Springer.

Basar, E., C. Basar-Eroglu, et al. (2000). "Brain oscillations in perception and memory." International Journal of Psychophysiology: 95-124.

Basar-Eroglu, C., D. Strüber, et al. (1996). "Frontal gamma band enhancement during multistable visual perception." International Journal of Psychophysiology **24**: 113-125.

Bavelier, D. and H. J. Neville (2002). "Cross-modal plasticity: where and how?" Nature Reviews Neuroscience **3**: 443-452.

Baylis, G. C., E. T. Rolls, et al. (1985). "Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey."Brain Research **342**: 91-102.

Baynes, K., M. G. Funell, et al. (1994). "Hemispheric contributions to the integration of visual and auditory information in speech perception." Perception and Psychophysics **55**(6): 633-641.

Beauchamp, M. S., K. E. Lee, et al. (2003). "FMRI responses to video and point-light displays of moving humans and manipulable objects." J Cogn Neurosci **15**(7): 991-1001.

Belin, P., R. J. Zatorre, et al. (2000). "Voice-selective areas in human auditory-cortex." Nature **403**: 309-312.

Benevento, A., J. Fallom, et al. (1977). "Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey." Experimental Neurology **57**: 849-872.

Bernstein, I. H. (1970). Can we see and hear at the same time? Acta Psychologica Attention and Performance III. A. F. Sanders. Amsterdam, North-Holland Publishing Company**:** 21-35.

Bernstein, I. H., R. Rose, et al. (1970). "Energy integration in intersensory facilitation." Journal of Experimental Psychology **86**(2): 196-203.

Bertelson, P. and G. Aschersleben (2003). "Temporal ventriloquism: crossmodal interaction on the time dimension: 1. Evidence from auditory-visual temporal order judgment." International Journal of Psychophysiology **50**: 147-155.

Bertelson, P. and M. Radeau (1981). "Cross-modal bias and perceptual fusion with

auditory-visual spatial discordance." Perception and Psychophysics **29**(6): 578-584.

Bertrand, O., Tallon-Baudry, C. (2000). “Oscillatory gamma activity in humans: a possible role for object representation”. International Journal of Psychophysiology, **38**: 211-223

Bhattacharya, J., L. Shams, et al. (2002). "Sound-induced illusory flash perception: role of gamma band responses." Neuroreport **13**(14): 1727-1730.

Binder, J. (2000). "The new neuroanatomy of speech perception." Brain **123**: 2371.

Binnie, C. A., Montgomery, A.A., et al. (1974). "Auditory and visual contributions to the perception of consonants." Journal of Speech and Hearing Research **17**: 619-630.

Blake, R. and Y. Yang (1997). "Spatial and temporal coherence in perceptual binding." Proceedings of the National Academy of Science **94**: 7115-7119.

Blakemore, S.-J. and J. Decety (2001). "From the perception of action to the understanding of intention." Nature Reviews Neuroscience **2**: 561-567.

Blamey, P. J., Cowan, R. S. et al. (1989). "Speech perception using combinations of auditory, visual, and tactile information." Journal of Rehabilitation Research and Development **26**(1): 15-24.

Bland, B. H. and S. D. Oddie (2001). "Theta bad oscillation and synchorny in the hippocampal formation and associated structures: the case for its role in sensorimotor integration." Behavioural Brain Research **127**: 119-136.

Blumstein, S. E. and K. N. Stevens (1979). "Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonant." Journal of the Acoustical Society of America **66**(4): 1001-1017.

Bodner, M., J. Kroger, et al. (1996). "Auditory memory cells in dorsolateral prefrontal cortex." Neuroreport **7**(12): 1905-1908.

Braida, L. D., K. Sekiyama, et al. Integration of audiovisually compatible and incompatible consonants in identification experiments.

Brecht, M., R. Goebel, et al. (2000). "Synchronization of visual responses in the superior colliculus of awake cats." Neuroreport **12**(1): 43-47.

Brecht, M., R. Goebel, et al. (2001). "Synchronization of visual responses in the

superior colliculus of awake cats." Neuroreport **12**(1): 43-7.

Brecht, M., W. Singer, et al. (1998). "Correlation analysis of corticotectal interactions in the cat visual system." J Neurophysiol **79**(5): 2394-407.

Brown, A. E. and H. K. Hopkins (1966). "Interaction of the auditory and visual sensory modalities." Journal of the Acoustical Society of America **41**(1): 1-6.

Brown, H. and S. Kosslyn (1993). "Cerebral lateralization." Current Opinion in Neurobiology **3**: 183-186.

Bruce, C. C., R. Desimone, et al. (1986). "Both striate cortex and superior colliculus contribute to visual properties of neurons in superior temporal polysensory area of macaque monkey." Journal of Neurophysiology **55**(5): 1057-1075.

Brunia, C. H. M., de Jong, B. M. et al. (2000). "Visual feedback about time estimation is related to right hemisphere activation measured by PET." Experimental brain Research **130**: 328-337.

Burns, M. M. and H. Moskowitz (1971). "Response time to a first signal as a function of time relationship to a second signal and mode of presentation." Perceptual and Motor Skills **32**: 811-816.

Bushara, K. O., J. Grafman, et al. (2001). "Neural correlates of auditory-visual stimulus onset asynchrony detection." Journal of Neuroscience **21**(1): 300-304.

Bushara, K. O., T. Hanakawa, et al. (2003). "Neural correlates of cross-modal binding." Nature Neuroscience **6**(2): 190-495.

Buzsáki, G. (2002). "Theta oscillations in the hippocampus." Neuron **33**: 325-340.

Byrne, R. W. (1995). The thinking ape. Oxford, New York, Tokyo, Oxford University Press.

Callan, D. E., A. M. Callan, et al. (2001). "Multimodal contribution to speech perception revealed by independent component analysis: a single-sweeo EEG case study." Cognitive Brain Research **10**: 349-353.

Calvert, G. A. (1997). "Activation of auditory cortex during silent lipreading." Science **276**: 893-596.

Calvert, G. A. (2001). "Crossmodal processing in the human brain: insights from functional neuroimaging studies." Cerebral Cortex **11**: 1110-1123.

Calvert, G. A., M. J. Brammer, et al. (1999). "Response amplification in sensory-specific cortices during cross-modal binding." Neuroreport **10**(12): 2619-2623.

Calvert, G. A., M. J. Brammer, et al. (1998). "Crossmodal Identification." Trends in Cognitive Sciences **2**(7): 247-253.

Calvert, G. A. and R. Campbell (2003). "Reading speech from still and moving faces: the neural substrates of visible speech." Journal of Cognitive Neuroscience **15**(1): 57-70.

Calvert, G. A., R. Campbell et al. (2000). "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex." Current Biology **10**: 649-657.

Calvert, G. A., P. C. Hansen, et al. (2001). "Detection of audio-visual integration sites in humans by application of electrophysiological criteria in teh BOLD effect." NeuroImage **14**: 427-438.

Campbell, C. and D. W. Massaro (1997). "Perception of visible speech: influence of spatial quantization." Perception **26**: 627-644.

Campbell, R. (1986). "Face recognition and lipreading." Brain **109**: 509-521.

Campbell, R. (1988). Tracing Lip Movements: Making Speech Visible. Visible Language**:** 32-57.

Campbell, R. (1989). Lipreading. Handbook of Research on Face Processing. Y. A.W. and E. H.D., Elsevier Science Publishers**:** 187-233.

Campbell, R. (1992). Lip-reading and the modularity of cognitive function: neuropsychological glimpses of fractionation from speech and faces. Analytic approaches to human cognition. A. J., H. D, J. d. M. J and R. M., Elsevier Science Publishers**:** 275-289.

Campbell, R. (1998). "Everyday speechreading: understanding seen speech in action." Scandinavian Journal of Psychology **39**(3): 163-167.

Campbell, R., J. Garwood, et al. (1990). "Neuropsychological studies of auditory-visual fusion illusions. Four case studies and their implications." Neuropsychologia **28**(8): 787-802.

Campbell, R., J. Zihl, et al. (1997). "Speechreading in the akinetopsic patient, L.M."

Brain **120**: 1793-1803.

Campbell, R., Dodd, B., & Burnham, D. (Eds.) (1998). Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech. Psychology Press Ltd., East Sussex, UK.

Caplan, J. B., J. R. Madsen, et al. (2003). "Human theta oscillations related to sensorimotor integration and spatial learning." The Journal of Neuroscience **23**(11): 4726-4736.

Carlesimo, G. A., P. Turriziani, et al. (2003). "Brain activity during intro- and cross-modal priming: new empirical data and review of the literature." Neuropsychologia **42**: 14-24.

Carmon, A. and I. Nachshon "Effect of unilateral brain damage on perception of temporal order." Cortex **7**: 410-418.

Celesia, G. G. (1976). "Organization of auditory cortical areas in man." Brain **99**: 403-414.

Chee, M. W. L., K. M. O'Craven, et al. (1999). "Auditory and visual word processing studied with fMRI." Human Brain Mapping **7**: 15-28.

Chomsky, N. (2000). Recent contributions to the theory of innate ideas. Minds, brains and computers The foundation of cognitive science, an anthology. Malden, Oxford, Blackwell**:** 452-457.

Chomsky, N. and M. Halle (1968). The sound pattern of English. New York, Evanston, London, Harper & Row.

Cigánek, L. (1966). "Evoked potentials in man: interaction of sound and light." Electroencephalography and Clinical Neurophysiology **21**: 28-33.

Cohen, Y. E., A. P. Batista, et al. (2002). "Comparison of neural activity preceding reaches to auditory and visual stimuli in the parietal reach region." Neuroreport **13**(6): 891-4.

Coles, M. G. H., G. O. M. Henderikus, et al. Mental chronometry and the study of human information processing.

Colin, C., M. Radeau, et al. (2002). "Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory." Clinical Neurophysiology **113**: 495-506.

Colombo, M., H. R. Rodman, et al. (1996). "The effects of superior temporal cortex lesions on the processing and retention of auditory information in monkeys *(Cebus apella)*." The Journal of Neuroscience **16**(14): 4501-4517.

Conrey, B. L. and D. B. Pisoni (2003). "Audiovisual asynchrony detection for speech and nonspeech signals". In *AVSP 2003*, 25-30.

Corrales, T. A. and J. I. Aunón (2000). "nonlinear system identification and overparameterization effects in multisensory evoked potential studies." IEEE Transactions on Biomedical Engineering **47**(4): 472-486.

Craik, K. (1943). Hypothesis on the nature of thought. The Nature of Explanation. Cambridge, Cambridge University Press**:** Chapter 5.

Cummins, R., and D. D. Cummins (2000) Minds, Brains, and Computers. The Foundation of Cognitive Science. An Anthology. Blackwell.

Curran, T., D. L. Schacter, et al. (1999). "Cross-modal priming and explicit memory in patients with verbal production deficits." Brain and Cognition **39**: 133-146.

Cutting, J.E. & Pisoni, D.B. (1978). An information-processing approach to speech perception. In J.F. Kavanagh & W. Strange (Eds.), Speech and language in the laboratory, school, and clinic, (pp. 38-72). Cambridge, MA: MIT Press.

Czigler, I. and L. Balázs (2001). "Event-related potentials and audiovisual stimuli: multimodal interactions." Neuroreport **12**(2): 223-226.

da Silva, L. F. (1991). "Neural mechanisms underlying brain waves: from neural membranes to networks." Electroencephalography and Clinical Neurophysiology **79**: 81-93.

Davis, A. E. and J. A. Wada (1977). "Hemispheric asymmetries of visual and auditory information processing." Neuropsychologia **15**: 799-806.

de Gelder, B. and P. Bertelson (2003). "Multisensory integration, perception and ecological validity." Trends in Cognitive Science.

de Gelder, B., K. B. E. Böcker, et al. (1999). "The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses." Neuroscience Letters **260**: 133-136.

Desimone, R. and C. G. Gross (1979). "Visual areas in the temporal cortex of the macaque." Brain Research **178**: 363-380.

Desmond, M. (1967). L'éthologie des primates. Bruxelles, Weidenfeld, Nicolson.

Dijkstra, T., U. H. Frauenfelder, et al. (1993). "Bidirectional grapheme-phoneme activation in a bimodal detection task." Journal of Experimental Psychology: Human Perception and Psychophysics **19**(5): 931-950.

Dinse, H. R., K. Krüger, et al. (1997). "Low-frequency oscillations of visual, auditory and somatosensory cortical neurons evoked by sensory stimulation." International Journal of Psychophysiology **26**: 205-227.

Dinse, H. R., F. Spengler, et al. (1993). Dynamic aspects of cortical function: processing and plasticity in different sensory modalities. Brain Theory. A. A., Elsevier Science Publishers.

Dixon, N. F. and L. Spitz (1980). "The detection of auditory visual desynchrony." Perception **9**: 719-721.

Dodd, B. (1979). "Lip reading in infants: attention to speech presented in- and out-of-synchrony." Cognitive psychology **11**: 478-484.

Dodd, B. and R. Campbell (1987). Hearing by Eye: The Psychology of Lip-Reading. London, Lawrence Erlbaum.

Dougherty, W. G., G. B. Jones, et al. (1971). "Sensory integration of auditory and visual information." Canadian Journal of Psychology **25**(6): 476-485.

Downar, J., A. P. Crawley, et al. (2001). "The effect of task-relevance on the cortical response to changes in visual and auditory stimuli: an event-related fMRI study." NeuroImage **14**: 1256-1267.

Downar, J., A. P. Crawley, et al. (2002). "A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities." Journal of Neurophysiology **87**: 615-520.

Downar, J., D. J. Mikulis, et al. (2000). "A multimodal cortical network for the detection of changes in the sensory environment." Nature Neuroscience **3**: 277-283.

Driver, J. (1996). "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading." Nature **381**: 66-67.

Driver, J. and C. Spence (2000). "Multisensory perception: beyond modularity and convergence." Current Biology **10**(20): R731-735.

Duffau, H., P. Gatignol, et al. (2003). "The articulartory loop: study of the subcortical connectivity by electrostimulation." Neuroreport **14**(15): 2005-2008.

Ehret, G. (1997). "The auditory cortex." journal of Comparative Physiology A **181**: 547-557.

Eijkman, E. and J. H. Vendrik (1965). "Can a sensory system be specified by its internal noise?" Journal of the Acoustical Society of America **37**(6).

Eimer, M., D. Cockbrun, et al. (2001). "Cross-modal links in endogenous spatial attention are mediated by common external locations: evidence from event-related potentials." Experimental Brain Research **139**: 398-411.

Eimer, M. and E. Schröger (1998). "ERP effects of intermodal attention and cross-modal links in spatial attention." Psychophysiology **35**: 313-327.

Engel, A. K., P. R. Roelfsema, et al. (1997). "Role of the temporal domain for response selection and perceptual binding." Cerebral Cortex **7**: 571-582.

Erber, M. P. (1978). "Auditory-visual speech perception of speech with reduced optical clarity." Journal of Speech and Hearing Research **22**: 213-223.

Ehrenfels C. von (1890/1988) On "Gestalt Qualities". In B. Smith (Ed. & Trans.) Foundations of Gestalt Theory. Wien: Philosophia Verlag. pp 82-117.

Evans, E. F. (1992). "Auditory processing of complex sounds: an overview." Philosophical Transactions of the Royal Society of London B: 295-306.

Falchier, A., S. Clavagnier, et al. (2002). "Anatomical evidence of multimodal integration in primate striate cortex." The Journal of Neuroscience **22**(13): 5749-5759.

Farah, M. (1990). Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision. Cambridge/Bradford Books, MIT Press.

Faulkner, A. and S. Rosen (1999). "Contributions of temporal encoding of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception." Journal of the Acoustical Society of America **106**(4): 2063-2073.

Felleman, D. J. and D. C. van Essen (1991). "Distributd hierarchical processing in the primate cerebral cortex." Cerebal Cortex **1**(1-47).

Ffytche, D. H., C. N. Guy, et al. (1995). "The parallel visual motion inputs into areas

V1 and V5 of human cerebral cortex." Brain **118 (Pt 6)**: 1375-94.

Fort, A., C. Delpuech, et al. (2002). "Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans." Cerebal Cortex **12**: 1031-1039.

Fort, A., C. Delpuech, et al. (2002). "Early auditory-visual interaction in human cortex during nonredundant target identification." Cognitive Brain Research **14**: 20-30.

Foxe, J. J., I. Morocz, A., et al. (2000). "Multisensory auditory-somatosensory interactiosn in early cortical processing revealed by high-density electrical mapping." Cognitive Brain Research **10**: 77-83.

Foxe, J. J., G. R. Wylie, et al. (2002). "Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study." Journal of Neurophysiology **88**: 540-543.

Frassinetti, F., N. Bolognigni, et al. (2002). "Enhancement of visual perception by crossmodal visuo-auditory interaction." Experimental Brain Research **147**: 332-343.

Fries, P., J. H. Reynolds, et al. (2001). "Modulation of oscillatory neuronal synchronization by selective visual attention." Science **291**: 1560-1563.

Friston, K. (2002). "Functional integration and inference in the brain." Progress in Neurobiology **68**: 113-143.

Fritz, J., S. Shamms, et al. (2003). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex." Nature Neuroscience **6**(11): 1216-1223.

Fuster, J. M., M. Bodner, et al. (2000). "Cross-modal and cross-temporal association in neurons of frontal cortex." Nature **405**: 347-351.

Gabrieli, J., Russell, P. and J. Desmond (1998). “The role of prefrontal cortex in language and memory.” Proc. Natl. Acad. Sci. USA **95**: 906-913.

Geldard, F. A. and C. Sherrick, E. (1972). "The cutaneous "Rabbit": a perceptual illusion." Science **176**: 178-179.

Ghazanfar, A. A. and N. K. Logothetis (2003). "Facial expressions linked to monkey calls." Nature **423**: 937-938.

Giard, M. H. and F. Peronnet (1999). "Auditory-visual integration during multimodal object recognition in Humans: a behavioral and electrophysiological study." Journal of Cognitive Neurscience **11**(5): 473-490.

Gibbon, J., C. Malapani, et al. (1997). "Toward a neurobiology of temporal cognition: advances and challenges." Current Opinion in Neurobiology **7**: 170-184.

Gibson, E. J. (1969). Principles of perceptual learning and development. New York, Appleton - Century - Crofts.

Gibson, J. J. (1966). The Senses Considered as Perceptual Systems. Boston, Houghton Mifflin.

Gibson, J. R. and J. H. R. Maunsell (1997). "Sensory modality specificity of neural activity related to memory in visual cortex." Journal of Physiology: 1263-1275.

Gielen, S. C. A. M., R. A. Schmidt, et al. (1983). "On the nature of intersensory facilitation of reaction time." Perception and Psychophysics **34**(2): 161-168.

Giraud, A.-L., C. J. Price, et al. (2001). "Cross-modal plasticity underpins language recovery after cochlear implantation." Neuron **30**: 657-663.

Girin, L., J.-L. Schwartz, et al. (2001). "Audio-visual enhancement of speech in noise." Journal of the Acoustical Society of America **109**(6): 3007-30220.

Gold, J., P. J. Bennett, et al. (1999). "Signal but not noise chances with perceptual learning." Nature **402**: 176-178.

Gonzalo, D., T. Shallice, et al. (2000). "Time-dependent changes in learning auditory-visual associations: a single trial fMRI study." Neuroimage.

Graf, O., A. O. Shimamura, et al. (1985). "Priming across modalities and priming across category levels: extending the domain of preserved function in amnesia." Journal of Experimental Psychology: Learning, Memory, and Cognition **11**: 386-396.

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

Grant, K. W. (2002). "Measures of auditory-visual integration for speech understanding: a theoretical perspective." Journal of the Acoustical Society of

America.

Grant, K. W. (2003). Auditory supplements to speechreading. ATR Workshop "Speech Dynamics by Ear, Eye, Mouth and Machine", Kyoto, Japan.

Grant , K. W. and S. Greenberg (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. Auditory-Visual Speech Pocessing, Scheelsminde, Denmark.

Grant, K. W. and P.-F. Seitz (1998). "Measures of auditory-visual integration in nonsense syllables." Journal of the Acoustical Society of America **104**(4): 2438-2450.

Grant, K. W. and P.-F. Seitz (1998). The use of visible speech cues (speechreading) for directing auditory attention: reducing temporal and spectral uncertainty in auditory detection of spoken sentences. 16th International Congress on Acoustics, 135th Meeting of the Acoustical Society of America, Washington, Seattle.

Grant, K. W. and P.-F. Seitz (2000). "The use of visible speech cues for improving auditory detection of spoken sentences." Journal of the Acoustical Society of America **108**(3): 1197-1207.

Grant, K. W. and P. F. Seitz (1998). "Measures of auditory-visual integration in nonsense syllables and sentences." J.Acoust.Soc.Am. **104**(4): 2438-2450.

Grant, K. W., V. van Wassenhove, et al. (2003). Discrimination of auditory-visual synchrony. Auditory-Visual Speech Processing, St Jorioz, France.

Grant, K. W. and B. E. Walden (1995) Predicting auditory-visual speech recognition in hearing-impaired listeners. XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden, Vol. 3, 122-129.

Grant, K. W. and B. E. Walden (1996). "Spectral distribution of prosodic information." Journal of Speech and Hearing Research **39**: 228-238.

Grant, K. W. and B. E. Walden (1996). "Evaluating the articulation index for auditory-visual consonant recognition." Journal of the Acoustical Society of America **100**(4): 2415-2424.

Grant , K. W., B. E. Walden, et al. (1998). "Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition and

auditory-visual integration." Journal of the Acoustical Society of America **103**(5): 2677-2690.

Grant, K. W., B. E. Walden, et al. (1998). "Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration." J.Acoust.Soc.Am. **103**(5): 2677-2690.

Green, K. P. (1996) "Studies of the McGurk effect: implications for theories of speech perception." Proceedings

Green, K. P. (1987). "The use of auditory and visual information in phonetic perception.". Hearing by Eye: The Psychology of Lip-Reading. Hillsdale, NJ., Lawrence Erlbaum Associates.55-77

Green, K. P. and A. Gerdeman (1995). "Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels." Jounal of Experimental Psychology: Human Perception and Performance **21**(6): 1409-1426.

Green, K. P. and P. Kuhl (1989). "The role of visual information in the processing of place and manner features in speech perception." Perception and Psychophysics **45**: 34-42.

Green, K. P., P. K. Kuhl, et al. (1991). "Integrating speech information across talkers, and sensory modality: Female faces and male voices in the McGurk effect." Perception and Psychophysics **50**(6): 524-536.

Green, K. P. and L. W. Norrix (1997). "Acoustic cues to place of articulation of the McGurk effect: the role of the lease bursts, aspiration, and formant transition." Journal of Speech, Language, and Hearing Research **40**: 646-665.

Greenberg, S. (1998). "A syllabic-centric framework for the evolution of spoken language." Brain and Behavioral Science **21**: 267-268.

Grossman, E., M. Donnelly, et al. (2000). "Brain areas involved in perception of biological motion." Journal of Cognitive Neuroscience **12**: 711-720.

Gunji, A., R. Kakigi, et al. (2003). "Cortical activities relating to modulation of sound frequency: how to vocalize?" Cognitive Brain Research **17**: 495-506.

Gunji, A., S. Koyama, et al. (2003). "Magnetoencephalographic study of the cortical activity elicited by human voice." Neuroscience Letters **348**: 13-16.

Hackett, T. A., I. Stepniewska, et al. (1998). "Thalamocortical connections of the parabelt auditory cortex in macaque monkeys." The Journal of Comparative Neurology **400**: 271-286.

Hackett, T. A., I. Stepniewska, et al. (1998). "Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys." The Journal of Comparative Neurology **394**: 475-495.

Hackett, T. A., I. Stepniewska, et al. (1999). "Prefrontal connections of the parabelt auditory cortex in macaque monkeys." Brain Research **817**: 45-58.

Hall, D. A. (2003). "Auditory pathways: are 'what' and 'where' appropriate?" Curr Biol **13**(10): R406-8.

Halle, M. and K. N. Stevens (1962). "Speech Recognition: A Model and a Program for Research." I.R.E. Trans. Inf. Theory IT-8 (2), 155-159.

Handy, T. C., M. S. Gazzaniga, et al. (2003). "Cortical and subcortical contributions to the representation of temporal information." Neuropsychologia **41**: 1461-1473.

Harrington, D. L., K. Y. Haaland, et al. (1998). "Cortical networks underlying mechanisms of time perception." The Journal of Neuroscience **18**(3): 1085-1095.

Harrington, L. K. and C. K. Peck (1998). "Spatial disparity affects visual-auditory interactions in human sensorimotor processing." Experimental Brain Research **122**: 247-252.

Harth, E., K. P. Unnnikrishnan, et al. (1987). "The inversion of sensory processing by feedback pathways: a model of visual cognitive functions." Science **237**.

Hartmann, G. W. (1974). Gestalt psychology; a survey of facts and principles. Westport, Connecticut, Greenwood Press.

Hay, J.C., Pick, H.L., Jr. and Ikeda, K. (1965) “Visual cpture produced by prusm spectacles.” Psychon. Sci. **2**: 215-216.

Heil, P. (1997). "Auditory cortical onset responses revisited I. First spike-timing." Jounal of Physiology**?**(?): 2616-2641.

Heil, P. (1997). "Auditory cortical onset responses revisited II. Response strength." Journal of Physiology**?**(?): 2642-2660.

Helfer, K. S. (1997). "Auditory and auditory-visual perception of clear and conversational speech." Journal of Speech, Language, and Hearing Research **40**: 432-443.

Heller, J., J. A. Hertz, et al. (1995). "Information flow and temporal coding in primate pattern vision." Journal of Computational Neuroscience **2**: 175-193.

Hershenson, M. (1962). "Reaction time as a measure of intersensory facilitation." Journal of experimental Psychology **63**(3): 289-293.

Hikosaka, K., E. Iwai, et al. (1988). "Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey." Journal of Neurophysiology **60**(5): 1615-1637.

Hirsh, I. J. (1967). Information processing in input channels for speech and language: the significance of serial order of stimuli. Brain Mechnanisms Underlying Speech and Language.

Hirsh, I. J. (1974). Temporal order and auditory perception. Sensation and Measurement. H. R. Moskowitz. Dordrecht, Holland, D. Reidel Publishing Company**:** 251-258.

Hirsh, I. J. (1975). "Temporal aspects of hearing." The Nervous System **3**: 157-162.

Hirsh, I. J. and C. E. J. Sherrick (1961). "Perceived order in different sense modalities." Journal of Experimental Psychology **62**(5): 423-432.

Hohnsbein, J., M. Falkenstein, et al. (1991). "Effects of crossmodal divided attention on late ERP components. I. Simple and choice reaction tasks II. Error processing in choice reaction tasks." Electroencephalography and Clinical Neurophysiology **78**: 438-446.

Hopfield, J. J. and C. D. Brody (2001). "What is a moment? Transient synchrony as a collective mechanismfor spatiotemporal integration." Proceedings of the National Academy of Science **98**(3): 1282-1287.

Howard III, M. A., I. O. Volkov, et al. (1996). "A chronic microelectrode investigation of the tonotopic organization of human auditory cortex." Brain Research **724**: 260-264.

Howard, I. P. and W. B. Templeton (1966). Human Spatial Orientation. Wiley.

Howard, M. A., I. O. Volkov, et al. (2000). "Auditory cortex on the human posterior

superior temporal gyrus." Journal of Comparative Neurology **416**: 79-92.

Howard, M. W., D. S. Rizzuto, et al. (2003). "Gamma oscillations correlate with working memory load in humans." Cerebal Cortex **13**: 1369-1374.

Hubel, D. H. and T. N. Wiesel (1977). "Ferrier lecture:functional architectire of macaque monkey visual cortex." Proceedings of the Royal Society of London: series B **198**: 1-59.

Hyvärinen, J. and Y. Shelepin (1979). "Distribution of visual and somatic functions in the parietal associative area 7 of the monkey." Brain Research **169**: 561-564.

Iacoboni, M., R. P. Woods, et al. (1998). "Bimodal (auditory and visual) left frontoparietal circuitry for sensorimotor integration and sensorimotor learning." Brain **121**: 2135-2143.

Izhikevich, E. M., N. S. Desai, et al. (2003). "Bursts as a unit of neural information: selective communication via resonance." Trends in Neurosciences **26**(3): 161-167.

Ivry, R. B. and L. C. Robertson (1998) The Two Sides of Perception. Cambridge: MIT Press, Series in Cognitive Neurosciences.

Jiang, W., M. T. Wallace, et al. (2001). "Two cortical areas mediate multisensory integration in superior colliculus neurons." Journal of Neurophysiology: 506.

Jones, J. A. and K. Munhall (1997). "The effects of separating auditory and visual sources on audiovisual integration of speech." Canadian Acoustics **25**(4): 13-19.

Jordan, T. R., M. V. McCotter, et al. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. Perception and Psychophysics. **62:**1394 -1404.

Kaas, J. H. and C. E. Collins (2001). "The organization of sensory cortex." Current Opinion in Neurobiology **11**: 498-504.

Kaas, J. H. and T. A. Hackett (2000). "Subdivisions of auditory cortex and processing streams in primates." Proceedings of the National Academy of Science **97**(22).

Kaas, J. H., T. A. Hackett, et al. (1999). "Auditory processing in primate cerebral cortex." Current Opinion in Neurobiology **9**(2): 164-170.

Kanwisher, N., J. McDermott, et al. (1997). "The fusiform face area: a module in human extrastriate cortex specialized for face perception." Journal of Neuroscience **17**: 4302-4311.

Keysers, C., E. Kohler, et al. (2003). "Audiovisual mirror neurons and action recognition." Exp Brain Res **153**(4): 628-36.

Kinchla, R. A., J. Townsend, et al. (1966). "Influence of correlated visual cues on auditory signal detection." Perception and Psychophysics **1**: 67-73.

Klucharev, V., R. Möttönen, et al. (2003). "Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception." Cogniitive Brain Research **18**: 65-75.

Kohler, E., C. Keysers, et al. (2002). "Hearing sounds, understanding actions: action representation in mirror neurons." Science **297**(5582): 846-8.

Kristofferson, A. B. (1980). "A quantal step function in duration discrimination." Perception and Psychophysics **27**(4): 300-306.

Kubovy, M. and D. Van Valkenburg (2001). "Auditory and visual objects." Cognition **80**(1-2): 97-126.

Kuhl, P. and A. N. Meltzoff (1984). "The intermodal representation of speech in infants." Infant Behavior and Development **7**: 361-381.

Kuriki, S. and M. Murase (1989). "Neuromagnetic study of the auditory responses in right and left hemispheres of the human brain evoked by pure tones and speech sounds." Experimental Brain Research **77**: 127-134.

Kubovy, M., and Van Valkenburg, D. (2001). Auditory and visual objects. Cognition **80**: 97–126.

Lachaux, J.-P., E. Rodriguez, et al. (1999). "Measuring phase synchrony in brain signals." Human Brain Mapping **8**: 194-208.

Làdavas, E. and F. Pavani (1998). "Neuropsychological evidence of the functional integration of visual, auditory and proprioceptive spatial maps." Neuroreport **9**(1195): 1200.

Laurienti, P. J., J. H. Burdette, et al. (2002). "Deactivation of sensory-specific cortex by cross-modal stimuli." Journal of Cognitive Neuroscience **14**(3): 420-429.

Lebib, R., D. Papo, et al. (2003). "Evidence of a visual-to-auditory cross-modal

sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation." Neuroscience Letters **341**: 185-188.

Lebib, R., D. Papo, et al. (2003). Early processing of visual speech information modulates processing of auditory speech input at a pre-attentive level: evidence from event-related potential data. Proceedings of AVSP, St Jorioz, France.

Lee, K.-H., L. M. Williams, et al. (2003). "Synchronous gamma activity: a review and contribution to an integrative neuroscience model of schizophrenia." Brain Research Reviews **41**: 57-78.

Leinonen, L., J. Hyvärinen, et al. (1980). "Functional properties of the neurons in the temporo-parietal association cortex of awake monkey." Experimental Brain Research **39**: 203-215.

Lestienne, R. (2001). "Spike timing, synchronization and information processing on the sensory side of the central nervous system." Progress in Neurobiology **65**: 545-591.

Leug, L. S. (1997). "Generation of theta and gamma rhythms in the hippocampus." Neuroscience and Behavioral Reviews **22(**2): 275-290.

Lewald, J., W. H. Ehrenstein, et al. (2001). Spatio-temporal constraints for auditory-visual integration. Behavioural Brain Research. **121:** 69-79.

Lewicki, M. S. (2002). "Efficient coding of natural sounds." Nature Neuroscience **5(**4): 356-363.

Lewkowicz, D. J. (1999). The development of temporal and spatial intermodal perception. Cognitive contributions to the perception of spatial and temporal events. T. B. Gisa Aschersleben, Jochen Müsseler. Amsterdam, Elsevier. **129:** 395-420.

Lewkowicz, D. J. (2000). "The development of intersensory temporal perception: an epigenetic systems/limitations view." Psychological Bulletin **126(**2): 281-308.

Lewkowicz, D. J. (2002). "Heterogeneity and heterochrony in the development of intersensory percpetion." Cognitive brain Research **14**: 41-63.

Lewkowicz, D. J. and G. Turkewitz (1980). "Cross-modal equivalence in early infancy: auditory-visual intensity matching." Developmental Psychology **16**:

597-607.

Liberman, A. M., F. S. Cooper, et al. (1967). "Perception of the speech code." Psychological Review **74**(6): 431-461.

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

Liberman, A. M. and D. H. Whalen (2000). "On the relation of speech to language." Trends in Cognitive Sciences **4**(5): 187-195.

Liégeois, C., J. B. de Graaf, et al. (1999). "Specialization of left auditory cortex for speech perception in man depends on temporal coding." Cerebral Cortex **9**: 484-496.

Liégeois-Chauvel, C., A. Musolino, et al. (1991). "Localization of the primary auditory area in man." Brain **114**: 139-153.

Lopes, d. S. F. (1991). "Neural mechanisms underlying brain waves: from neural membranes to networks." Electroencephalography and clinical neurophysiology **19**: 81-93.

Lovelace, C. T., B. E. Stein, et al. (2003). "An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection." Cognitive Brain Research **17**: 447-453.

Loveless, N., S. Levänen, et al. (1996). "Temporal integration in auditory sensory memory: neuromagnetic evidence." Electroencephalography and Clinical Neurophysiology **100**: 220-228.

Ludman, C. N., A. Q. Summerfield, et al. (2000). "Lip-reading ability and patterns of cortical activation studied using fMRI." British Journal of Audiology(34): 225-230.

Luo, Y.-j. and J.-h. Wei (1999). "Cross-modal selective attention to visual and auditory stimuli modulates endogenous ERPs components." Brain Research **842**: 30-38.

Macaluso, E., Frith, C.D., & Driver, J. (2000). “Modulation of human visual cortex by

crossmodal spatial attention. Science **289**: 1206-1208.

Macaluso, E., C. D. Frith, et al. (2002). "Crossmodal spatial influences of touch on

extrastriate visual areas take current gaze direction into account." Neuron **34**: 647-658.

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

MacDonald, J., A. Soren, et al. (2000). Hearing by eye: how much spatial degradation can be tolerated? Perception. **29:**1155- 1168.

Mackick, S. L. and M. S. Livingston (1998). "Neuronal correlates of visibility and invisibility in the primate visual system." Nature Neuroscience **1**: 144-149.

MacLeod, A. and Q. Summerfield (1990). "A procedure for measuring auditory and audio-visual speech perception thresholds for sentences in noise: rationale, evaluation, and recommendations for use." British Journal of Audiology **24**: 29-43.

Maioli, M. G., C. Galletti, et al. (1984). "Projections from the cortex of the superior temporal sulcus to the dorsal lateral geniculate and pregeniculate nuclei in the macaque monkey." Archives Italiennes de Biologie **122**: 301-309.

Maiste, A. C., A. S. Wiens, et al. (1995). "Event-related potentials and the categorical perception of speech sounds." Ear and Hearing **16**(1): 68-90.

Marinkovic, K., R. P. Dhond, et al. (2003). "Spatiotemporal dynamics of modality-specific and supramodal word processing." Neuron **38**: 487-497.

Marks, L. (1978). The unity of the senses: Interrelations among the modalities. New York, Academic Press.

Massaro, D. W. (1987). Speech Perception by Ear and Eye: a paradigm for psychological inquiry. Hillsdale, Lawrence Erlbaum Associates, Inc.

Massaro, D. W. (1998). Perceiving Talking Faces. Cambridge, MIT Press.

Massaro, D. W., M. M. Cohen, et al. (1996). "Perception of asynchronous and conflicting visual and auditory speech." Journal of the Acoustical Society of America **100**(3): 1777-1786.

Maunsell, J. H. R., J. M. Ghose, et al. (1999). "Visual response latencies in the magnocellular and parvocellular LGN neurons in macaque monkeys." Visual Neuroscience **16**: 1-14.

McDonald, J. J. and L. M. Ward (2000). "Involuntary listening aids seeing: evidence

from human electrophysiology." Psychological Science **11**(2): 167-171.

McGrath, M. and Q. Summerfield (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America. **77:** 678-684.

McGurk, H. and J. MacDonald (1976). "Hearing lips and seeing voices." Nature **264**: 746-747.

Mehta, A. D., I. Ulbert, et al. (2000). "Intermodal selective attention in monkeys. II: physiological mechanisms of modulation." Cereb Cortex **10**(4): 359-70.

Mehta, B. and S. Schaal (2002). "Forward models in visuomotor control." J Neurophysiol **88**(2): 942-53.

Mehta, M. R. (2001). "Neuronal dynamics of predictive coding." Neuroscientist **7**(6): 490-5.

Mehta, M. R., A. K. Lee, et al. (2002). "Role of experience and oscillations in transformin a rate code into a temporal code." Nature **417**: 741-746.

Meltzoff, A. N. (1999). "Origins of theory of mind, cognition and communication." Journal of Communication Disorders **32**: 251-26.

Meredith, A. M. and H. R. Clemo (1989). "Auditory cortical projection from the anterior ectosylvian sulcus (field AES) to the superior colliculus in the cat: and anatomical and electrophysiological study." Journal of Comparative Neurology **289**: 687-707.

Meredith, A. M., J. W. Nemitz, et al. (1987). "Determinants of multisensory interaction in superior colliculus neurons I. temporal factors." The Journal of neuroscience **7**(10): 3215-3229.

Meredith, M. A. (2002). "On the neuronal basis of multisensory convergence: a brief overview." Cognitive Brain Research **14**: 31-40.

Meredith, M. A. and B. E. Stein (1983). "Interactions among converging sensory inputs inthe superior colliculus." Science **221**: 389-391.

Mesulam, M. M. (1998). "From sensation to cognition." Brain **121**: 1013-1052.

Meyer, G. F., Wuerger, S.M. (2001). "Cross-modal integration of auditory and visual motion signals." Neuroreport **12**(11): 2557-2560.

Miall, R. C. (2003). "Connecting mirror neurons and forward models." Neuroreport

14(16): 1-3.

Molfese, D. L. (1978). "Neuroelectrical correlates of categorical speech perception in adults." Brain and Language **5**: 25-35.

Molfese, D. L. (1978). "Left and right hemisphere involvement in speech perception: electrophysiological correlates." Perception and Psychophysics **23**(3): 237-243.

Molfese, D. L. (1980). "Hemispheric specialization of temporal information: implications of voicing cues during speech perception." Brain and Language **11**: 285-299.

Molfese, D. L. (1984). "Left hemisphere sensitivity to consonant sounds not displayed by the right hemisphere: electrophysiological correlates." Brain and Language **22**: 109-127.

Molfese, D. L., R. A. Burhke, et al. (1985). "The right hemisphere and temporal processing of consonant transitions durations: electrophysiological correlates." Brain and Language **26**: 49-62.

Molholm, S., W. Ritter, et al. (2002). "Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study." Cognitive Brain Research **14**: 115-128.

Moore, D. R. (2000). "Auditory neuroscience: is speech special?" Current Biology **10**(10): R362-R364.

Morein-Zamir, S., S. Soto Faraco, et al. (2003). "Auditory capture of vision: examining temporal ventroiloquism." Brain Research.

Möttönen, R., C. M. Krause, et al. (2002). "Processing of changes in visual speech in the human auditory cortex." Cognitive Brain Research **13**: 417-425.

Munhall, K., P. Gribble, et al. (1996). "Temporal constraints the Mcgurk effect." Perception and Psychophysics **58**(3): 351-362.

Munhall, K. and Y. Tohkura (1998). "Audiovisual gating and the time course of speech perception." Journal of the Acoustical Society of America **104**(1): 530-539.

Näätänen, R. (1992). Attention and brain function. Hillsdale, NJ, Lawrence Erlbaum Ass.

Näätänen, R. (1995). "The mismatch negativity: a powerful tool for cognitive neuroscience." Ear & Hearing **16**: 6-18.

Näätänen, R. and T. Picton (1987). "The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure." Psychophysiology **24**(4): 375-422.

Nelken, I., Y. Rotman, et al. (1999). "Responses of auditory-cortex neurons to structural features of natural sounds." Nature **397**: 154-157.

Neville, H. J., D. Bavelier, et al. (1998). "Cerebral organization for language in deaf and hearing subjects: biological constraints and effects of experience." Protocols of the National Academy of Sciences **95**: 922-929.

Nyberg, L., R. Habib, et al. (2000). "Reactivation of encoded related brain activity during memory retrieval." Proceedings of the National Academy of Science **97**(20): 11120-11124.

Oram, M. W. and D. I. Perett (1996). "Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey." Journal of Neurophysiology **76**(1): 109-129.

Oram, M. W., D. Xiao, et al. (2002). "The temporal resolution of neural codes: does response latency have a unique role?" Proceedings of the Royal Society of London series B **357**: 987-1001.

Oray, S., Z.-L. Lu, et al. (2000). "Modification of sudden onset auditory ERP by involuntary attention to visual stimuli." International Journal of Psychophysiology **43**: 213-224.

Otten, l. J., C. Alain, et al. (2000). "Effects of visual attentional load on auditory procesing." Neuroreport **11**(4): 875-880.

Pandey, P. C., H. Kunov, et al. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. The Journal of Auditory Research. **26:** 27-41.

Pandya, D. N. (1995). "Anatomy of the auditory cortex." Rev. Neurol. **151**(8-9): 486-494.

Pandya, D. N. and B. Seltzer (1982). "Associations areas of the cerebral cortex." Trends in Cognitive Sciences(Nov): 386.

Pandya, D. N. and B. Seltzer (1982). "Association areas of the cerebral cortex." Trends in Neuroscience: 386-390.

Pantev, C., M. Hoke, et al. (1988). "Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields." Electroencephalography and Clinical Neurophysiology **69**: 160-170.

Panzeri, S., S. R. Schultz, et al. (1999). "Correlations and the encoding of information in the nervous system." Proceedings of the Royal Society of London series B **266**: 1001-1012.

Paré, M., R. C. Richler, et al. (2003). "Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect." Perception and Psychophysics **65**(4): 553-567.

Patton, P., K. Belkacem-Boussaid, et al. (2002). "Multimodality in the superior colliculus: an information theoric approach." Cognitive Brain Research **14**: 10-19.

Penner, M. J. (1977). "Detection of temporal gaps in noise as a measure of the decay of auditory sensation." J Acoust Soc Am **61**(2): 552-7.

Pessoa, L., S. Kastner, et al. (2003). "Neuroimaging studies of attention: from modulation of sensory processing to top-down control." The Journal of Neuroscience **23**(10): 3990-3998.

Philips, C., T. Pellathy, et al. (2000). "Auditory cortex accesses phonological categories: an MEG mismatch study." Journal of Cognitive Neuroscience **12**.

Pickett (1999). The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology, MA. Allyn & Bacon.

Pisoni, D. B. and P. A. Luce (1986). Trading relations, acoustic cue integration and context effects in speech perception. The Psychophysics of Speech Perception. M. E. H. Schouten and B. C. J. Moore.

Plomp R. (1964). “Rate of decay of auditory sensation”. J Acoust Soc Am **36**: 277-282.

Poeppel, D. and A. Marantz (2000). Cognitive neuroscience of speech processing. In Miyashita, Marantz, O’Neil (eds). Image, language, brain. Cambridge, MA: MIT Press.

Poeppel, D. (2001). “Pure word deafness and the bilateral processing of the speech code.” Cognitive Science **21** (5): 679-693.

Poeppel, D. (2003). “The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’” Speech Communication **41**: 245-255.

Poremba, A., R. C. Saunders, et al. (2003). "Functional mapping of the primate auditory system." Science **299**: 568-572.

Price, C. J. (2000). "The anatomy of language: contributions from functional neuroimaging." Journal of Anatomy **197**: 335-359.

Puce, A., T. Allison, et al. (1998). Temporal cortex activation in humans viewing eye and mouth movements. The Journal of Neuroscience. **18:** 2188-2199.

Radeau, M. (1974). "Adaptation au déplacement prismatique sur la base d'une discordance entre la vision et l'audition." Année Psychologique **74**: 23-34.

Radeau, M. (1985). "Signal intensity, task context, and auditory-visual interactions." Perception **4**: 571-577.

Radeau, M. and P. Bertelson (1987). "Auditory-visual interaction and the timing of inputs -*Thomas revisited (1941)*." Psychological Research **49**: 17-22.

Raij, T., K. Uutela, et al. (2000). "Audiovisual integration of letters in the human brain."Neuron **28**(617-625).

Rao, R. P. N. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." Nature Neuroscience **2**(1): 79-87.

Rauschecker, J. P. (1998). "Parallel processing in the auditory cortex of primates." Audiology Neurootology **3**: 86-103.

Rauschecker, J. P. (1998). "Cortical processing of complex sounds." Current Opinion in Neurobiology **8**: 516-521.

Rauschecker, J. P. (2000). "Auditory cortical plasticity: a comparison with other sensory systems."

Rauschecker, J. P. and B. Tian (2000). "Mechanisms and streams for processing of "what" and "where" in auditory cortex." Proceedings of the National Academy of Science **97**(22): 97.

Rauschecker, J. P., B. Tian, et al. (1995). "Processing of complex sounds in the macaque nonprimary auditory cortex." Science **268**: 111-114.

Rauschecker, J. P., B. Tian, et al. (1997). "Serial and parallel processing in rhesus monkey auditory cortex." The Journal of Comparative Neurology **382**: 89-103.

Recanzone, G. H. (1998). "Rapidly induced auditory plasticity: The ventriloquism aftereffect." PNAS **95**: 869-875.

Recanzone, G. H. (2003). "Auditory influences on visual temporal rate perception." Journal of Neurophysiology **89**: 1078-1093.

Repp, B. H. and S. Bentin (1984). "Parameters of spectral/temporal fusion in speech perception." Perception and Psychophysics **36**(6): 523-530.

Rieke, F., D. Warland, et al. (1999). Spike Exploring the neural code. Cambridge, Massachussets, London, England.

Ritz, R. and T. S. Sejnowski (1997). "Synchronous oscillatory activity in sensory systems: new vistas on mechanisms." Current Opinion in Neurobiology **7**: 536-546.

Robert-Ribes, J., J.-L. Schwartz, et al. (1998). "Complemantarity and synergy in bimodal speech; Auditory, visual, and audio-visual identification of French oral vowels in noise." Journal of the Acoustical Society of America **103**(6): 3677-3689.

Roberts, M. and Q. Summerfield (1981). "Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory." Perception and Psychophysics **30**(4): 309-314.

Rockland, K. S. and H. Ojima (2003). "Multisensory convergence in calcarine visual areas in macaque monkey." International Journal of Psychophysiology **50**(2003).

Rodriguez, E., N. George, et al. (1999). "Perception's shadow: long-distance synchronization of human brain activity."

Romo, R., A. Hernández, et al. (2003). "Correlated neuronal discharges that increase coding efficiency during perceptual discrimination." Neuron **38**: 649-657.

Rosen, S. (1992). "Temporal information in speech: acoustic, auditory and linguistic

aspects." Philosophical Transaction of the Royal Society of London, series B **336**: 367-373.

Rosenblum, L., M. A. Schmuckler, et al. (1997). The McGurk effect in infants. Perception and Psychophysics. **59:**347 -357.

Rosenblum, L. and D. A. Yakel (2001). "The McGurk effect from single and mixed speaker stimuli." Acoustics Research Letters Online **2**(2): 67-72.

Rosenblum, L. D. and C. A. Fowler (1991). "Audiovisual investigation of the loudness-effort effect for speech and nonspeech events." Journal of Experimental Psychology: Human Perception and Performance **17**(4): 976-985.

Rosenblum, L. D. and H. M. Saldaña (1992). "Discrimination tests of visually influenced syllables." Perception and Psychophysics **52**(4): 461-473.

Rosenblum, L. D. and H. M. Saldaña (1996). "An audiovisual test of kinematic primitives for visual speech perception." Journal of Experimental Psychology: Human Perception and Performance **22**(2): 318-331.

Saito, Y. and T. Isa (2003). "Local excitatory network and NMDA receptor activation generate a synchronous and bursting command from the superior colliculus." The Journal of Neuroscience **23**(13): 5854-5864.

Sakowitz, O. W., R. Q. Quiroga, et al. (2001). "Bisensory stimulation increases gamma-responses over multiple cortical regions." Cognitive Brain Research **11**: 267-279.

Saldaña, H. M. and L. D. Rosenblum (1993). "Visual influences on auditory pluck and bow judgements." Perception and Psychophysics **54**(3): 406-416.

Sams, M. and R. Aulanko (1991). "Seeing speech: visual information from lip movements modifies activity in the human auditory cortex." Neuroscience Letters **127**: 141-147.

Sams, M., P. Manninen, et al. (1998). "McGurk effect in Finnish syllables,isolated words, and words in sentences: effects of word meaning and sentence context." Speech Communication **26**: 75-87.

Sauvé, K. (1999). "Gamma-band synchronous oscillations: recent evidence regarding their functional significance." Consciousness and Cognition **8**: 213-224.

Schack, B., N. M. Vath, et al. (2002). "Phase-coupling of theta-gamma EEG rhythms during short-term memory processing." International Journal of Psychophysiology **44**: 143-163.

Schacter, D. L. and R. L. Buckner (1998). "Priming and the brain." Neuron **20**: 185-195.

Schadlen, M. N. and J. A. Mocshon (1999). "Synchrony unbound: a critical evaluation of the temporal binding hypothesis." Neuron **24**: 67-77.

Scheier, C., D. J. Lewkowicz, et al. (under review). "Sound induces perceptual reorganization of an ambiguous motion display in human infants."

Schroeder, C. E., J. Smiley, et al. (2003). "Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing." International Journal of Psychophysiology **50**(2003).

Schröger, E. and A. Widmann (1998). "Speeded responses to audiovisual signal changes result from bimodal integration." Psychophysiology **35**: 755-759.

Schulman, G. L., M. Corbetta, et al. (1997). "Top-down modulation of early sensory cortex." Cerebal Cortex **7**: 193-206.

Scott, S. K. and I. S. Johnsrude (2003). "The neuroanatomical and functional organization of speech perception." Trends in Neurosciences **26**(2): 100-107.

Segalowitz, S. J. and H. Cohen (1989). "Right Hemisphere sensitivity to speech." Brain and Language **37**: 220-231.

Sekihara, K., S. S. Nagarajan, et al. (1999). "Time-frequency MEG MUSIC algorithm." IEEE transactions on medical imaging **18**(1).

Sekuler, R. and A. Sekuler (1997). "Sounds alter visual motion perception." Nature **385**: 308.

Seldon, L. H. (1984). The anatomy of speech perception. Cerebral Cortex. E. G. J. Alan Peters. New York, Plenum Press.

Sergent, J. (1984). "Processing of visually presented vowels in the cerebral hemispheres." Brain and Language **21**: 136-146.

Servos, P., R. Osu, et al. (2002). "The neural substrates of biological motion perception: an fMRI study." Cerebral Cortex **12**: 772-782.

Shams, L., Y. Kamitani, et al. (2000). "What you see is what you hear." Nature **408**:

408.

Shams, L., Y. Kamitani, et al. (2002). "Visual illusion induced by sound." Cognitive Brain Research **14**: 147-152.

Shams, L., Y. Kamitani, et al. (2001). "Sound alters visual evoked potentials in humans." Neuroreport **12**(17): 4.

Sharma, A. and M. F. Dorman (1999). "Cortical auditory evoked potential correlates of categorical perception of voice-onset time." Journal of the Acoustical Society of America **16**(2): 1078-1083.

Sharma, J., V. Dragoi, et al. (2003). "V1 neurons signal acquisition of an internal representation of stimulus location." Science **300**: 1758-1763.

Shin, J. (2002). "A unifying theory on the relationship between spike trains, EEG, and ERP based on the noise shaping/predictive neural coding hypothesis." BioSystems **67**: 245-257.

Shinozaki, N., H. Yabe, et al. (2003). "Spectrotemporal window of integration of auditory information in the human brain." Cognitive Brain Research.

Shipley, T. (1964). "Auditory flutter-driving of visual flicker." Science **145**: 1328-1330.

Shore, D. I., E. Spry, et al. (2002). "Confusing the mind by crossing the hands." Cognitive Brain Research **14**: 153-163.

Schulman, G. L., M. Corbetta, et al. (1997). "Top-down modulation of early sensory cortex." Cerebral Cortex **7**: 193-206.

Simos, P. G., D. L. Molfese, et al. (1997). "Behavioral and electrophysiological indices of voicing-cue discrimination: laterality patterns and development." Brain and Language **57**: 122-150.

Singer, W. and C. M. Gray (1995). "Visual feature integration and the temporal correlation hypothesis." Annual Reviews of Neuroscience **18**: 555-586.

Skipper, J., V. van Wassenhove, et al. (2003). Cognitive Neuroscience Society, San Francisco.

Slutsky, D. A. and G. H. Recanzone (2001). "Temporal and spatial dependency of the ventriloquism effect." Neuroreport **12**(1): 7-10.

Soroker, N., N. Calamaro, et al. (1995). ""McGurk illusion" to bilateral administration of sensory stimuli in patients with hemispatial neglect." Neuropsychologia **33**(4): 461-470.

Sparks, D. L. (1986). "Translation of sensory signals into commands for control of saccadic eye movements: role of primate superior colliculus." Physiological Reviews **66**(1): 118-171.

Spelke, E. S. (1981). "The infant's acquisition of knowledge of bimodally specified events." Journal of Experimental Child Psychology **31**: 279-299.

Spence, C., R. Baddeley, et al. (2002). "Multisensory temporal order judgments: when two locations are better than one." Perception and Psychophysics.

Spence, C., D. I. Shore, et al. (2001). "Multisensory prior entry." Journal of Experimental Psychology: General(1-39).

Spence, C. and S. Squire (2003). "Multisensory integration: maintaining the perception of synchrony." Current Biology **13**: R519-R521.

Spencer, K. M., P. G. Nestor, et al. (2003). "Abnormal neural synchrony in schizophrenia." The Journal of Neuroscience **23**(19): 7407-7411.

Stein, B. E. (1998). "Neural mechanisms for synthesizing sensory information and producing adaptive behaviors." Experimental Brain Research **123**: 124-135.

Stein, B. E. and M. O. Arigbede (1972). "Unimodal and multimodal response properties of neurons in the cat's superior colliculus." Experimental Neurology **36**: 179-196.

Stein, B. E. and H. Gallagher (1981). "Maturation of cortical control over the superior colliculus cells in cat." Brain research **223**: 429-435.

Stein, B. E., P. J. Laurienti, et al. (2000). Neural mechanisms for integrating information from multiple senses. IEEE.

Stein, B. E., N. L. London, et al. (1996). "Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis." Journal of Cognitive Neuroscience **8**(6): 497-506.

Stein, B. E. and A. M. Meredith (1993). The merging of the senses. Cambridge, MIT Press.

Steinschneider, M., J. C. Arezzo, et al. (1990). "Tonotopic features of speech-evoked

activity in primate auditory cortex." Brain Research **519**: 158-168.

Steinschneider, M., D. Reser, et al. (1995). "Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey." Brain Research **674**: 147-152.

Steinschneider, M., C. E. Schroeder, et al. (1994). "Speech-evoked activity in primary auditory cortex: effects of voice onset time." Electroencephalography and Clinical Neurophysiology **92**: 30-43.

Steinschneider, M., C. E. Schroeder, et al. (1995). "Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns." Brain and Language **48**: 326-340.

Steinschneider, M., I. O. Volkov, et al. (1999). "Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex." Journal of Neurophysiology **82**(5): 2346-2357.

Steriade, M. and F. Amzica (1996). "Intracortical and corticothalamic coherency of fast spontaneous oscillations." Proceedings of the National Academy of Science **93**: 2533-2538.

Stevens, K. N. (1960). "Toward a model of speech perception." Journal of the Acoustical Society of America **32**: 45-55.

Stevens, K. N. and M. Halle (1967). Remarks on analysis by synthesis and distinctive features. Moels for the perception of speech and visual form. W. Whaten-Dunn. Cambridge MA, MIT Press**:** 88-102.

Stone, J. V., N. M. Hunkin, et al. (2001). "When is now? Perception of simultaneity." Proceedings of the Royal Society of London series B **268**: 31-38.

Stratton, G. M. (1897). "Vision without the inversion of the retinal image." Psychological Review **4**: 341-360, 463-481.

Studdert-Kennedy, M. (1983). "On learning to speak." Human Neurobiology **2**: 191-195.

Studdert-Kennedy, M. (2002). "How did language go discrete?"

Suaseng, P., W. Klimesch, et al. (2002). "The interplay between theta and alpha oscillations in the human electroencephalogram reflects the transfer of information between memory systems." Neuroscience Letters **324**: 121-124.

Sugita, Y. and Y. Suzuki (2003). "Audiovisual perception: implicit estimation of sound-arrival time." Nature **421**: 911.

Sumby, W. and I. Pollack (1954) "Visual contributions to speech intelligibility in noise." Journal of the Acoustical Society of America (26).

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R., editors, Hearing by Eye: The Psychology of Lipreading, pages 3-51. Lawrence Earlbaum, Hillsdale, New Jersey.

Summerfield, Q. (1992). "Lipreading and audio-visual speech perception." Philosophical Transaction of the Royal Society of London, series B **335**: 71-78.

Summerfield, Q. (2000). Lipreading and audio-visual speech perception. Lipreading and audio-visual speech perception**:** 71-78.

Takegata, R. and T. Morotomi (1999). "Integrated neural representation of sound and temporal features in human auditory sensory memory: an event-related potential study." Neuroscience Letters **274**: 207-210.

Tallon-Baudry, C. (2001). "Oscillatory synchrony and human visual cognition." Jourbal of Physiology Paris **97**: 355-363.

Tallon-Baudry, C. and O. Bertrand (1999). "Oscillatory gamma activity in humans and its role in object representation." Trends in Cognitive Sciences.

Tastevin, J. (1937). En partant de l'expérience d'Aristote. L'encéphale**, 1**: 57-84.

Teder-Sälejärvi, W. A., J. J. McDonald, et al. (2002). "An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recrordings." Cognitive Brain Research **14**: 106-114.

Tesche, C. D. and J. Karhu (1999). "Interactive processing of sensory input and motor output in the human hippocampus." Journal of Cognitive Neuroscience **11**(4): 424-436.

Thomas, G.J. (1940). "Experimental study of the influence of vision on sound localization." J. Exp. Psychol. **28**: 163 -177.

Tiippana, K., Andersen., T. and M. Sams (2004). "Visual attention modulates audiovisual speech perception." European Journal of Cognitive Psychology

16: 457-472.

Tononi, G., G. M. Edelman, et al. (1998). "Complexity and coherency: integrating information in the brain." Trends in Cognitive Sciences **2**(12): 474-484.

Treisman, A. (1996). "The binding problem."Current Opinion in Neurobiology **6**: 171-178.

Treisman, M., N. Cook, et al. (1994). "The internal clock: electroencephalographic evidence for oscillatory processes underlying time perception." The Quaterly Journal of Experimental Psychology **47A**(2): 241-289.

Treisman, M., A. Faulkner, et al. (1990). "The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency." Perception **19**: 705-743.

Ungerleider, L. G. and Mishkin, M. (1982) “Two cortical visual systems”. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Eds.), Analysis of Visual Behavior. The MIT Press: Cambridge, Mass. 1982, pp. 549-586.

Vaina, L. M., J. Solomon, et al. (2001). "Functional neuroanatomy of biological motion perception in humans." Proceedings of the National Academy of Science **98**(20): 11656-11661.

VanRullen, R. and C. Koch (2003). "Is perception discrete or continous?" Trends in Cognitive Science **7**(5): 207-213.

VanRullen, R. and C. Koch (2003). "Visual selective behavior can be triggered by a feed-forward process." Journal of Cognitive Neuroscience **15**(2): 209-217.

Vatikiotis-Bateson, E. Audio-visual speech production: some issues for recognition.

Vatikiotis-Bateson, E., I.-M. Eigsti, et al. (1998). "Eye movement of perceivers during audiovisual speech perception." Perception and Psychophysics **60**(6): 926-940.

Viemeister, N. F. and G. H. Wakefield (1991). "Temporal integration and multiple looks." Journal of the Acoustical Society of America **90**(2): 858-865.

Vitkovich, M. and P. Barber (1994). "Effect of video frame rate on subject's ability to shadow one of two competing verbal passages." Journal of Speech and Hearing Research **37**(5): 1204-1212.

von der Malsburg, C. (1995). "Binding in models of perception and brain function."

Current Opinion in Neurobiology **5**: 520-526.

von Stein, A., P. Rappelsberger, et al. Synchronization between temporal and parietal cortex during multimodal object processing in man. Cerebral Cortex. **9:** 137-150.

von Stein, A. and J. Sarnthein (2000). "Different frequencies for different scales of cortical integration: from local gamma to long range alpha/theta synchronization." International Journal of Psychophysiology **38**: 301-313.

Wada, Y., N. Kitagawa, et al. (2003). "Audio-visual integration in temporal perception." International Journal of Psychophysiology **20**: 117-124.

Walker, S., V. Bruce, et al. (1995). "Facial identity and facial speech processing: familiar faces and voices in the McGurk effect." Percept Psychophys **57**(8): 1124-33.

Walker-Andrews, A. S. (1986). "Intermodal perception of expressive behaviors: relation of eye and voice?" Developmental psychology **22**: 373-377.

Wallace, M. T., M. A. Meredith, et al. (1998). "Multisensory integration in the superior colliculus of the alert cat." Journal of Neurophysiology **80**(2): 1006-1009.

Wallace, M. T., R. Ramachandran, et al. (2004). "A revised view of sensory cortical parcellation." Proceedings of the National Academy of Science **101**(7): 2167-2172.

Wallace, M. T. and B. E. Stein (1997). "Development of multisensory neurons and multisensory integration in cat superior colliculus." The Journal of Neuroscience **17**(7): 2429-2444.

Wallace, M. T. and B. E. Stein (2000). "Onset of cross-modal synthesis in the neonatal superior colliculus is gated by the development of cortical influences." Journal of Neurophysiology **83**(6): 3578-3582.

Wang, X.-J. and G. Buzsàki (1996). "Gamma oscillations by synaptic inhibition in a hippocampal interneuronal network model." Journal of Neuroscience **16**(20): 6402-6413.

Watanabe, J. and E. Iwai (1991). "Neuronal activity in visual, auditory, and polysensory areas in the monkey temporal cortex during visual fixation task."

Brain Research Bulletin **26**: 583-592.

Welch, R. B. (1999). Meaning, attention and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. Cognitive contributions to the perception of spatial and temporal events. T. B. Gisa Aschersleben, Jochen Müsseler. Amsterdam, Elsevier. **129:** 371-387.

Welch, R. B., L. D. DuttonHurt, et al. (1986). "Contributions of audition and vision to temporal rate perception." Perception and Psychophysics **39**(4): 294-300.

Welch, R. B. and D. H. Warren (1980). "Immediate Perceptual Response to Intersensory Discrepancy." Psychological Bulletin **88**(3): 638-667.

Winkler, I. and I. Czigler (1998). "Mismatch negativity: deviance detection or the maintenance of the 'standard'." Neuroreport **9**(17): 3809-13.

Winkler, I., O. Korzyukov, et al. (2002). "Temporary and long term retention of acoustic information." Psychophysiology **39**: 530-534.

Wolpert, D. M., Z. Ghahramani, et al. (1995). "An internal model of sensorimotor integration." Science **269**.

Wolpert, D. M., Z. Gharamani, et al. "Forward dynamic models in human motor control: psychophysical evidence."

Wolpert, D. M., C. R. Miall, et al. (1998). "Internal models in the cerebellum." Trends in Cognitive Sciences **2**: 9.

Woods, D. L., C. Alain, et al. (1993). "Frequency-related differences in the speed of human auditory processing." Hearing Research **66**: 46-52.

Woods, D. L., K. Alho, et al. (1992). Intermodal selective attention I. Effects on event-related potentials to lateralized auditory and visual stimuli. Electroencephalography and Clinical Neurophysiology. **82:**341- 355.

Woods, D. L., K. Alho, et al. (1993). "intermodal selective attention: evidence for processing in tonotopic auditory fields." Psychophysiology **30**: 287-295.

Wright, T. M., K. A. Pelphrey, et al. (2003). "Polysensory interactions along lateral temporal regions evoked by audiovisual speech." Cerebral Cortex **13**: 1034-1043.

Yabe, H., M. Tervaniemi, et al. (1998). "Temporal window of integration of auditory information in the human brain." Psychophysiology **35**: 615-619.

Yordanova, J., V. Kolev, et al. (2002). "Wavelet entropy analysis of event-related potentials indicates modality-independent theta dominance." Jounal of Neuroscience Methods **117**: 99-109.

Yost, W. A. (1991). "Auditory image perception and analysis: the basis for hearing." Hearing Research **56**: 8-18.

Yvert, B., A. Crouzeix, et al. (2001). "Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in Humans." Cerebal Cortex **11**: 411-423.

Yvert, B., C. Fischer, et al. (2002). "Simultaneous intracerebral EEG recordings of early auditory thalamic and cortical activity in human." European Journal of Neuroscience **16**: 1146-1150.

Zampini, M., S. Guest, et al. (2002). "Audiovisual simultaneity judgments." Quaterly Journal of Experimental Psychology: section A.

Zeki, S., R. J. Perry, et al. (2003). "The processing of kinetic contours in the brain." Cerebal Cortex **13**: 189-202.

Zhou, Y.-D. and J. M. Fuster (2000). "Visuo-tactile cross-modal associations in cortical somatosensory cells." Proceedings of the National Academy of Science **97**(17): 9777-9782.

Zwiers, M. P., A. J. Van opstal, et al. (2003). "Plasticity in human sound localization induced by compressed spatial vision." Nature Neuroscience **6**(2): 175-181.