MULTISENSORY INTEGRATION IN SPEECH AND MOTION PERCEPTION

By

Lars Arne Ross

Submitted to the Graduate Faculty in Psychology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
The City University of New York
2008

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirements for the degree of doctor of Philosophy.

Executive Officer:	Thesis Advisor:
_____	_____
Signature	Signature
Maureen O'Connor	John J. Foxe
_____	_____
Name Date	Name Date

Supervisory Committee

Charles Schroeder
Robert Melara
Dave Saint-Amour
Mark McCourt

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

MULTISENSORY INTEGRATION IN SPEECH AND MOTION PECEPTION

BY

LARS ARNE ROSS

This thesis is concerned with the interaction of the senses in the perception of events in the external environment. It is organized in four parts. In the first part I (chapter I) will give a general introduction into the field of multisensory integration. The second part (Chapter II and III) comprises two behavioral studies that explore the benefit of the exposure to visual articulation in the perception of speech in noisy environmental conditions. In the first study (chapter II) we replicate the substantial benefit that visual articulation ads to the perception of speech, but also show that this benefit does not obey a previously established (and commonly accepted) rule of multisensory integration called "*the principle of inverse effectiveness."* The first study will lay the ground- work for the second clinical investigation (Chapter III) where I will show that patients with schizophrenia do not experience the same benefit from visual speech as healthy controls and that this multisensory decrement is especially pronounced in situations where controls usually experience the greatest benefit.

The third section of this thesis (Chapter IV) will report an investigation of the effect of auditory stimulation in the perception of dynamic visual events. We

used the “bistable motion illusion”, where a brief sound affects the perception of the motion direction of two moving visual objects, as a paradigm to study the effects of the auditory stimulus on visual processing. The use of event-related potentials allowed insight into the spatio-temporal dynamics of brain processes underlying this illusion.

Finally, in the fourth and last part of this thesis (Chapter V) I will give an overview of selected projects that directly or indirectly emerged from the work of this thesis. These projects are at various stages of progress and are performed with my direct participation. This section is meant as an “outlook” to my present and future scientific work in the domain of multisensory integration and are supposed to serve as a guide for the reader to understand the thesis work in a larger context.

DEDICATION

Ich widme die Frucht der beruflichen Arbeit meiner „New Yorker Jahre“ zwischen

Dezember 2001 und Maerz 2008

meiner Familie,

meinem viel zu frueh verstorbenen Bruder Christian

und meiner besten Freundin Andie Iatrou.

ACKNOWLEDGEMENTS

I have had the privilege to work with a group of people, extraordinary in their skill and character alike, and many of which are now among my best friends. My first, sincerely felt gratitude goes to my supervisor and friend John Foxe who has been substantially influential in my development as a scientist and person over the last five years. Due to his mentorship and company I can say that the journey towards my graduation was rich in invaluable (both easy and difficult) professional and personal lessons. I am deeply indebted to him and ineffably grateful to have met his wonderful family.

Let me thank my friends and colleagues Dave Saint-Amour, Simon Kelly, Manuel Gomez-Ramirez and Sophie Molholm for all their help, their scientific and technical skill and company. Let me also thank the students, post-docs and technicians at the cognitive neuroscience laboratories at NKI and CCNY. I wish them the best for their personal and professional future.

My sincere appreciation also goes toward my previous supervisors Arthur Spielman, John Antrobus and David Lewkowicz for teaching me about their field of research and for broadening my scope. My mentors at the Humboldt-University Berlin Werner Sommer and Joerg Sangals deserve my gratitude for introducing me to the field of Cognitive Neuroscience and electrophysiology and from which I learned to conduct electrophysiological experiments. Also in

Germany, Gernot Laemmler deserves my gratitude for his excellent supervision and training in neuropsychological testing during my clinical internship.

My tanks also go to Ekkehart Trenkner at the College of Staten Island who arranged for and provided me with much needed financial support in my early days as a doctoral student in New York.

Finally, I thank my best friend Andie Iatrou for her tireless work on the formatting of the thesis, and her patience with me during the final steps towards its completion.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER II

CHAPTER III

CHAPTER IV

LIST OF FIGURES

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

CHAPTER I

GENERAL INTRODUCTION

1. Multisensory integration in human experience- some historical notes.

For human kind, one of the most fascinating questions has been how we and other living species perceive and represent the "outside world." The curious observer is ineluctably fascinated by the intricate complexity of the sensory systems and their power to represent the environment. Tuned to different kinds of environmental energy, each system provides a qualitatively distinct sensory specific perceptual experience (quale). The modality-specificity of sensation and perception may be, in fact, one of the reasons why the effects of multisensory (multisensory) stimulation have long been overlooked. Watching a speaker's articulations, for instance, does not change the subjective experience of spoken speech as being auditory in nature, even though many scientists now believe that the underlying representation of speech is amodal. Only in rare situations, usually under artificial, often experimental circumstances, incongruent or asynchronous cross-sensory combinations raise awareness about the impact of multisensory stimulation. Such a situation may arise when we lower the volume on the TV during a news show and attempt to lip- or speechread from the speaker's face. In this case it often appears as if we "hear" the words that we attempt to identify from the movements of the lips, despite the lack of any auditory input. The reception of input in one modality (visual) creates a sensory impression in another modality (auditory). Or consider the experience of a rollercoaster ride by watching it on one of these movie theaters with all-around

screens. The vestibular and visceral experience delivered by the visual modality alone (and too a lesser extent by the auditory modality) can be nearly as strong as on the rollercoaster itself.

One of the most well-known multisensory illusions, the McGurk effect was discovered accidentally in the 1970's by Harry McGurk who was interested in whether auditory and visual modalities were differentially dominant during an infant's perceptual development. McGurk asked his technician to create a film where an actor pronounced the syllable "ga" (velar) during an auditory recording of a consonant "ba" (bilabial). After initially suspecting a technical error they found that adults tend to perceive "da" as a result of combining the visually and auditorily presented information. The subsequent publication "Hearing lips and seeing voices" in the journal *Nature* became one of the most influential publications in the field of multisensory integration (McGurk and MacDonald, 1976) and the McGurk effect, in its several variations, became one of the most studied paradigms in the disciplines of speech and language science, psychology, neuroscience, and engineering.

Under normal conditions, incongruencies between sensory modalities are rare, if not absent. In a curious condition called "synesthesia", the affected individual reports an involuntary, automatic sensory experiences during the exposure to particular stimuli in another sensory modality. For instance, letters, words, digits or music may evoke vivid experiences of color when they are seen or heard. The phenomenon has been reported since the 1700's, and in 1812 "color hearing" ("audition coloré") was described by Georg Sachs as a medical

pathology (Mahling, 1926). Later, it was believed that synesthesia was a normal condition of the brain and that these cases are just extreme cases of a general law that states that "*all our sense- organs influence each other's sensations*" (William James, 1890, Vol. II, page 29).

Later, it became fashionable in the polite society to have synesthetic experiences. In the second half of the nineteenth century synesthesia was investigated more systematically and early investigators asked whether there was a systematic mapping between synesthetes cross-sensory associations, a notion that was later rejected (Marks, 1978). In what is probably the earliest experimental investigation of cross-sensory processes, the audiologist Urbantschitsch (1888) showed quite remarkably that visual stimuli near or below the perceptual threshold became visible when a sound generated by a tuning fork was presented. In one of the many variations of this experiment, he also found that near- threshold sounds became audible when paired with a visual (light) stimulus. This and other findings led the early visionary psychologist William James to believe that synesthesia was a physiological phenomenon resulting from a unified and Gestalt-like operation of the senses. The following is a suggestion by William James for an investigative approach of the study of multisensory processes[1]:

[1] Of course James had no elaborate concept of what we call "multisensory integration" today. However, it is evident in his writings that he had at least an intuitive understanding that integrative processes of sensory modalities are an inherent physiological property of the brain.

"You cannot build up one thought or one sensation out of many; and only direct experiment can inform us of what we shall perceive when we get many stimuli at once." (Vol 2, page 30, 31).

The early interest in synesthesia triggered a lively debate among early scientists who discussed whether the synesthetic experience reflected a direct sensory experience or whether it was accomplished through an "association" by mental processes. These opposing ideas foreshadowed the later (and still active) debate between a modality–specific view and an approach favoring the idea of "unified" or "supramodal sensory operations." The notion of the "unity of the senses" was further explored by Gestalt psychologists such as Von Hornbostel and Werner (Von Hornbostel, 1927) and experimentally investigated by Zietz (1931). In an early multisensory experiments, Zietz asked participants about their perceptions of afterimages of photic stimuli in a dark room during the presentation of varying auditory stimuli.

After having experienced a "dip" in interest, synesthesia was subject to a renaissance over the last three decades, especially since the invention of new brain imaging technology (e.g. Paulesu et al. 1995). While the neuronal origins of synesthetic experience are still unknown, synesthesia may hold important cues for the understanding of normal multisensory integration in humans and is therefore is an integral part of current investigations in this domain.

In the example of ventriloquism we can see that the artificially manipulated disparity of sound and motion produce the impression of the dislocation of sound.

Ventriloquism was considered an act of “witchcraft” in ancient times as the ventriloquist alleged to communicate with the dead through his/her stomach. Starting in the 16th century this became a performance art that reached its peak popularity in the middle of the last century. During this stage-performance a person manipulated his/her voice and articulation so that it appeared as if the voice originated from a puppet who’s gestures and articulations were actually operated by the ventriloquist. Only much later was it appreciated that a fundamental integrative mechanism in the brain underlied this illusion (Howard and Templeton, 1966). When a visual stimulus is presented simultaneously with an auditory stimulus, but at a disparate spatial location, the resulting percept is that both stimuli originate from the location of the visual stimulus.

In our experience it is rare to receive information about events through only one sensory channel. This ubiquity of the multisensory events may be another reason why the converging attributes of the brain usually go unnoticed by the perceiver. This, in turn, may have contributed to the historical negligence to study of the interaction of the senses. On the contrary, scientific endeavor has tried to understand the workings of sensation and perception by focusing on the investigation of each sensory system in isolation. However, research over the past few decades through an exponentially growing number of experiments and publications across disciplines, has consistently revealed that multisensory stimulation affects neural mechanisms in ways that are inconsistent with the notion that the sensory systems operate independently of one another. Vertebrate brains have apparently evolved to take advantage of the fundamental

multisensory nature of the environment, and the multisensory effects have been found across a variety of species such as cats, rats, primates and humans using a large variety of experimental methodologies (see Stein et al., 2004 for a review).

2. Sensory convergence or segregation[2]- which came first?

Complex organisms have evolved to process different kinds of environmental energy and to receive information about the environment through multiple sensory channels. Signals that are transmitted through two or more sensory channels provide both supplementary as well as redundant information about the environment. Considering the highly evolved sensory segregation in complex organisms may lead the observer to intuitively conclude that crosstalk between the senses was an evolutionarily late consequence of sensory segregation. On the contrary, the "unity of the senses" can be traced back to the earliest phases of evolution (Stein and Meredith, 1993).

The different sensory modalities are thought to have evolved from an undifferentiated "supramodal" sensory system, a system that is not selective to what it responds to. In such a system, different effective sensory stimuli, regardless of whether they are chemical, thermal, mechanical or radiant, are thought to have equivalent consequences on the cellular membrane. Despite the

[2] At a neural level, the term multisensory convergence refers to the processing of information from different senses by neurons that respond to input from each of the involved modalities. The term segregation refers to the independence of information processing in different sensory modalities
.

fact that evidence for a single, nonspecific ancestral sensory system is lacking, the differentiation of sensory systems, specialized to be sensitive to just one specific type of environmental energy, is believed to reflect an evolutionary process (Marks, 1978, Stein and Meredith, 1993). Although the phylogenetic development of different receptor systems in extant eukaryotes is inconsistent with the view of a supramodal system, single-celled eukaryotes are supramodal in the sense that they integrate the transduced energy from different receptors into the unified language of a change in membrane potential this representing the most elemental form of multisensory integration. As organisms became more complex and their sensory specialization and segregation increased, integrative mechanisms evolved as well and remained a fundamental property of living organisms.

Primitive invertebrate multicellular species exist in a broad variety of forms and adaptations and generalized statements about the level of segregation of their sensory organs and integration between them are difficult to make. In the most simple multi-cellular species represented by loose networks of cells, such as sponges, simple stimulus identification can take place. These organisms can withdraw from harmful stimuli but despite the involvement of different receptors, many do not segregate the inputs according to different modalities. However, in other invertebrates we find nerve cells, synapses and nerve nets that are specialized to detect and transmit information from different sensory modalities such as rotation and vibration of the body and photic stimuli. While this degree of specialization suggests segregation of the sensory modalities it remains unclear

whether it was actually achieved in these species despite the existence of a simple nervous systems. In most species that have been studied, nerve impulses from different receptor systems pass readily among neural and non-neural cells via electrotonic coupling so that stimuli form different sensory modalities have access to large parts of the organism (De Ceccatty, 1974; Anderson and Schwab, 1982).

It is likely that sensory segregation emerged with the first encephalized species (Stein and Meredith, 1993). Elongated axons that had already begun developing in coelenterates then provided transmission of information over a long distance without the involvement of intervening cells, a basic prerequisite for the selectivity of inputs and outputs. While it is not known which metazoan first developed a genuine separation of senses, the flatworm (platyhelminth) is believed to be a likely candidate (Stein and Meredith, 1993). It has a symmetrical anatomy with an encephalized central nervous system and peripheral plexus. Neurons in its peripheral plexus habituate differentially to vibration and light (while maintaining bimodal sensitivity!) and interneurons in its central nervous system are found to be tactile- or vibration-sensitive. A mixture of sensory segregation and integration has also been documented in advanced invertebrates such as crustaceans. The setae of crayfish, for example, respond to chemical and mechanical stimuli, and concurrent stimulation in both modalities enhances the firing in an axon, innervating the peripheral receptor organ.

From this stage in evolution on a basic architecture of sensory segregation and integration is observable; The transduction of sensory inputs was

accomplished by elaborate modality-specific organs that then relayed input to a central nervous structure for further analysis via separate pathways. Finally, signals are sent to the effector organs via well-defined output pathways. Generally, the higher the degree of afferent separation, the more integration occurs on a central level and more toward the peripheral central organs (Wiersma and Mill, 1965; Tautz, 1987). The convergence of modality-specific afferents with central and efferent neurons is therefore a general characteristic of complex invertebrates.

The segregation of afferent input is most expressed in vertebrates, and, thus far, the sensory systems and their integration have been researched most extensively in mammals. It is generally accepted that sensory segregation in mammalian species is preserved in the sensory projection pathways. However, the vestibular system is a notable exception, since visual and somatosensory input converge early along the sensory pathways. Convergence, according to the most conservative view, occurs at higher sensory association sites in the cortex. There are, however, many multisensory convergence sites below the cortical level, most notably the superior colliculus (SC). Because of its prominent role as a convergence site it will be described in more detail below.

We find multisensory neurons in the reticular formation where they play a role in the regulation of arousal. As we discuss the role of the SC in the regulation of detection and gaze orientation, this will make intuitive sense given the strong relationship between arousal and orientation responses. Subcortical

sites of multisensory convergence can also be found in the locus coeruleus and the external nucleus of the inferior colliculus.

The coexistence of sensory specific pathways and multisensory structures is particularly apparent in the thalamus. Here, the nuclei of the primary projection pathways (e.g. the lateral geniculate nucleus (visual), the medial geniculate nucleus (auditory), and the ventrobasal complex (somatosensory)) that preserve sensory specific information live in close proximity with structures in the posterior and lateral thalamus where multisensory convergence is common. This allows the cortex to receive input that is already multisensory in nature.

The parallel existence of brain regions of multisensory convergence, and structures with a relative (but not absolute as we will see further below) preservation of sensory specificity at brain stem level also exists in the cortex. In primates, several cortical polysensory areas have been identified in the superior temporal region (e.g. Beneveneto et al., 1977), the ventral intraparietal area or VIP (Duhamel et al., 1998), the lateral intraparietal area or LIP (Grunewald et al., 1998), and in the prefrontal cortex (e.g. Graziano et al., 1999). Convergence in polysensory association cortices is believed to subserve a large variety of “higher” functions such as perception of objects (Newell, 2004), the association of auditory and visual speech (Bernstein et al., 2004 for a review), attention (Eimer, 2004), memory processes (Lehmann and Murray, 2005; Murray et al., 2005), and emotion (O’Doherty, 2004).

3. Multisensory integration in mammals

3.1 The superior colliculus as a model of multisensory integration: A brief overview

In the brain, regions with neurons with converging input from different senses can be found on various levels of the neuraxis. With the ground-laying work in cats by the group around Stein and Meredith (e.g. 1988, 1993; Meredith and Stein, 1983, 1986) a midbrain structure called the superior colliculus (SC) has become one of the most extensively studied sites of multisensory convergence. This twin- protuberance together with the two inferior hills (inferior colliculi) forms the roof (tectum) of the midbrain. The multisensory properties of this structure are now well known thanks to the long and intensive investigative efforts of the Stein/Meredith group and others, and the SC has since been serving as a general model of multisensory integration in anatomy, physiology and behavior.

The SC controls orientation towards targets that involve shifts of eye-gaze and positioning of head and limbs. Its seven alternating cellular and fibrous layers are operationally subdivided into a superficial layer (layers I-III) and a deep layer (layers IV-VII). Both superficial and deep layers contain visuotopically arranged neurons that respond to visual input, but it is the deep layers that also receive ascending afferents from other sensory modalities (somatosensory, auditory) and from motor-related structures. As well as from other descending afferents from the cortex. The deep laminae send efferents to the brain stem and to the spinal chord thereby controlling orientation of peripheral structures such as

sensory organs (eyes and ears). This SC coordinates input from the sensory systems and output to the motor systems. This is accomplished by convergence of sensory input with descending input from the cortex. There are a variety of convergence patterns some with matching sensory representations where retinal input converges onto neurons that also receive input from visual association cortices and others with input from different modalities that converge onto the same neurons (e.g. visual inputs from the retina and auditory inputs from the inferior colliculi). The structures involved and the patterns of convergence are manifold, thereby the reader is referred to Stein and Meredith (1993) for a comprehensive description.

For the purpose of a brief overview, it should suffice to say that in mammals, input from different sensory modalities converge in a map-like fashion, which suggests that multisensory neurons have their receptive fields in the same relative locations in visual, auditory and body-space. A multisensory neuron in a rostral location of the SC receives input from neurons receive input from central visual and auditory space as well as somatosensory input from anterior locations on the body (e.g. the face). Neurons in more caudal locations in the SC converge input from more peripheral visual and auditory space and more caudal body parts. These sensory representations are in topographic alignment with motor representations involved in moving eyes, ears, head and limbs toward targets, allowing for a flexible orientation with a common eye-centered frame of reference. While some aspects of these delicate alignments of sensory and motor maps are thought to reflect an innate, hard-wired circuitry, others are

known to be the result of postnatal experience (King et al., 2004; Gutfreund and Knudsen, 2004; Wallace and Stein, 1997; Wallace, 2004a, 2004b, 2006).

3.2 Properties of multisensory integration

The innate properties of the SC as an integrative structure within the CNS and its ability to be modified by experience (presumably via corticotectal efferents) allows the brain to integrate information from different sensory channels based on whether they belong to the same external event. The association between input from different sensory channels is not intrinsic to the stimulus and due to the segregation of the sensory organs, the brain has to reassemble this information. The most salient cues for two or more inputs in different sensory modalities to belong together (because they were generated by the same external event) is their spatial and temporal coincidence. As we will see, neurons in the SC respond to stimuli that coincide in space and time in a special manner. In extracellular electrophysiological recordings, stimulus combinations produce enhanced responses as measured in response reliability, number of evoked impulses, peak impulse frequency and duration of the discharge train (Meredith and Stein, 1986).

Multisensory enhancements found in monkeys, cats and rodents (see Stein et al., 2004 for a review) can vary greatly and can sometimes exceed the sum of the unisensory responses (superadditive). Observations such as these led Stein and Meredith (1983) to operationally define a multisensory effect as a

response that exceeds the response elicited by the most effective unisensory stimulus. This mechanism seems to be an efficacious way to index that two stimuli that were processed in different sensory modalities originated from the same external event. The majority of the multisensory neurons in the SC (65%) respond to two concurrent unisensory inputs by summing the modality-specific inputs. In contrast, responses of a multisensory neuron can be unaffected (no enhancement) or even show depressed responses when the two stimuli fall into different receptive fields, especially when the second stimulus falls into an inhibitory region that borders the receptive field of the multisensory neuron. Twenty two percent of the multisensory neurons in the SC show this property and fire in a frequency that falls below the sum of the respective unisensory inputs (subadditive).

The result of the described enhancement is the increased likelihood of the detection of an event and, as a consequence, the generation of a faster and more accurate response to the event. This property becomes especially important when the respective unisensory inputs are weak. Indeed, about 15% of the multisensory neurons in the SC show superadditive responses. The majority of these superadditive responses occurred when the constituent unisensory stimuli were minimally effective in evoking responses. On the other hand, SC neurons showed the most subadditive responses when the unisensory stimuli were effective on their own. This pattern, termed “the principle of inverse effectiveness,” is considered one of the fundamental principles of multisensory integration. It has also been found in multisensory neurons in the neocortex of

animals (Kayser et al. 2005; Lakatos et al. 2007) and has been inferred in brain imaging (Callan 2003; Calvert and Lewis 2004), although imaging data of multisensory neurons have to be interpreted with great caution (Laurienti, Perrault et al. 2005). The parallels seen between multisensory integration at subcortical and cortical levels have been suggested to reflect a general operational mode of multisensory integration throughout the brain (Stein et al. 2004). This principle will be revisited in the second and third and fifth chapters of this thesis.

It is common practice in dissertations and other academic writings to report excerpts from "The Principles of Psychology" by William James (1890), as his work heralded forthcoming discoveries in the fields of psychology and neuroscience. In following with this tradition I present the following statement as it indicates that James seemed to have a suspected the existence of a neural operation similar to inverse effectiveness:

"The law is this, that a stimulus which would be inadequate by itself to excite a nerve-centre to effective discharge may, by acting with one or more other stimuli (equally ineffectual by themselves alone) bring the discharge about." Page 82

He believed that this relationship that is found in nerves may be a general principle that can also be found in behavior as well as other psychological phenomena. And although he did not explicitly make a reference to stimulation in different senses and focused more on effects of the temporal summation of weak

stimuli in one modality, inverse effectiveness in multisensory stimulation is implicit in his examples that follow:

"Bruecke noted that his brainless hen, which made no attempt to peck at the grain under her very eyes, began picking if the grain were thrown on the ground with force, so as to produce a rattling sound." Page 84

"A patient who cannot name an object simply shown him, will name it if he touches as well as sees it." Page 85

And for an example of a multisensory effect (and maybe foreseeing the discovery of the temporal rule):

"If a car- horse balks, the final way of starting him is by applying a number of customary incitements at once. If the driver uses reins and voice, if one bystander pulls at his head, another lashes his hind quarters, and the conductor rings the bell, and the dismounted passengers ahove the car, all at the same moment, his obstinacy generally yields, and he goes on his way rejoicing." Page 84

As these examples suggest, inverse effectiveness has significant implications for behavior, especially for orienting and detection responses. This should be apparent in the anatomical connectivity of the SC. Indeed, the SC is connected to the effector organs via the predorsal bundle, a fiber bundle that

leaves the SC and reaches the periphery via the brainstem and spinal chord. As the anatomical connectivity and the patterns of multisensory integration in the SC predict, behavioral correlates have been identified in animals and humans (Stein et al., 1989; Diederich and Colonius, 2004). In cats, spatially and temporally coincident stimuli, such as a flash and a tone on a perimetry device, evoked significantly more correct orientation responses than the unisensory stimuli alone. On the other hand, a spatially disparate auditory stimulus depressed orientation to a visual stimulus. Correct responses were superadditive more often when the unisensory stimuli were not likely to evoke correct responses, and subadditive responses were found when cats were likely to respond correctly to a unisensory stimulus. This can be regarded as a behavioral equivalent of the inverse effectiveness rule.

Recent evidence from lesion studies in cats (Jiang et al, 2001) suggests that multisensory neurons in the SC work in concert with cortical multisensory neurons. In fact, cortical sites with a high incidence of multisensory neurons, such as the anterior ectosylvian sulcus and the lateral suprasylvian sulcus in cats, seem to have a modulatory influence on the integrative attributes of the SC. It has been shown that SC neurons lose their multisensory properties when these cortical regions are deactivated (Jiang et al. 2001), while the responses to unisensory stimuli remain intact. Moreover, this loss of multisensory properties has been shown to translate to behavior (Jiang et al., 2002). Deactivation of either cortical region had little effect on unisensory orientation behavior, but multisensory response enhancement was lost and response depression to

spatially disparate stimuli was degraded. This cortical influence of multisensory responsiveness of the SC provides an excellent model for the influence of experience on multisensory integration in the SC.

Taken together, the SC is an established site of MI integration where sensory specific information converges in a rule- based manner thereby organizing detection and orientation responses. The SC is widely connected throughout the brain and it seems likely that multisensory responsiveness in the SC is tuned by influenced from higher order multisensory integration regions in the cortex

3.3 Multisensory integration in the cortex

Research on the SC has provided us with a model that shows how convergence patters in neurons can explain detection and orientation in animals. It remains to be shown how multisensory stimulation affects behavior and perception in humans. The effect of multisensory stimulation on human perception and behavior is well established and interactions between modalities have been shown for virtually every combination of sensory modalities. The research on perceptual and behavioral consequences of multisensory stimulation is vast and the reader is referred to Calvert, Spence and Stein (2004) for comprehensive reviews. This introduction is limited to the description of topics more closely related to the forthcoming chapters and to the experiments of this dissertation, which include the physiological basis of multisensory integration and the perception of speech and dynamic events. It should be noted, however, that

multisensory effects are ubiquitous and are found in almost every domain of brain function involving visual object perception, spatial perception of vision and sound, attention, memory, and speeded classification, just to name a few. All of these functions, regardless whether they involve sensory, perceptual or executive processes involve the cortex. It remains to be discussed how this abundance of integration is accomplished in the face of the apparent segregation of the sensory systems.

Investigations of the functional neuroanatomy of the cortex have revealed areas that are specialized to process direct input from the receptive organs of specific sensory modalities. This modularity, apparent in the primary sensory projection pathways, is likely to be the reason for the high precision and efficiency (as can be measured behaviorally) of our sensory systems as well as for the unique sensory experience that emerges from their activities. Traditional scientific opinion offers the intuitive notion that one of the main common attributes of our highly specialized sensory systems is their operational independence. A large body of research in monkeys shows that each sensory system is associated with an interconnected network of cortical areas that are organized in a hierarchical fashion from simple to more complex processing (Felleman and Van Essen, 1991). Several cortical areas, in higher order stages in the information processing hierarchy, in the superior temporal sulcus/gyrus (STS/G) (Beneveneto et al., 1977), the parietal cortex (Seltzer et al., 1980) and the frontal lobe (Bodner et al., 1996) have been identified to receive input from more than one sensory modality. This as well as current knowledge about any

anatomical cross-connections between modalities at earlier stages, has led to the belief that the convergence of sensory input does not occur until later in the information processing hierarchy and in higher-order sensory association cortices, when the processing of information in each constituent sense is terminated. As we will see, this view, or at least its exclusiveness, has been challenged by more recent findings that give rise to a new understanding of the cortical processing of sensory inputs (see Schroeder and Foxe, 2004; Foxe and Schroeder, 2005, for reviews).

Subsequent imaging work in humans has revealed multisensory integration in lower-sensory regions (Calvert et al., 1999; Lewis et al., 2000; King et al., 2001; Macaluso et al., 2000). The preferred explanation for these findings, which is consistent with hierarchical models, is that multisensory effects in "sensory specific" cortex are a result of feedback input from higher order multisensory regions. Indeed, evidence for such feedback processes exists from human event related potential (ERP) studies (Murray et al., 2002) and intracranial recordings in monkeys (Schroeder and Foxe, 2002) that support the important role of feedback in multisensory integration. This, however, at least theoretically, does not preclude the existence of crosstalk between sensory modalities before higher-order multisensory regions are activated. Moreover, since the haemodynamic signal is an indirect measure of cell activity and reflects activation only many seconds after the underlying neural event occurs, studies using haemodynamic measures themselves are not suited to provide support for models that require precise timing information to be validated.

Such temporal resolution was provided by studies using electrophysiological techniques. A study by Giard and Peronnet (1999) showed audiovisual (AV) interactions in the ERP as early as 40 ms after stimulus onset in the occipital cortex. At this time the first visual inputs arrive in the occipital cortex (Foxe and Simpson, 2002). The timeframe and locus of this effect strongly suggest that the multisensory interaction was not caused by a feedback from higher order multisensory cortical areas but was more likely due to a crosstalk between early auditory and low-level visual areas. A similar experiment was repeated in our laboratory using a simpler stimulus design, more sophisticated EEG-recording techniques, and the collection of chronometric behavioral measures (Molholm et al., 2002). Reaction time data were collected to establish behavioral confirmation of a genuine multisensory interaction. That is, multisensory stimulation generally results in an acceleration of reaction time (RT) in comparison to reactions to unisensory stimuli which is called the “redundant target effect” (Todd, 1929). This effect could solely be due to a “race” between the sensory specific processing hierarchy until a decision criterion is reached and a subsequent motor response is initiated. In the case of multisensory stimulation two stimuli enter the race and have a higher likelihood (as indexed by probability summation) of reaching a response criterion earlier. If empirical data show multisensory RT’s that are faster than probability summation predicts, then the race model is regarded as violated and it is likely that a genuine multisensory interaction has taken place (Murray et al., 2001).

The investigators found an AV interaction in the ERP around 46 ms, at the onset of the C1 component. This component of the visual evoked potential has been implicated as the first input to visual cortices, a timeframe that renders a feedback explanation unlikely. Topographic mapping revealed that this interaction was maximal not in V1 but over the right parieto-occipital scalp. In agreement with the earlier study by Giard and Perronet (1999), the authors concluded that an early stage of the dorsal visual pathway was a likely site for this interaction. The behavioral measures also suggested a genuine multisensory effect showing a violation of the race model.

It has been mentioned that the absence of knowledge of anatomical cross-connections between primary sensory areas was one of the primary reasons for the dominance of the traditional, hierarchical models of sensory processing and multisensory integration. The empirical evidence on this matter has changed over the past years and recent findings from anatomical tracing studies should be reported here in brief. Rockland and Ojima (2003) were able to show direct projections from auditory parabelt regions to V1 and V2. More evidence comes from a study by Falchier et al. (2002) who reported direct input from the primary auditory cortex (A1) and the caudal parabelt regions of the auditory association cortex into the peripheral V1. The termination profiles of these inputs are multilaminar, which are typical for feedback and lateral connections. Further, it was estimated that peripheral V1 received substantial input from visual motion-processing area MT+ and a direct input from the superior temporal polysensory area (STP). Given these recent advances, it seems likely that early multisensory

convergence in the visual cortex is accomplished by feedback and lateral input from the auditory cortex.

The question remains as to whether these lateral inputs are specific for the visual cortex or if similar multisensory convergence exists in the primary auditory cortex. Evidence for such interactions has been presented recently when Bizley et al. (2007) demonstrated that visual activity can influence the activity of units in the auditory cortex of anesthetized ferrets using single unit recordings. In the primary auditory cortex and the anterior auditory field, 15% of the units were reported to have non-auditory input, and this proportion increased in the higher-order auditory fields ventral to A1/AAF. Responses of multisensory neurons to pure tones did not differ from the responses of pure auditory neurons in the same fields. Further, neural tracer injections revealed direct input from V1 to A1 and, to an increasing extent to higher auditory areas. The authors concluded that it is likely that these anatomical cross-connections underlie the multisensory effects found in single cells in the primary auditory cortex.

There is also extensive research on the connectivity between the somatosensory and the auditory modalities and there is now evidence that somatosensory afferents terminate in an area called the caudo-medial auditory cortex, which is only one synapse away from A1 (Schroeder et al., 2001; Schroeder and Foxe, 2005, for a review). This somatosensory input terminated in layer four of the lamina and triggered responses in the extragranular laminae. This laminar activation profile is consistent with feedforward convergence of somatosensory and auditory input.

Taken together, it seems very likely that multisensory convergence in the mammalian cortex is not only accomplished in higher order multisensory association cortices, and then feeds back to subsequently influence sensory-specific processes, but also via lateral connections between sensory-specific cortices and feedforward input. The important question remains of what advantage such early convergence patterns represent. The functionality becomes immediately clear when one considers that (to remain with audiovisual interactions) sound waves and light travel at different speeds and that the modality-specific environmental energy arrives at our sensory organs at different times. In order to perceive these stimuli as belonging to the same event they need to be associated with one another. What was classically termed the "binding problem" becomes more complicated when considering the different signal transmission rates of the different sensory systems. Imagine a complex yet common-place environment like a street scene, where a multitude of stimuli from different modalities have to be "bound" in a matter of seconds. This would be a rather complicated endeavor if our brains were designed to associate stimuli "late" in multisensory association cortices. On the other hand, the idea that the brain associates incoming stimuli early, and possibly repeatedly before later associative stages are reached, seems advantageous. This may, in this way, create a coherent percept of the environment with beneficial consequences for behavior.

4. Multisensory integration of audiovisual speech in the human observer

Under noisy acoustic conditions, seeing the articulations of the speaker has a substantial benefit for the perception of the information. Under experimental conditions human observers can experience a multisensory enhancement of up to 11dB increase in signal-to-noise ratio (SNR) (Sumby and Pollack, 1954; McLoad and Summerfield, 1987). The visual signal is believed to not only restore missing auditory information but also interact with the residual auditory signal, resulting in superadditive multisensory enhancements (Breeuwer and Plomp, 1985). As we will see in Chapters two and three of this thesis, substantial multisensory enhancements are found even if the auditory signal by itself is entirely unintelligible and speechreading (or lipreading) by itself can only account for about half the gain. Even under perfectly clear listening conditions visual information strongly influences the perception of speech, as the McGurk effect most dramatically demonstrates. It seems clear that the influence of visual information from a talker's mouth and face plays a strong role in the perception and understanding of spoken language.

As in other fields of multisensory integration the observed integration of sensory input forces the question of where, when and how such integration is implemented in the brain of the observer. Research in this discipline has traditionally been dominated by larger-scale theoretical accounts of how speech-processing and language operates. These have derived their evidence mostly from behavioral work, but often with little reference to possible neural implementations (for overviews and taxonomy of models of audio-visual

integration see Summerfield, 1987 and Schwartz et al., 1998). Neurophysiological investigations have begun only relatively recently and first attempts have been made to integrate theoretical models of audiovisual integration and what is known about the neural correlates of AV-integration (for review see Bernstein et al., 2004). This section on audiovisual speech is intended to outline models of AV speech perception with specific reference to its neural representation and empirical evidence that informs such models. This section is limited to this selection because the focus of the experiments described in Chapters two and three, although behavioral, have strong implications for neuronal aspects of multisensory integration.

Speech perception theories generally assume a hierarchical process of stimulus analysis in which subsegmental speech features such as voicing, manner, place, and nasality, are extracted and then subsequently combined into consonant and vowel segments which are finally combined into word patterns. Several theories have been proposed that conceptualize how and when visual and auditory signals are combined or associated. A very general way of classifying models of AV-speech perception is to distinguish common format and modality-specific theories (see Bernstein et al., 2004 for a review, and for a detailed taxonomy of AV-integration theories see Schwarz et al., 1998). Common format theories of audiovisual speech perception suggest that at some point along the way the auditory and visual modality-specific stimulus information is transformed into an amodal representation. The neural implementation has been conceptualized as a convergence of modality specific information in

multisensory neurons. Extensive research with the McGurk effect has led many scientists to believe that visual speech gestures influence the perception of speech at early, subsegmental levels (prior to the identification of consonants and vowels) (see Green, 1998, for a review). Here, visual and auditory speech converge early, and the speech segments no longer comprise modality-specific perceptual information. One popular variant of common format theories is the idea that the perception of speech is achieved by same networks involved in speech production (Liebermann and Mattingly, 1985). Although intuitively appealing, neurophysiological and neuropsychological evidence (from studies in patients with aphasia) for this notion is yet missing (Bernstein, et al., 2004). On the other hand, according to modality-specific theories, auditory and visual speech information is processed by modality-specific networks and then associated. These models do not require neuronal convergence.

Comprehensive overviews of proposed models and their evidence exist elsewhere (Bernstein, 2004, Massaro, 2004, Fowler, 2004; Schwartz et al., 1998; Summerfield, 1987) and a detailed description cannot be attempted here. It should be noted however, that despite the fact that current opinion seems to favor common format approaches (but for a criticism of Massaro's FLMP see Gigerenzer, 1989, Schwartz, 2006), the way AV-speech signals are integrated remains unresolved. There is evidence for common format as well as modality-specific models but, after all, these theories have to be informed by experiments investigating how AV speech perception is implemented by the brain. The knowledge that presently exists about the auditory and visual speech pathways

and the ways of their interaction is not sufficient to constrain the present, behaviorally informed models of AV-integration. The following is a summary of what is presently known about the auditory and visual speech processing pathways and their possible sites of interaction.

4.1 The auditory speech processing pathway

As revealed by tract-tracing studies in macaque monkeys, the auditory pathway ascends through the brain stem to the central nucleus of the inferior colliculus. From there, auditory information is relayed through the medial geniculate nucleus of the thalamus, the main relay station for auditory information, on its way to the core auditory cortex (A1). A collateral branch leaves the external branch of the inferior colliculus and is relayed through the superficial layers of the superior colliculus to terminate in the deep layers of the superior colliculus. Here, converging input ascends through non-primary thalamic nuclei (including the medial, lateral and inferior pulvinar and the anxillary nuclei of the medial geniculate) to the cortex in two types of projections. The first projection, relayed through the lateral and inferior pulvinar, projects to higher levels of the visual cortex (Brodmann areas 18 and 19), while the medial pulvinar and the anxillary medial geniculate nuclei project to higher-level auditory cortex (Brodmann areas 42 and 22; Brodmann, 1908). Interestingly, tangential projections from the named pathways reach the core auditory and the core visual cortices (Rezak and Beneveneto, 1979)

The core or primary auditory cortex receives the vast majority of the ascending auditory input (e.g. Rauschecker et al., 1997) relayed through the (ventral, parvocellular) medial geniculate nucleus and is homologue in monkeys and humans. The core area codes elemental stimulus features such as pitch, sharpness of frequency tuning, intensity tuning and binaural interaction (Recanzone et al., 1999) and has a tonotopic organization as shown in monkeys (Morel and Kaas, 1993) and humans using various imaging techniques (MEG: Elberling et al., 1982; PET: Lauter et al., 1985; fMRI: Wessinger et al., 1997). Importantly, the core was found to not respond differentially to speech versus non-speech stimuli (Binder et a., 2000; Celsis et al., 1999; Huckins et al., 1998; Scott et al., 2000).

Termed by Brodmann as “area 41” (Brodmann, 1908) and residing in the transverse temporal gyrus (Heschl’s gyrus), the core area projects heavily into “area 42”, the “belt area,” surrounding the core and representing a second synaptic level of the auditory processing pathway. The belt region is involved in the perception of sounds and has been shown to respond to species-specific calls (Rauschecker et al., 1995; Wang et al., 1995). As the auditory core, the belt area does not activate selectively to speech in human listeners (Binder et al., 2000; Steinschneider et al., 1999; Wise et al., 2001).

The extensive parabelt region (Brodmann “area 22”) that surrounds the core and the belt represents the third synaptic level and receives input from the belt region (but not the core!) (e.g. Luethke et al., 1989, Hackett et al., 1998). Like the belt region, the parabelt does not receive input from the medial

geniculate nucleus but from the anxillary geniculate nuclei and the medial pulvinar only. The parabelt region seems to process more complex stimuli such as speech.

Finally, a fourth synaptic level has been suggested to be represented by projections that pass from the parabelt to the superior temporal sulcus (e.g. Kaas and Hackett, 2000; Pandya et al., 1969). This region, comprising the superior temporal polysensory area (STP) is characterized by extensive multisensory integration for auditory visual and somatosensory information and is the fourth level of projections from all named senses. This area activates to a wide variety of auditory inputs, especially those related to speech.

Using fMRI, Binder et al. (2000) showed that tone stimuli activated early areas of the auditory pathway, whereas passive listening to speech activated also the parabelt and the parts of the STG. Word classification produced widespread activation of the parabelt STS/G, the mediotemporal gyrus and the temporo-parietal cortex (Wernicke's area). Similarly, intracranial recordings in humans have shown a passage of activation from Heschl's gyrus laterally to the superior temporal gyrus (Howard et al., 2000; Liegois-Chauvel et al., 1994).

Taken together, anatomical connectivity with results from human functional imaging studies suggest a hierarchical processing of auditory stimuli. However, in light of the discussion about early influences of auditory processing in A1 earlier in this paper, we have to question the absoluteness of this notion. The nature of this penetration of the early auditory processing hierarchy, on the other hand, demands future investigation.

4.2 The visual speech processing pathway

Visual information enters the cortex via V1, the exclusive recipient of the mangocellular and parvocellular projections from the lateral geniculate nucleus (Felleman and Van Essen, 1991). V1 is retinotopically mapped and neurons respond to low-level stimulus attributes like luminance, wavelength, spatial frequency, orientation, motion and binocular disparity. The second synaptic level is represented by visual association areas such as V2, V4 and the motion processing area V5 (MT) all being specialized in specific aspects of visual analysis. The phonemic analysis of visual articulation requires a more elaborate analysis than can be accomplished in cortical areas within the first two synaptic levels. The third and fourth synaptic levels include the fusiform face area that specifically activates to faces in imaging studies in humans (e.g. Haxby, 1991, 1996) (but not the identity of faces; see Tong et al., 2000) and areas that are instrumental in the assembly of visual objects from motion, texture or luminance contrast in the lateral occipital cortex (Grill Spector et al., 1998). Representations of object categories are believed to be contained in the inferotemporal cortex (e.g. Puce et al., 1996).

Visual articulation requires a more complex level of analysis. The phonemic content of the dynamic articulatory gestures has to be identified by the particular motion characteristics of the face and the entire head. This type of analysis is believed to be accomplished in areas specialized in so called

biological motion (BM) processing[3]. Functional imaging studies in humans have shown that the cortex in the posterior temporal sulcus/gyrus (STS/G) responds to biological motion (for a review see Allison et al., 2000; Puce and Perrett, 2003). In non-human primates the STS/G responds to face motion (e.g. Calvert and Campbell, 2003; Campbell et al., 2001; Puce et al., 2003) and has repeatedly been implicated in the process of speechreading in conjunction with the angular gyrus, posterior cingulate and the medial frontal and frontal poles (e.g. Puce et al., 1998; Calvert et al., 1999; Campbell et al., 2001). The STS/G activates to both visual and auditory speech (Calvert et al., 1997; MacSweeney et al., 2000, 2001) which gave rise to the idea that the influence of visual speech on auditory speech occurs via feedback of higher-order multisensory areas such as in the STS/G to the auditory cortex.

The STS/G has widespread connections to motion processing areas MT+ (Puce et al., 1998), the lateral occipital complex, the auditory cortex and frontal areas involved in motion processing. This connectivity pattern makes the STS/G, at least intuitively, a likely candidate for the processing of articulatory speech. Interestingly, areas in the STS/G have also been shown to be part of a network responsible for the judgment of gaze perception (Hoffman and Haxby, 2000). Gaze monitoring is a central aspect in the attentional regulation of social interaction, especially communication. Further, biological motion sites in the STS/G are connected with areas in the frontal lobes that regulate the

[3] BM is the perception of a moving object by the relative motion of its parts to each other. In comparison to the perception of form from motion where the object itself remains static and moves in relation to other objects or the environment, it involves the change of the form of the object itself.

programming and execution of motor responses. Theoretically speaking, this opens the possibility for a motor component in perceptual speech, as Liebermann proposed in his motor theory of speech perception where perception is coded in the format of intended articulatory gestures. (Liebermann and Mattingly, 1985)

4.3 Interaction of visual and auditory speech

Given what we know about the auditory and the visual pathways, some of which was described in the brief synopsis above, there is a range of possibilities of how, when and where auditory and visual information can interact and converge onto common cells. Auditory and visual information converge in the SC and can therefore influence cortical processing via ascending pathways to auditory and visual association cortices and (conversely) in a modulatory function in V1 and A1 (Bernstein, 2004; Cauller and Connors, 1994; Rezak and Benevento, 1979). Further, earliest cortical interactions are possible via lateral connections from visual to auditory cortices and, arguably, even as early as A1 (e.g. Schroeder and Foxe, 2005). AV interactions are also likely via feedback connections from higher order multisensory association cortices to the earlier auditory cortex (Calvert, 1997; 2004). Finally, it is possible that no crosstalk occurs between the visual and the auditory speech pathways until later and that information is processed in higher order speech-specific multisensory convergence networks (Mesulam, 1998; Bernstein, 2004).

This multitude of ways for AV interactions to be implemented raises the question of whether we can impose any theoretical or empirical constraints. In what follows I will attempt to draw some, in varying degrees speculative conclusions from what is known about the auditory and the visual speech pathways. I will include some evidence from functional imaging and electrophysiological studies and will hopefully arrive at a more informed model of AV-integration of speech that will provide a framework for the first two empirical Chapters of this thesis.

First, it is reasonable to assume that visual and auditory speech information is not processed differently than non-speech stimuli at subcortical and early cortical levels (A1 and belt and V1). Most likely, auditory speech and visual articulation are analyzed like non-speech visual and auditory information at early levels of analysis. Second, visual articulation can contain information with different levels of complexity. For example, the timing and probably the extent of mouth opening represents visual articulatory information that is available to the system at early levels of analysis and may be delivered to auditory processing sites via lateral connections from the visual to auditory cortices in a feedforward manner, and quite possibly at first synaptic levels of analysis (Foxe et al., 2000, 2002; Foxe and Schroeder, 2005; Giard and Peronnet, 1999; Molholm et al., 2002; Murray et al., 2005; Schroeder and Foxe, 2002, 2005). On the other hand, phonetic information in the visual stimulus, such as place of the closure of the vocal tract (manner) or the level of the raising of the velum (nasality) is likely to require a more elaborate analysis of the visual system, possibly involving BM

processing in the STS/G. At first glance, it may look as if this higher-order visual information contributes to the speech percept at late stages possibly via convergence with auditory information in multisensory integration sites in the STS. However, depending on the particular AV-speech stimulus, the visual information often precedes the auditory input. This enables articulatory information to arrive in cortical sites of higher order analysis early enough to modulate auditory processing before it reaches the STS/G. Hence, the phonetic content in visual speech could modulate auditory processing at stages prior to its arrival in multisensory cortices like the STS/G via feedback.

Electrophysiological investigations using the McGurk-effect have provided useful insight into the timing and location of AV-effects in speech (Sams et al., 1991; Möötönen et al., 2002; Colin et al., 2002, 2004; Saint-Amour, 2007). A well known electrophysiological component called the mismatch negativity (MMN) can be evoked by rarely occurring stimulus deviants in a train of frequently presented stimuli (standards) (e.g. Naatanen, 2001; Ritter, 1995). The MMN is believed to index auditory sensory memory.

Most interestingly, the MMN can also be evoked with the McGurk-effect. Here, instead of a change in physical stimulation, an illusory change induced by the incongruent visual articulation evokes the MMN. Several important findings should be highlighted here. First, the visual stimuli did not evoke a MMN alone and the illusory MMN appears to be synchronized to the auditory stimulus (Colin et al., 2002; Saint-Amour et al., 2007), suggesting that the visual stimuli need a suitable auditory context in order to trigger an MMN. Second, like the MMN

evoked by a change in the auditory modality alone, the McGurk-MMN appeared to be generated in the primary auditory cortex with a left-hemispheric dominance (Saint-Amour et al., 2007) suggesting that auditory processes are modulated by visual input to the primary auditory cortex. Third, in the study by Saint-Amour et al. the McGurk-MMN onsets around 175 ms after the onset of the auditory stimulus (Saint-Amour et al., 2007), and earlier onsets around 150 ms have been reported in other studies (Colin et al., 2002; Möötönen et al., 2002). This has been interpreted as evidence for an early integration of auditory and visual speech before phonetic analysis. However, the visual deviant in these studies onsets as early as 200-300 ms before the auditory stimulus, giving the visual information ample time to travel through the visual system into higher order polysensory areas such as STS/G. As mentioned before, the level of analysis of the visual signal necessary to evoke the McGurk-effect requires a relatively elaborate analysis of the information delivered by the moving face. Given these constraints, it is likely that the illusory auditory percept in the McGurk effect is evoked by visual information on phonemic levels, affecting auditory processing at early, possibly pre-phonetic levels of auditory analysis in the primary auditory cortex.

Taken together, in audiovisual speech, acoustic and visual information are probably integrated at different levels of analysis. While pre-phonetic visual information can modulate acoustic processing in a feedforward manner, higher-level articulatory phonetic information is likely to influence auditory processing through backprojections to auditory cortices. However, this information is

probably available to the auditory system at early levels of analysis due to the prior availability of the visual stimulus.

We have seen how physiological information with high temporal acuity can constrain current models of audiovisual information in speech. With the discussion of recent electrophysiological data on the McGurk-effect, I have provided a framework within which we can understand the influence of visual articulation on auditory speech information, which will be the focus of the forthcoming two Chapters of this thesis.

CHAPTER II

Do you see what I'm saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments

Lars A. Ross [a,b], Dave Saint-Amour [b], Victoria Leavitt [b,c],

Daniel C. Javitt [a,b], and John J. Foxe [a,b,c,*]

[a] *Program in Cognitive Neuroscience, Department of Psychology*
The City College of the City University of New York
138th Street & Convent Avenue
New York, NY 10031, USA

[b] *The Cognitive Neurophysiology Laboratory*
Nathan S. Kline Institute for Psychiatric Research,
Program in Cognitive Neuroscience and Schizophrenia
140 Old Orangeburg Road
Orangeburg, NY 10962, USA

[c] *Program in Neuropsychology, Department of Psychology*
Queens College of the City University of New York
65-30 Kissena Boulevard
Flushing, NY 11367, USA

Acknowledgements: Support for this work was provided by grants to JJF from the National Institute of Mental Health (MH65350) and the National Institute on Aging (AG22696). The authors would like to express their sincere thanks to Dr. Sophie Molholm for her ever-valuable comments on earlier versions. We would also like to thank our good friend Dr. Alex Meredith for his challenging comments and two anonymous reviewers for their helpful suggestions.

Published in the journal Cerebral Cortex May 2007; 17:1147-1153

1. Abstract

Viewing a speaker's articulatory movements substantially improves a listener's ability to understand spoken words, especially under noisy environmental conditions. It has been claimed that this gain is most pronounced when auditory input is weakest, an effect that has been related to a well-known principle of multisensory integration - *inverse effectiveness.* In keeping with the predictions of this principle, the present study showed substantial gain in multisensory speech enhancement at even the lowest signal-to-noise ratios (SNRs) used (-24 dB), but it was also evident that there was a 'special zone' at a more intermediate SNR of –12dB where multisensory integration was additionally enhanced beyond the predictions of this principle. As such, we show that *inverse effectiveness* does not strictly apply to the multisensory enhancements seen during audio-visual speech perception. Rather, the gain from viewing visual articulations is maximal at intermediate SNRs, well above the lowest auditory SNR where the recognition of whole words is significantly different from zero. We contend that the multisensory speech system is maximally tuned for SNRs between extremes, extremes where the system relies on either the visual (speech-reading) or the auditory modality alone, forming a window of maximal integration at intermediate SNR levels. At these intermediate levels, the extent of multisensory enhancement of speech-recognition is considerable, amounting to more than a threefold performance improvement relative to an auditory-alone condition.

2. Introduction

Speech, surely one of the most complex inputs that the human brain must decode, is fundamentally perceived as an auditory experience. Yet, research has shown that there are often profound influences from the visual system on this ostensibly auditory perception, and that in some cases, visual inputs can even override the veridical inputs of the auditory system (e.g. McGurk and MacDonald, 1976; Spence and Driver, 2000). Functional imaging studies have bolstered this view by identifying regions of the brain that show integrative processing for the combination of visible and heard speech (e.g. Calvert et al., 2000; Calvert and Campbell, 2003). Our ability to decode speech is even more remarkable when one considers the relative ease with which humans can understand a speaker under what are often highly adverse listening conditions (e.g. the factory floor, the noisy Manhattan sidewalk, the holiday office party). To accomplish this, it is clear that viewing a speaker's articulatory movements, actually watching the actions of the mouth, provides critically important complementary information and serves to augment and enhance our auditory capabilities (see e.g. Sumby and Pollack, 1954; Grant and Seitz, 2000). Indeed, it has been shown that viewing the ancillary head movements that typically accompany speech also provide linguistic information (e.g. Munhall et al., 2004).

In seminal work, Meredith and Stein (1986) delineated a set of principles of multisensory integration, one of which they termed "*inverse effectiveness*" (see also Stein and Meredith, 1993). That is, during recordings in multisensory neurons of the cat superior colliculus (SC), they repeatedly found that maximal

multisensory response enhancements occurred under circumstances where the constituent unisensory stimuli were minimally effective in evoking responses. In other words, multisensory enhancement was greatest when unisensory stimuli were at their weakest, clearly a very useful property in a structure specialized for orientation. That is, potential ambiguity from weak sensory input (e.g. the direction of an object or event) can be compensated for by inputs from a second or third sensory system, providing a significant advantage for the organism. Behavioral work confirmed these properties (Stein, et al., 1988).

At higher intensities, when the unisensory stimuli evoke more robust responses of their own, redundant information is provided by both unisensory inputs and the need for multisensory integration is considerably lessened, and this is paralleled by a drop in the amplitude of multisensory enhancements seen in responses of SC neurons. Analogous to this principle of *inverse effectiveness*, classical multisensory speech studies, undertaken with both hearing-impaired and normal hearing listeners, claimed that the gain from seeing a speaker's articulations was inversely related to the signal-to-noise ratio (e.g. Sumby and Pollack, 1954, O'Neill, 1954, Neely, 1956, Erber, 1969, 1971, 1975, Binnie et al., 1974, McCormick, 1979), and more recent neurophysiological investigations have made similar claims (e.g. Callan et al., 2001, 2003; see also Calvert and Lewis, 2004).

A close reading of the early behavioural studies (Sumby and Pollack, 1954; Erber, 1969, 1975) reveals a potential flaw in this assertion. That is, these early studies used a delimited set of word stimuli that were presented to the

subjects prior to and during the experiments in the form of checklists. It is very likely that this manipulation greatly facilitated word recognition overall, leading to artificially high improvements in speech recognition due to speech-reading. Further, this manipulation would have had particularly large effects at low signal-to-noise ratios (SNR). That is, it is easier to guess a word based on very sparse input when the options are limited to relatively short lists. In the more recent neurophysiological investigations, we also come upon the claim that multisensory speech-processing regions in the human superior temporal gyrus/sulcus (STS/STG) respond according to the principle of *inverse effectiveness*. In an elegant study, Callan et al. (2003) showed that STS activation to audio-visual speech signals was significantly enhanced when the speech signal was embedded in noise in comparison to a condition without noise. They interpreted this finding as evidence for the operation of *inverse effectiveness* in STS but unfortunately, only a single noise-level was used. In fact, behavioral testing in an auditory-alone condition showed that their subjects were able to recognize more than 20% of the words at this noise level, indicating a relatively high baseline level of intelligibility. If the principle of *inverse effectiveness* can in fact be applied to speech processing regions, then the prediction would be that the audiovisual gain should increase with decreasing SNR's, but this was not tested in the Callan study.

We contend that in classical behavioral studies (Sumby and Pollack, 1954, Erber 1969, 1975) the large gain from visual input at low SNRs was enhanced because of the methodology used. Here, we used a modified design employing a

much larger stimulus set, so that each stimulus presentation was unique and no checklists were available to subjects. We predicted that, unlike these previous studies, we would find highest gain due to multisensory audio-visual enhancement at intermediate SNRs[4]. We reasoned this on the following grounds. While the SC is a structure specialized for detection and orientation, a function that clearly benefits greatly from *inverse effectiveness*, the speech recognition system is concerned with higher-level semantic recognition and not just the simple presence or absence of a speech stimulus. Thus, it is reasonable to expect considerable differences between the modes of processing within the speech recognition system and that found in the SC.

We set out to delineate the conditions under which visual articulatory cues have their greatest impact on speech-recognition. We reasoned that there would be a maximal window of multisensory integration, a range within which reliance on either pure auditory or pure visual inputs would produce sub-maximal recognition. That is, when auditory noise in the environment is of sufficient magnitude to mask speech signals, it is obvious that the speech recognition system would be biased to rely almost entirely on visual inputs (i.e. speech-reading) (see also Erber, 1969), and when the auditory signal is intelligible and unambiguous, the system would be biased towards substantial or even complete reliance on the auditory inputs themselves. Given these two extrema, it stands to

[4] Note that a somewhat similar prediction has been made by Bernstein and colleagues (2004) based on a review of classical studies (Sumby and Pollack, 1954 and MacLeod and Summerfield, 1987). They point out that although these studies claim to show enhancements at very low SNRs (-15 to –20 dB), enhancements at these levels are mainly due to high speech-reading scores, driven by the use of delimited wordlists.

reason that some interim value exists where speech-recognition will be weighted toward an equal reliance on both auditory and visual inputs and that this point on the continuum will be the point of highest multisensory integration. We sought to determine this maximal level and to ascertain whether multisensory AV speech mechanisms have a maximal tuning window.

3. Methods

3.1 Subjects

Twenty adults (9 females) between the ages of 18 and 59 (mean= 31, SD= 11.5) participated in this study as healthy volunteers. All participants were native English speakers with normal or corrected-to-normal vision, and had normal hearing and no history of neurological or psychiatric disorders according to self-report. The Institutional Review Boards of the Nathan Kline Institute for Psychiatric Research and of the City College of the City University of New York approved the experimental procedures, and each subject provided written informed consent.

3.2 Stimuli

Stimulus materials consisted of 525 simple monosyllabic words (taken from the online MRC Psycholinguistic database: http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm). The words were selected from a well-characterized normed set based on their written-word

frequency (Kucera and Francis, 1967)[5]. The face of a female speaker was digitally recorded articulating the words. These movies were digitally re-mastered, so that the length of the movie (1.3 sec) and the onset of the acoustic signal were highly similar across all words. Average voice onset occurred at 520ms after movie onset (SD= 30ms). The words were presented at an intensity of 50dB SPL. Sound level for each word was adjusted to 50 dB SPL using a measure of average intensity across the duration of word, measured using a Brüel and Kjær Type 2236 Sound Level Meter with the time constant set at "F". Seven different levels of pink noise were presented simultaneously with the presentation of the words at 50, 54, 58, 62, 66, 70 and 74dB SPL. Noise onset at the same time as the movie onset, 520ms before the voicing began. The signal-to-noise ratios (SNRs) were therefore 0, -4, -8, -12, -16, -20 and –24dB SPL. The movies were presented on a 21-inch computer monitor at a distance of 1.7m from the participant with a rate of 30 frames per second. The whole face of the speaker was visible and extended 6.3° horizontally and 7.6° vertically. The words were presented from a speaker situated in the center on top of the screen and the noise was presented from speakers flanking both sides of the screen.

[5] It should be pointed out that what constitutes everyday language usage has clearly changed to some extent since this normed word set was first established. Here, care was taken to only select words that all authors considered to be still in common use. In addition, the chosen words were distributed over conditions randomly and were checked for equal distributions afterwards.

3.3 Procedure

The main experiment consisted of two conditions: In the auditory-alone condition (A) 175 words (25 words per noise level) were presented in conjunction with a still image of the speakers face; in the audiovisual condition (AV) the speaker's face articulated another set of 175 words. Words were randomly assigned to all of the conditions and noise levels. Stimulus presentation of A and AV trials were also randomly intermixed. A subset of nine participants from the same pool of subjects received a third speech-reading-alone condition (V), where we used an additional 175 words. In this condition, the speakers face articulated the words but no auditory word-signal was present. Again, this condition occurred with all seven levels of noise and V trials were randomly intermixed with all other trials. Please note that the A condition where a visual stimulus (still image) is provided and the V condition where no auditory stimulus is present are not exactly equivalent. The crucial comparison, however, is not made between the A and the V condition but the A and the AV condition.

Participants were instructed to watch the screen and report which word they heard. If a word was not clearly understood they were asked to guess which word was presented. The experimenter recorded a response that exactly matched the target word presented as a correct answer while any difference to the target was recorded as an incorrect answer. Pacing of the experiment was under participant control; the participant initiated the next trial with a button press. The experiment consisted of five blocks with 105 words per block for participants

who received all three conditions and 70 words per block where only A and AV conditions were present.

After the experiment, participants were presented with the full list of words used in the experiment. Subjects were asked to indicate any words that they had not heard before and words that they had heard before but didn't know the meaning of. A list of pseudo-words was randomly intermixed with the words in the list as catch trials to control for the possible tendency of subjects to not report words they didn't know. This test was run to ensure that we hadn't inadvertently included any subject with unusually low vocabulary size. A cutoff criterion of 90% word identification was preset but none of the subjects fell below that criterion.

4. Results

A 2X7 repeated measures analysis of variance (RM-ANOVA) with the factors of condition (A and AV) and SNR (1-7) was employed to analyze the data. As expected, the level of noise affected recognition performance significantly, $F(6, 114) = 361.27$, $p < 0.001$, $\eta^2 = 0.99$, the lower the SNR, the fewer words that were recognized (see Fig. 1). In the auditory-alone (A) condition we can see a monotonic increase ranging from a recognition accuracy of essentially zero at an SNR of –24dB to 85% at an SNR of 0dB. An independent T-test revealed that the first SNR-level where word recognition was significantly different from zero was at –20dB, $t(19) = 3.27$, $p < 0.01$. The SNR where participants began to be able to make effective use of the auditory information must therefore be somewhere between –24dB and –20dB. Whereas participants recognized absolutely no

words in the auditory-alone condition at –24dB, performance jumped to 19% at this noise level in the AV condition. Note that speech-reading alone can only account for 8-9% of this performance boost (see below). At the other extreme (0db SNR) performance improved more modestly from 85% (A) to 95% (AV).

Overall, speech-recognition benefited substantially from the additional visual stimulation, $F(1, 114) = 77.63$, $p < 0.001$, $\eta^2 = 0.8$, and the interaction between both factors was also highly reliable, $F(6, 114) = 9.93$, $p < 0.001$, $\eta^2 = 0.78$, indicating that the performance enhancement due to visual stimulation was greater at certain SNR levels than others. This interaction held true even when we excluded the lowest (-24dB) and the highest (0dB) SNRs to account for floor and ceiling effects that might potentially have driven the interaction, $F(4, 76) = 8.22$, $p < 0.001$, $\eta^2 = 0.3$.

As noted above, only a subset of participants received the V-alone condition. In order to ensure that there was no difference between this smaller group and the rest of the population who did not receive this condition, we tested for any differences in performance between the two sub-populations on the two common conditions (A and AV). No significant effect of performance was found in the A and the AV conditions [$F(2, 17) = 0.53$, $p = 0.599$, $\eta^2 = 0.058$].

Figure 1 (top panel) displays the difference between conditions AV and A at all 7 noise levels employed. A series of protected comparisons (paired 2-tailed T-tests) revealed differences in recognition accuracy at all SNR levels and the p-values for these tests are indexed in Figure 1 by asterisks. Figure 1 (bottom panel) plots the absolute difference in recognition due to visual input (i.e. AV-A)

showing an inverted u-shaped relationship between the gain in recognition accuracy due to the additional visual stimulation and SNR. In order to characterize the gain in the AV condition we used the absolute difference in percent, in keeping with former studies of speech recognition performance (e.g. Sumby and Pollack, 1954; Erber, 1969, 1975; Callan et al., 2001, 2003). It should be noted though that the method by which gain should be appropriately characterized is a matter of ongoing debate, an issue we will return to in the discussion section below.

Using the measure of absolute gain, the largest benefit is found at the center of the curve at an SNR of –12dB with a clearly discontinuous gain of some 45% in recognition accuracy at this level. Interestingly, a post-hoc analysis revealed that age correlated significantly with speech-recognition performance during the AV condition [r = -0.21; p< 0.05] but not with A [r = -0.1; p> 0.05]. That is, in our sample, there appears to be a gradual change with age in the AV performance that is not due to hearing loss, which would be reflected, in lower auditory-alone scores. Work is presently underway to both confirm and explicate this age effect[6].

Recognition accuracy in the V condition [$\bar{x}$ = 9%; *SD*= 7%] stayed consistent over all noise levels. We also reasoned that if the gain from additional visual input was mainly dependent on speech-reading, then one would expect to find strong correlations between speech-reading performance and performance

[6] An upper age range cutoff of 60 years was used in the present study. The performance of all subjects in the A condition fell in the range of normal variation of this sample. Hence, there was no reason to exclude older subjects from the sample.

in the AV condition across all SNR levels. A Pearson product-moment correlation suggested a significant relationship between speech-reading (V) and averaged auditory-visual gain (AV-A) at the two lowest SNRs of –24 and –20 [n= 9 (r= 0.75, p< 0.05); (r= 0.74, p< 0.05)]. In contrast, no relationship was found at intermediate SNRs of -16, -12, -8 and -4dB [(r= -0.14, p= 0.71); (r= -0.21, p= 0.58); (r= 0.36, p= 0.34); (r= 0.41, p= 0.27)]. Surprisingly, a significant correlation was found at the 0dB SNR (r= 0.72, p< 0.05) which is likely due to a relationship between speech-reading ability and overall performance.

Note that results from the test of word-recognition showed that the amount of words that subjects were unfamiliar with was negligible (< 2%). All subjects identified the pseudowords correctly.

5. Discussion

Here, we explored the gain provided by visual articulatory information on the recognition of speech embedded in noise. In particular, we wished to establish the levels at which multisensory audiovisual interactions resulted in the greatest benefits. In the past it has been suggested that the gain provided by audio-visual stimulation continues to increase as the information available from the auditory modality decreases (Sumby and Pollack, 1954, Erber, 1969). Our thesis was that this finding was at least partly due to the fact that in these previous studies, subjects were exposed to the word material both before and during the experiment (by way of checklists) resulting in artificially high speech-reading

scores at low signal-to-noise ratios (SNRs). Here, we used a much larger pool of word stimuli, participants had no previous exposure to the material, words were used only once and no checklists were employed. This resulted in substantially different results than those previously reported. We found that word recognition through speech-reading was considerably poorer at low SNRs than previously shown (Sumby and Pollack, 1954, Erber, 1969). Rather, the absolute gain from AV stimulation, measured here using the original method of Sumby and Pollack (AV-A), was found to be maximal at an SNR-level of approximately –12dB or, where recognition accuracy in the auditory-alone condition was approximately 20 % (center of the curve in Fig. 2). In line with our prediction, this window of maximal integration is located between the extremes where observers have to rely mostly on speech-reading (-24dB) and where information from visual articulation is largely redundant to the auditory signal (0dB).

5.1 Alternate ways of characterizing the data

In the literature, the gain from multisensory in comparison to unisensory stimulation has been characterized in a variety of ways depending on the subject under study and the dependent measure used. Figure 2 shows several oft-used approaches applied to the present dataset, which rather unfortunately appear to lead to somewhat different conclusions concerning the locus of maximal gain. For example, Graph A in Figure 2 shows audiovisual benefit as gain in percent, clearly displaying some of the hallmarks of *inverse effectiveness*, with gain steadily increasing as SNR is lowered. The black trace in Graph A represents a

best-fit nonlinear regression that characterizes all the data points with an R^2 of 0.97. However, the data point at –12dB is not well explained by this curve fitting. Statistical testing of the residual values of each data point shows that the residual of the -12dB data point is significantly different to all other points ($z = 4.34$, $p < 0.0001$ - see Table 1). Refitting the curve excluding the data point at –12 dB (dashed orange trace) improves the fit to a near perfect R^2 of 0.997.

We would contend that this method (and to some degree, the measure of absolute difference used in the present study) is somewhat limited in its ability to appropriately characterize gain because of the inherent inverse relationship between performance in the auditory-alone condition and the maximum derived benefit when it is expressed as a percent-gain (see also Grant & Braida, 1996). In other words, there is a strong bias in favor of the benefit at lower SNRs due to a simple ceiling effect. To parse this in more concrete terms, one might ask whether it is meaningful to state that the gain at a very low SNR (e.g. A= 2%; AV=6%; gain =300%) can really be considered equivalent to the gain at a higher SNR (e.g. A= 30%; AV = 90%; gain= 300%)? That is, could an improvement in recognition from 1-in-50 words to 3-in-50 words really equate with an improvement from 12-in-50 to 36-in-50? The issue becomes particularly acute at higher SNRs where the “room for gain” diminishes rapidly (ceiling effect). For example, the 800% gain shown for a SNR of –20 dB in Graph A simply cannot be surpassed at any SNR where subjects perform above 12.5 % in the auditory-alone condition, even if subjects were to perform perfectly during the AV condition.

In the speech perception literature another definition of gain has therefore been more widely applied, one initially suggested by Sumby and Pollack (1954). It has the advantage of being independent of the auditory-alone (A) performance level and is defined as the difference score (AV-A) divided by the maximum improvement possible (100-A). This gain score is depicted in graph B of Figure 2. It shows the largest relative gain now located at the highest SNRs, in apparent opposition to the *inverse effectiveness principle.* Consistent with graph A, however, the datapoint at –12db again represents the only residual value that is significantly different from the others ($z = 3.39$, $p < 0.001$). Again, the curve fit is near perfect with $R^2 = 0.997$ if we exclude the datapoint at –12dB from the calculation.

Yet another way to characterize benefit is displayed in Graph C. Here, gain is plotted in decibels. That is, from the plot in Figure 1, one can read off the intensity level required in the auditory-alone (A) condition to achieve the same recognition accuracy seen in the AV condition at any given SNR. For example, from the graph, we can estimate that at –12 dB the gain from visual stimulation amounts to an increase in SNR of approximately 9dB, since AV recognition at –12dB is 69% and estimating the equivalent performance point on the auditory-alone curve leads to an estimate of –3dB SNR. As with the previous two methods, the datapoint at –12 dB is again the only point that is not well fit by the curve, with a significant residual deviation ($z = 2.0$, $p = 0.046$). Leaving out the –12dB datapoint as before improves the fit from an $R^2 = 0.71$ to an $R^2 = 0.92$. It should be noted that gain derived in this manner is likely to be dependent on the

specific experimental parameters and stimulus materials used. For instance, Sumby and Pollack's data showed an audiovisual gain that ranged from 5 to 22dB depending on the size of the set of bisyllabic words used. MacLeod &. Summerfield (1987) estimated a gain of 11dB for the contribution of vision to the perception of spoken sentences. Here, gain is maximal at the lowest SNR, and falls off with increasing SNR in general accordance with *inverse effectiveness*, but as before, there is once again a departure from a simple monotonic function at the –12dB SNR.

In sum, while the approach to appropriately characterizing gain is still a matter of some disagreement in the literature, no matter which method is applied here, the gain around the –12 dB data point appears to represent a "special zone" for audiovisual multisensory integration that does not accord with a strict interpretation of *inverse effectiveness*.

5.2 Previous neuroimaging studies

It is also noteworthy that the window of maximal integration in our study (at 20% intelligibility for the auditory-alone condition or –12dB) is located precisely where Callan et al. (2003) found increased activation in STS in their fMRI study. However, these authors interpreted their findings as evidence for the operation of *inverse effectiveness* in STS, whereas the present results show that the size of the gain due to AV interactions actually diminishes below this level. Given the present results we would predict that at lower SNRs, multisensory regions like the STS would likely show less AV multisensory integration. However, it is worth

stressing that significant and considerable AV enhancements continue to be present all the way down to the highest noise level used, even at a noise level at which words were completely unintelligible in the auditory-alone condition. This last point is remarkable in that it shows that under circumstances where word-recognition is impossible on the basis of the auditory input alone, adjunctive visual input can cause substantial recovery of function beyond what is possible by speech-reading alone. Thus, although we would predict a drop in STS AV activation at lower SNRs, we would also predict that multisensory enhancement will persist at noise levels considerably higher than those used in the Callan (2003) study.

Further, the present data clearly show that only part of the performance gain introduced by visual articulation can be attributed to pure speech-reading. It was found that the gain from additional visual stimulation was correlated with speech-reading performance at low SNRs where the auditory signal was low and speech-reading is likely to play an important role in the recognition of words in addition to any multisensory effect. At intermediate levels, however, where by far the greatest multisensory gain was seen, performance was not correlated with speech-reading scores.

5.3 Previous behavioral studies of audio-visual speech integration

As was pointed out in the introduction, a number of previous behavioral studies have investigated audio-visual gain in speech recognition. Here, we go into more detail regarding how the present results relate to this previous

literature. In an early study by O'Neill (1954), increased gain with decreasing SNR was found. However, SNRs stopped at a nominal –20dB where fully 25% of the words were still recognizable in the auditory-alone condition. As such, this study cut off at almost exactly the same SNR level at which we also found maximal gain and it is probable, had O'Neill used lower SNRs, that gain would have begun to decrease as was found here. Binnie et al. (1974**)** claimed to show the *"greatest visual complement occurring at poorer SNRs"*. Unfortunately, this contention was simply not supported by their data, where a closer look shows that word recognition was not greatest at the lowest SNR used (41% correct at -18dB) but occurred at a more intermediate level (50% at –12dB)[7]. In a similar vein, McCormick (1979) states that the contribution of vision is inversely related to the SNR, again despite the fact that gain is maximal in his data where intelligibility in the auditory-alone condition is at 24% and clearly decreases at lower SNRs.

A number of other studies have also investigated the contribution of vision to speech perception in noise with more complex stimuli such as sentences (e.g. MacLeod and Summerfield, 1987, Grant and Seitz, 2000). In the study by McLeod and Summerfield (1987) an average audiovisual benefit of 11 dB was found. This study was explicitly designed to overcome the difficulties associated with the use of percent-correct performance as a dependent measure, which can often give rise to floor or ceiling effects. Performance was assessed using the threshold, measured in decibels, at which a number of target words were

[7] Note that speechreading performance in this study was very high (43%) due to the use of a very limited list of simple phonemic stimuli.

detected within a given sentence. Audiovisual gain was then calculated as the difference in decibels between the auditory-alone and audiovisual conditions. This design precludes assessment of audiovisual gain across a spectrum of SNRs and therefore does not directly address the issue of interest here[8]. Grant and Seitz (2000) also found that visible speech improved the detection of speech but likewise, audiovisual gain was not assessed across different SNRs. These authors used a very limited stimulus set of only three target sentences. A general problem with the use of sentences is that contextual cues such as prosody, syntax and context are likely to influence detection and recognition, factors that are difficult to control between conditions. In an earlier study by Grant and Braida (1991) subjects had to recognize target words in sentences over a variety of SNRs in wideband noise and filtered speech. Although the primary goal of the study was not the investigation of the locus of the AV benefit, the authors conclude that their findings are consistent with previous studies (O'Neill, 1954, Sumby and Pollack, 1954, Massaro, 1987) indicating that the absolute contribution of speech- reading to audition is maximal when the auditory channel is greatly degraded. A closer look at the performance curves (% correct) for their A and AV conditions, however, reveals that the gain is maximal at about –10dB and decreases at lower SNRs.

Taking these earlier behavioral studies together, we believe that a somewhat consistent pattern emerges. Some who have claimed that the greatest

[8] One serious drawback of this method is that the same sentence has to be presented repeatedly starting at different SNRs for audiovisual and auditory-alone conditions, possibly confounding thresholds between conditions

gains are to be found at the lowest SNRs actually didn't test sufficiently low SNRs to warrant such a contention, whereas others who made similar claims are often not supported by their own data.

6. Conclusions

These results show that there is a delimited window within which visual speech signals enhance auditory speech comprehension maximally and that this occurs, not when the unisensory (auditory) input is weakest as would be the prediction under a strict interpretation of the '*inverse effectiveness*' principle, but rather, at more intermediate values. Nonetheless, it is clear from these data that substantial gain in multisensory speech enhancement is found at even the lowest SNRs, largely in keeping with the general predictions of *inverse effectiveness*, but it is also evident that there is an optimally tuned window, a 'special zone' if you will, around an SNR of –12dB where multisensory integration is additionally enhanced beyond the predictions of this principle. We would like to clarify that the originators of the *inverse effectiveness* principle did not explicitly predict that it would also apply to higher-level functions such as speech recognition (Stein and Meredith, 1993), but speech researchers have certainly applied the principle since then (e.g. Callan et al., 2001, 2003; see Calvert and Lewis, 2004) and early AV speech researchers implied a similar mode of operation (e.g. Sumby and Pollack, 1954; Erber, 1975). Of course, the principle of *inverse effectiveness* was devised based on observations in a subcortical structure specialized for detection

and orientation. That is, the job of the SC is thoroughly different to that of the speech recognition system. Since the SC is specialized for initiating saccades to events or objects and is an essential component of the brain's orienting (and early warning system), *inverse effectiveness* for weak unisensory inputs is a highly useful property in this structure. The speech system, on the other hand, is not primarily concerned with detection. Rather, its purpose is recognition, which involves the classification of complex waveforms and their integration into semantically meaningful units.

This, however, does not imply that all cortical multisensory mechanisms will not obey this principle. For example, recent electrophysiological studies in both humans (e.g. Foxe et al., 2000; Molholm et al., 2002, 2004; Murray et al., 2005) and non-human primates (e.g. Schroeder and Foxe, 2001, 2004; Schroeder et al., 2004; Foxe and Schroeder, 2005) have shown extremely rapid multisensory interactions in early sensory processing regions, and these early interactions may well be involved in detection and orientation processes, like those of the SC, where *inverse effectiveness* would be highly advantageous. However, this remains to be directly assessed.

It is reasonable to propose that there are minimal levels of auditory input necessary before recognition can be most effectively enhanced by concurrent visual input. These data suggest that the speech recognition system appears to be maximally tuned for multisensory integration at SNR levels that contain these minimal levels of input – that is, there is a window of maximal multisensory integration at intermediate levels. As such, we contend that maximal

multisensory tuning for speech-recognition does not strictly adhere to the *inverse effectiveness* principle.

Table 1.

Table shows observed, predicted and residual values for the fitted curves (black traces) of the three gain estimates (Gain in %, AV-A/100-A and gain in dB). *Z*-scores are derived from the means and the standard deviations of the residuals. Significant *z*-scores are marked in bold.

Table 1: Test of the residual values

	Observed	Predicted	Residual	z-score	p-value
Gain in %					
-20	766	759.85	6.42	0.15	0.88
-16	200	246.56	-46.56	-1.23	0.23
-12	**187**	**102.04**	**84.96**	**4.34**	**0.0000**
-8	66	61.33	4.67	0.11	0.91
-4	35	49.86	-14.86	-0.35	0.72
0	12	46.63	-34.63	-0.86	0.39
AV-A/100-A					
-24	19	17.47	1.53	0.28	0.78
-20	24	24.07	-0.07	-0.01	0.99
-16	27	34.65	-7.65	-1.66	0.1
-12	**59**	**48.01**	**10.99**	**3.39**	**0.001**
-8	55	60.47	-5.47	-1.07	0.28
-4	67	67.19	-0.19	-0.03	0.97
0	66	65.15	0.85	0.18	0.88
Gain in dB					
-24	10.8	10.16	0.64	0.42	0.68
-20	8.5	8.99	-0.49	-0.55	0.58
-16	6	7.56	-1.56	-1.86	0.06
-12	**8.7**	**6.77**	**1.93**	**2.0**	**0.046**
-8	6.2	5.12	1.08	0.83	0.41
-4	4	4.71	-0.71	-0.75	0.45

Figure 1.

The top panel depicts the percentage of correctly identified words (% correct) depending on the signal to noise ration (SNR) for the auditory-alone (A: dashed line) and the audiovisual (AV: solid line) conditions. Significant differences between both conditions are indexed with stars ($p < 0.05$*; $p < 0.001$***). The bottom panel shows the multisensory gain as the difference (AV-A) in speech-recognition accuracy as a function of level of SNR (solid line). The dotted line represents performance in pure speech-reading (V) in percent correct.

Figure 1: Results

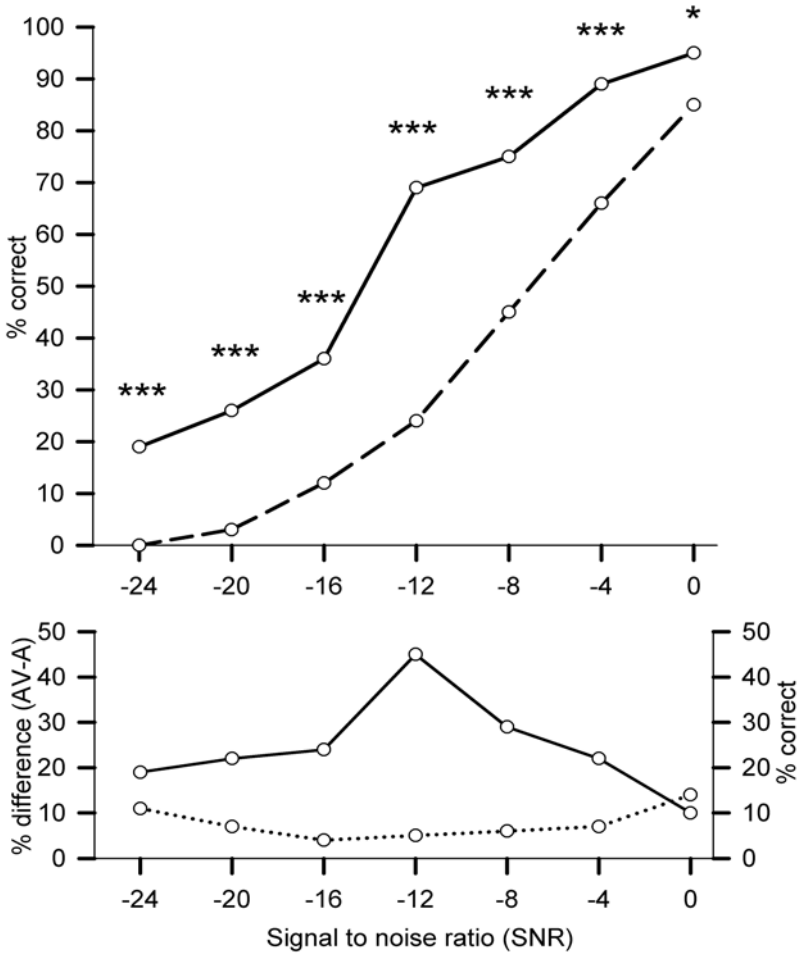
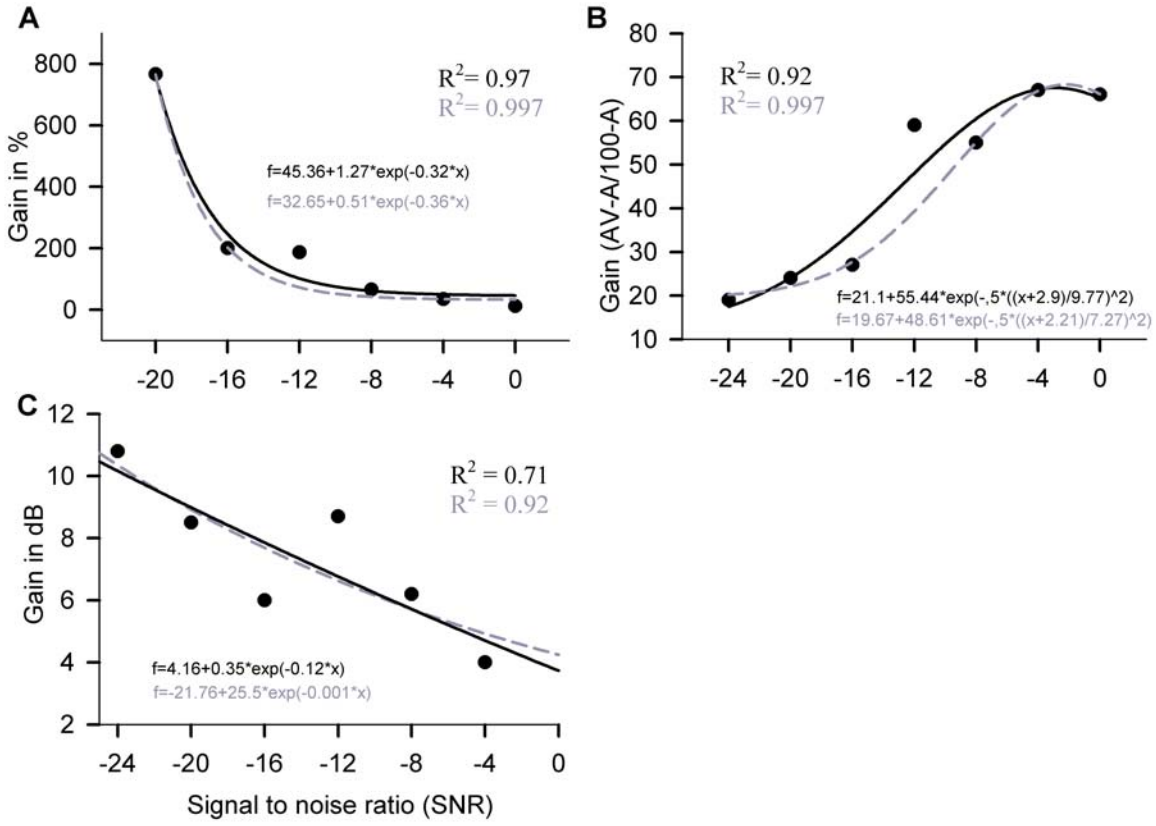


Figure 2.

Graphs show fitted curves (black traces) for three additional methods of defining the audio-visual gain function. Graph A shows gain in percent ((AV-A)*100/AV). Graph B displays gain corrected for the ceiling effect (AV-A/100-A) and Graph C shows gain in dB. The grey-colored dashed traces represent the curve fits when the value at -12 dB is excluded from the calculation. Note that in Graph A, the –24dB datapoint is excluded since performance in the auditory-alone condition was at zero.

Figure 2: Ways of characterizing audiovisual benefit in our data



References

Bernstein, L. E., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? In G. A. Calvert, C. Spence, & B. E. Stein, (Eds.),*The handbook of multisensory processes* (pp. 203-223). Cambridge, MA: Bradford, MIT Press.

Binnie, C. A., Montgomery, A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *J Speech Hearing Res, 17,* 616.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr Biol, 1011,* 649-657.

Calvert, G., Lewis, & W. L. (2004). Hemodynamic studies of audiovisual interactions. In G. A. Calvert, C. Spence, & B. E. Stein, (Eds.), *The handbook of multisensory processes* (pp. 203-223). Cambridge, MA: Bradford, MIT Press.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *J Cogn Neurosci*, 1, 151: 57-70.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodalcontribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Brain Res Cogn Brain Res, 10,* 349-353.

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

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hearing Res, 12,* 423-425.

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *J Speech Hearing Res, 143,* 496-512.

Erber, N. P. (1975). Auditory-visual perception in speech. *J Speech and Hearing Dis, 40,* 481-492.

Foxe, J. J., Morocz, I. A., Higgins, B. A., Murray, M. A., Javitt, D. C., & Schroeder, C.E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing. *Brain Res Cogn Brain Res, 10,* 77-83.

Foxe, J. J., Schroeder, C. E. (2005). The case for a feedforward component in multisensory integration mechanisms. *NeuroReport, 16,* 419-423.

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *J Acoust Soc Am, 89,* 2952-2960.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *J Acoust Soc Am, 100,* 2415-2424.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am, 108,* 1197-1208.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological Inquiry*. Hillsdale, NJ: Erlbaum.

McCormick, B. (1979). Audio-visual discrimination of speech. *Clin Otolaryngol Allied Sci, 45,* 355-361.

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

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Res Cogn Brain Res, 369,* 350-354.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Brain Res Cogn Brain Res, 14,* 121-134.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cereb Cort, 144,* 452-465.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol Sci, 152,* 133-137.

Murray, M. M., Molholm, S., Michel, C. M., Ritter, W., Heslenfeld, D. J., Schroeder, C. E., Javitt, D. C., & Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cereb Cort, 157,* 963-974.

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *J Acoust Soc Am, 26,* 212.

O'Neill, J. J. (1954). Contributions of the visual component of oral symbols to speech comprehension. *J Speech and Hearing Dis, 19,* 429.

Schroeder, C. E., & Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the Macaque neocortex. *Brain Res Cogn Brain Res, 14,* 195-207.

Schroeder, C. E., & Foxe, J. J. (2004). Multisensory convergence in early cortical Processing. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 295-309). Cambridge, MA: Bradford, MIT Press.

Schroeder, C. E., Molholm, S., Lakatos, P., Ritter, W., & Foxe, J. J. (2004). Human-simian correspondence in the early cortical processing of multisensory cues. *Cog Proc, 53,* 140-151.

Spence, C., & Driver, J. (2000). Attracting attention to the illusory location of a sound: Reflexive crossmodal orienting and ventriloquism. *NeuroReport, 11,* 2057-2061.

Stein, B. E., Huneycutt, W. S., & Meredith, M. A. (1988). Neurons and behavior: The same rules of multisensory integration. *Brain Res Cogn Brain Res, 448,* 355-358.

Stein, B. E., Meredith, M. A. (1993). *The merging of the senses*. Cambridge, Massachusetts: MIT Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The J Acoust Soc Am, 26,* 212-215.

CHAPTER III

Impaired Multisensory Processing in Schizophrenia: Deficits in the Visual Enhancement of Speech Comprehension Under Noisy Environmental Conditions.

Lars A. Ross [a,b,c], Dave Saint-Amour [b,d],

Victoria M. Leavitt [b,e], Sophie Molholm [a,b],

Daniel C. Javitt [a,b], and John J. Foxe [a,b,e,*]

[a] *Program in Cognitive Neuroscience, Department of Psychology*
The City College of the City University of New York
138th St. & Convent Avenue
New York, New York 10031, USA

[b] *The Cognitive Neurophysiology Laboratory*
Nathan S. Kline Institute for Psychiatric Research
Program in Cognitive Neuroscience and Schizophrenia
140 Old Orangeburg Road
Orangeburg, New York 10962, USA

[c] *Ramapo College of New Jersey*
505 Ramapo Valley Road
Mahwah, New Jersey 07430, USA

[d] *Département d'ophtalmologie*
Université de Montréal
C.P. 6128 Succ. Centre Ville
Montréal, Québec, H3C 3J7

[e] *Program in Neuropsychology, Department of Psychology*
Queens College of the City University of New York
65-30 Kissena Boulevard
Flushing, New York 11367, USA

Published in the journal Schizophrenia Research in Dec. 2007, 97, 173-183.

1. Abstract

Background: Viewing a speaker's articulatory movements substantially improves a listener's ability to understand spoken words, especially under noisy environmental conditions. In this study we investigated the ability of patients with schizophrenia to integrate visual and auditory speech. Our objective was to determine to what extent they experience benefit from visual articulation and to detail under what listening conditions they might show the greatest impairments.

Methods: We assessed the ability to recognize auditory and audiovisual speech in different levels of noise in 18 patients with schizophrenia and compared their performance with that of 18 healthy volunteers. We used a large set of monosyllabic words as our stimuli in order to more closely approximate performance in everyday situations.

Results: Patients with schizophrenia showed deficits in their ability to derive benefit from visual articulatory motion. This impairment was most pronounced at signal-to-noise levels where multisensory gain is known to be maximal in healthy control subjects. A surprising finding was that despite known early auditory sensory processing deficits and reports of impairments in speech processing in schizophrenia, patients' performance in unisensory auditory speech perception remained fully intact.

Conclusions: Thus, the results showed a specific deficit in multisensory speech processing in the absence of any measurable deficit in unisensory speech processing and suggest that sensory integration dysfunction may be an important and, to date, rather overlooked aspect of schizophrenia.

2. Introduction

The integration of heard speech signals with the seen articulatory movements of a speaker's face and mouth is essential for everyday communication, as seeing a speaker's face substantially facilitates the recognition of spoken words, especially under noisy listening conditions (e.g. Erber, 1969; Grant, 2000; Munhall et al., 2004a; O'Neill, 1954; Ross et al., 2006; Sumby and Pollack, 1954). The brain processes underlying this multisensory speech integration are presently under intense investigation (Bernstein et al., 2004; Calvert, 2001; Calvert and Campbell, 2003; Campbell, 2004; Callan et al., 2003; Munhall, 2002, 2004b) and investigators have now begun to explore whether there is a specific role for multisensory processes in some of the perceptual deficits seen in disorders such as autism (Iarocci and McDonald, 2006a; Kern, 2002) and schizophrenia (de Gelder et al., 2003). In schizophrenia, past research has established the existence of robust deficits within-modality where early auditory and visual sensory processing has been shown to be impaired (e.g. Butler, 2006; Foxe et al., 2001, 2005; Schwartz et al., 2001). Given these early unisensory deficits, there is good reason to predict that multisensory processes, which clearly rely on the fidelity of early sensory inputs from the respective unisensory systems, will show similar if not greater impairment. Given extensive physiological evidence that multisensory integration can act as a non-linear gain mechanism (Foxe and Schroeder 2005; Meredith and Stein, 1986; Molholm et al., 2004, 2006; Stein et al., 2001, 2002; Schroeder and Foxe, 2005), it seems a

reasonable prediction that impairment of multisensory processing might well be especially impaired in this population[9].

Indeed, recent evidence does suggest processing deficits in schizophrenia. In a cleverly constructed study, DeGelder and colleagues (de Gelder et al., 2003) used a variant of the so-called "McGurk illusion"(McGurk and MacDonald, 1976; Saint-Amour et al., 2006) to assess whether patients have deficits in integrating auditory and visual speech. For the reader unfamiliar with the McGurk illusion, the following example will be helpful. When participants attend to a video of a speaker articulating the syllable /ga/ while listening to the incongruent auditory syllable /ba/, the listener typically reports the perception of the fused syllable /da/, and this occurs despite that fact that the /da/ syllable was neither heard nor seen. There are numerous other examples of these phonemic fusions and the illusion is very strong such that even when the listener is fully apprised of the 'trick,' it is difficult or even impossible to suppress it (Massaro, 1998). In DeGelder's experiment, patients were much less susceptible to these illusory fusions than healthy participants, whereas performance in an audiovisual control task involving spatial localization of sounds remained unimpaired. The authors hypothesized that if there was a general deficit in multisensory integration, patients would show decrements in both tasks. The results favored the notion of an isolated deficit related to the integration of phonetic information. However, somewhat contradictory evidence comes from a study by Surguladze

[9] Note that the authors do not mean to imply that sensory integration deficits are peculiar to schizophrenia. They have also been implicated in a number of other clinical populations such as in autism (see Iarocci and McDonald, 2006) and in certain neurological patients (e.g. Rorden et al., 1999; Munhall et al., 2002).

et al., where schizophrenia patients and controls showed similar susceptibility to fusions in a McGurk-type experiment (Surguladze et al., 2001).

In both of these previous studies, the premise was that susceptibility to McGurk fusions would index an intact audiovisual integration system, although it should be also be pointed out that a small proportion of healthy control subjects do not experience McGurk fusions. Nonetheless, the vast majority of normal observers do in fact perceive these fusions and so these studies took advantage of this fact to assess whether, on average, patients would experience lower levels of fusion. The McGurk-task, however, where mostly simple syllables are used, could be considered a rather indirect and non-ecological means of assessing multisensory performance[10]. Due in part to the rather artificial nature of the McGurk-task, we reasoned that testing patients with schizophrenia on an audiovisual task using real words as opposed to syllables would provide a better test of their abilities for audiovisual integration of speech in real-life situations. More importantly, it has also been suggested that an impairment in auditory speech recognition in general (Hoffman et al., 1999; Lebib et al., 2003) and the integration of auditory and visual speech in particular (Surguladze et al., 2001) is most likely to manifest itself in situations where the auditory signal is degraded. We would therefore expect a deficit in speech processing to predominate when patients are asked to identify speech under noisy environmental conditions that are more typical of normal everyday social situations.

[10] It should be mentioned that the McGurk effect has also been shown using words (Dekle et al., 1992) but that only syllables have been used in patients.

Furthermore, we expected to find the most robust deficit in multisensory speech perception under environmental conditions where healthy control subjects usually experience the most benefit from seeing the speaker's articulations. In a recent experiment from our laboratory (Ross et al., 2007), we showed that the gain derived from viewing visual articulations is maximal at intermediate signal-to-noise ratios (SNRs) in healthy volunteers. Here, we investigated the ability of patients with schizophrenia to integrate visual and auditory speech. Our objective was to determine to what extent they experience benefit from visual articulation and to detail under what listening conditions (SNRs) they might show the greatest impairments. For that, we assessed their ability to recognize auditory and audiovisual speech in different levels of noise and compared their performance with that of healthy volunteers. We used a large normed set of monosyllabic words as our stimuli in order to more closely approximate performance in everyday situations without delivering semantic, grammatical or prosodic context.

3. Methods

3.1 Subjects

Informed consent was obtained from 18 patients (1 woman, mean age: 39, SD: 10.6) meeting the DSM-IV criteria for schizophrenia (n=15) or schizoaffective disorder (n=3) and 18 healthy volunteers (7 women, mean age: 35, SD: 11.6) at the Nathan Kline Institute (NKI) for Psychiatric Research (Orangeburg, NY). NKI's Institutional Review Board approved all procedures. Please refer to Table 1

for the sample characteristics of the patients with schizophrenia. All patients and controls had normal or corrected-to-normal vision and reported normal hearing. Patients' diagnoses were obtained using the Structured Clinical Interview for DSM-IV (First et al., 1997) and all available clinical information. All patients were receiving antipsychotic medications at the time of testing. Chlorpromazine equivalents were 1194 ± 435 mg per day. Equivalents were calculated using conversion factors described previously (Hyman et al., 1995; Peuskens and Link, 1997; Jibson and Tandon, 1998; Woods, 2003). None of the healthy volunteers had a history of Axis I psychiatric disorder as defined by the DSM IV.

3.2 Stimuli

Stimulus material consisted of 525 simple monosyllabic words (taken from the MRC Psycholinguistic database). The words were selected from a well-characterized normed set (Kucera and Francis, 1967) based on their written-word frequency. The face of a female speaker was digitally recorded articulating the words. These movies were digitally re-mastered, so that the length of the movie (1.3 sec) and the onset of the acoustic signal were highly similar across all words. Average voice onset occurred at 520ms after movie onset (SD= 30 ms). The words were presented at 50 dBA FSPL. Seven different levels of pink noise were presented simultaneously with the presentation of the words at 50, 54, 58, 62, 66, 70 and 74 dBA FSPL. Noise onset at the same time as the movie onset, 520 ms before the voicing began. The signal-to-noise ratios (SNRs) were therefore 0, -4, -8, -12, -16, -20 and –24 dBA FSPL. The movies were presented

on a 21-inch computer monitor at a distance of 1.7m from the participant. The whole face of the speaker was visible and extended 6.3° horizontally and 7.6° vertically. The words were presented from a speaker situated in the center on top of the screen and the noise was presented from speakers flanking both sides of the screen.

3.3 Procedure

The main experiment consisted of two conditions. In the auditory-alone condition (A) 175 words (25 words per noise level) were presented in conjunction with a still image of the speaker's face; in the audiovisual condition (AV) the speaker's face articulated another set of 175 words. Words were randomly assigned to all of the conditions and noise levels and reassigned several times to all conditions across subjects during the course of data collection. Stimulus presentation of A and AV trials were also randomly intermixed. A subset (from the same pool of subjects) of nine controls and three patients received a third speechreading-alone condition (V), where we used an additional 175 words. In this condition, the speaker's face articulated the words but no auditory word-signal was present. Again, this condition occurred with all seven levels of noise and V trials were randomly intermixed with all other trials.

Participants were instructed to watch the screen and report which word they heard. The experimenter assured that eye fixation was maintained by reminding participants, if necessary. If a word was not clearly understood they were asked to guess what word was presented. The experimenter was seated at

approximately 1.7m distance from the participant at a 45° angle to the participant- screen axis. The experimenter recorded a response that exactly matched the target word presented as a correct answer while any other response was recorded as an incorrect answer. Pacing of the experiment was under participant control by initiating the next trial with a button press. The experiment consisted of five blocks with 105 words per block for participants who received all three conditions and 70 words per block where only A and AV conditions were present.

After the experiment, participants of both groups were presented with the full list of words used in the experiment. Subjects were asked to characterize the words in terms of familiarity and knowledge of meaning. A list of pseudo-words was randomly intermixed with the words in the list as catch trials to control for the possible tendency of subjects to not want to report words they didn't know. This test was run to ensure that we had not inadvertently included any subject with unusually low vocabulary size. Word knowledge below 90% of the words in the list was used as a cutoff criterion to exclude subjects from the analysis. None of the subjects fell below that criterion and no significant differences in word knowledge were found between groups.

4. Results

A 2X7X2 repeated measures analysis of variance (RM-ANOVA) with the factors of condition (A and AV) and SNR level (1-7) and the between groups factor patients (P) vs. controls (C) was employed to analyze the data. Overall, the level

of noise affected recognition performance significantly in both conditions, $F(1, 34) = 1871.4$, $p < 0.001$, $\eta^2 = 0.98$; the lower the SNR, the fewer words that were recognized (see Fig. 1). In the auditory-alone (A) condition we can see a monotonic increase ranging from a recognition accuracy of essentially zero at an SNR of –24dB to 85% at an SNR of 0dB. Patients and controls showed no differences in the A condition.

In both groups, speech-recognition benefited substantially from the additional visual stimulation, $F(1, 34) = 136.9$, $p < 0.001$, $\eta^2 = 0.8$. However, AV-curves differed between groups which was reflected by a significant interaction between the factors of condition and group $F(1, 34) = 4.06$, $p = 0.05$, $\eta^2 = 0.11$. A series of protected comparisons (two tailed t-tests, $\alpha = 0.05$) between groups at each SNR level revealed that this difference is of significant magnitude at the intermediate SNR (-12dB) $t(34) = 2.08$, $p < 0.05$. At this SNR the difference between performance in the A condition and the AV condition was maximal as the bottom panel in Fig. 1 shows. Looking more closely at this point of maximal gain for each group, one can see that control subjects showed a gain of some 46% in recognition accuracy, going from approximately 25% to 71% performance. Patients on the other hand, showed a more modest gain of 29%, improving to just 56.5% performance in the multisensory condition. Data on speechreading performance were unfortunately only obtained for 3 patients (Mean over 7 noise levels: 8.7%, SD: 5.67) and 9 controls (Mean over 7 noise levels: 10.7%, SD: 3.32). No trend in differences was observed between the groups. Note that the mean performance value for the small sample of patients

falls well within the 95% confidence interval around the mean speechreading performance level found in controls, which ranges from 4.4% to 16.9%. In addition, a two-tailed Pearson correlational analysis revealed no overall relationship between speechreading ability and AV-gain in control subjects ($p = 0.22$).

In the control population the gain at -12dB represents a significant deviation from the gain at other SNR's, representing a maximal window of integration (Ross et al., 2007). In a *post-hoc* analysis, we tested here whether the gain at this SNR was also significantly more prominent in the patient population. For this, we compared the gain at -12dB with the average gain at the two flanking levels (-8dB and -16dB) where the next two highest gains were observed. A simple 2-tailed paired sample t-test was employed ($\alpha = 0.05$). As expected, the difference was significant for control subjects ($t(17) = 4.71$, $p < 0.001$) but this was also the case for patients, albeit to a lesser degree ($t(17) = 2.331$, $p = 0.032$). Thus, although patients show a significant impairment at this level relative to controls, they do appear to show residual specialized tuning at this intermediate SNR, a result that will bear replicating.

Finally, we assessed whether there was a relationship between the strength of antipsychotic agents that the patients received as measured in chlorpromazine equivalents (Range: 500- 2000 mg; Mean: 1196 mg) and recognition performance in the AV- condition. Chlorpromazine equivalents were available for 16 of the patients. Pearson bivariate correlation coefficients (two-

tailed, α = 0.05) did not reach significance at any of the seven noise levels with an average p-value of 0.45.

5. Discussion

Here, we set out to assess the integrity of multisensory audiovisual processing in patients with schizophrenia, with an emphasis on determining whether patients would benefit similarly to healthy controls by seeing speakers' articulations while trying to recognize spoken words embedded in various levels of background noise. A rather surprising finding was that despite very well-characterized deficits in early unisensory auditory processing (e.g. Javitt et al., 1993, 1995, 1997,1998, 2000b; Michie et al., 2002; Rosburg et al., 2004; Salisbury et al., 2002) and quite a number of reports of deficits in aspects of speech perception (e.g. Baltaxe and Simmons, 1995; Bull and Venables, 1974; Cannon et al., 2002; Condray, 2005; Lee et al., 2004; Leitman et al., 2005; Titone and Levy, 2004), patients in this study did not show any deficits whatsoever in recognizing auditorily presented words when they were embedded in noise. On the other hand, the data do show that patients have impairment in the gain that is commonly observed when auditory information is accompanied by visual articulation, an impairment that is most pronounced at an SNR-level where the gain of audiovisual stimulation is known to be maximal in healthy control subjects (Ross et al., 2007). Thus, the results show a specific deficit in multisensory speech processing in the absence of any measurable deficit in unisensory (auditory-alone) speech processing and

suggest that sensory integration dysfunction may be an important and, to date, rather overlooked aspect of schizophrenia.

5.1 Possible Neural Substrates

There are several mechanisms involving different brain structures that have been suggested to underlie the successful integration of audiovisual speech and possible reasons for its malfunction are therefore manifold. One of them is the visual biological motion processing system, which has already been shown to be impaired in schizophrenia (Kim et al., 2005). In contrast to more basic motion stimuli, which are mainly processed in area V5, biological motion contains information about the identity of the moving stimulus, actions, intentions and even emotions. As a necessary sub-process for audiovisual integration, lip-motion obviously represents such a biological motion process. In fact, the anatomical substrate of biological motion, including lip-motion, is found in the posterior superior temporal cortex (STS/G) (Allison, 2000), and this region has also been shown to contain abnormalities in schizophrenia (Shenton, 2001).

In humans and primates this region is involved in the analysis of complex biological stimuli, such as hand motion and communication and their imitation (Iacoboni and Rizzolatti, 2001). It receives projections from the medial superior temporal (MST) motion processing area and is therefore part of a functional network that uses motion information provided by the dorsal visual stream for the identification of complex motion, which would clearly include articulatory movement of the lips. The dorsal visual stream has extensive magnocellular

inputs that rapidly conduct low-resolution visual information to cortex (e.g. V5 receives direct input from V1) providing information about motion and spatial organization of stimuli. Patients with schizophrenia have shown clear dysfunction in tasks involving the magnocellular system (see e.g. Kim et al., 2006; Schwartz et al., 2001) and electrophysiological studies have shown early dorsal visual stream processing deficits in schizophrenia (Doniger et al., 2002; Foxe et al., 2001; Foxe et al., 2005; Yeap et al., 2006; Butler et al., 2007) that may well be the origin for impairment of "upstream" function involving motion perception (Chen et al., 2005), biological motion (Kim et al., 2005) and consequently audiovisual integration of speech. As such, dysfunction of dorsal stream visual processing, and by inference, of biological motion processing, might well be a source of upstream deficits in the multisensory integration of speech. However, such links remain speculative and will need to be explicitly tested in future studies.

It should also be noted that results from speechreading studies in patients, including the small sample tested here, are not entirely consonant with this hypothesis, with research in this area producing somewhat non-uniform results. DeGelder and colleagues did find reduced ability in their patients to identify syllables that were speechread (de Gelder et al., 2003). This deficit, however, did not correlate with audiovisual integration as measured by fusions in the McGurk illusion, and so the authors concluded that the multisensory integration decrement found in their study was not due to reduced speechreading ability. A more detailed examination of their results is merited here. The speechreading

decrement for patients in their study amounted to 14% accuracy in comparison to healthy subjects. If a speechreading deficit of similar magnitude existed in our patient group, we could clearly infer that speechreading would only account for a minute proportion of the observed audiovisual integration deficit we find. In our study, healthy controls, and for that matter the three patients we tested, recognized 8.7 % of the words when delivered in the visual modality alone (i.e. when they were speech read). A decrease in accuracy by 14% (i.e. from 8% down to 7%) would be insufficient to explain the difference in multisensory gain of 17% that we did find between patients and controls at the –12dB noise level. Further, if a deficit in speechreading was the source of the patients' failure to experience normal gain from visual input then one might expect such a deficit to impact gain at all SNR levels to the same extent. The gain curve in Figure 1 shows otherwise, where the gain at higher SNRs (i.e. –8dB, -4dB and 0dB) shows no differences between patients and controls. (It should be noted here, however, that at 0dB SNR gain was restricted by AV- performance reaching ceiling levels). Further, our study suggests no or, at best, modest loss of speechreading ability, which again, would not be sufficient to explain a 17% loss in AV- gain.

It is also the case that other studies have found no deficit in speechreading ability at all for patients with schizophrenia. In a carefully screened set of patients, Myslobodsky and colleagues found no differentiation between patients and controls when they were asked to speech read spoken words (Myslobodsky et al., 1992). The authors attribute a mild deficiency in the

speechreading of spoken sentences in their patients to a secondary coping strategy rather than to speechreading itself. Similarly, only very small differences were found between patients and controls in speechreading (Schonauer et al., 1998). Although in the present study only three patients were tested for speechreading, they also did not show any obvious deficit. In a similar vein, Cienkowski and Carney (2002) showed that speechreading ability in both young and elderly subjects was unrelated to audiovisual integrations in a McGurk task. Gagne et al. (1995) investigated the effect of conversational versus "clear" speech on speech intelligibility and reported no correlations between performance in the visual-alone and audiovisual conditions. In line with these findings, we also found no relationship between speechreading performance and AV- gain in either patients or controls. Overall, it seems highly unlikely in our opinion that an isolated deficit in speechreading could be the major source of the multisensory deficit observed here.

So, if we can largely rule out that the present deficit is primarily driven by unisensory visual deficits, and since it is also clear that there is no deficit at all in unisensory auditory speech recognition under the present circumstances, the inevitable conclusion is that these data have uncovered an isolated multisensory integration deficit for audiovisual speech recognition. This in turn suggests that the neural substrate of this deficit is likely to be a higher-order speech integration region, of which the superior temporal sulcus/gyrus (STS/G) would appear to be the likeliest candidate (Calvert and Campbell, 2003; Surguladze et al., 2001).

One obvious avenue for further study would be to assess STS/G activity as a function of the extent of the deficit in audiovisual integration seen in patients.

A reviewer of an earlier version of this paper raised the important issue of whether the decrements seen in audiovisual gain might be due in part to a failure to maintain proper fixation on the face during the task. The experimenters closely monitored eye-fixation throughout the experiment and no systematic deviations in eye fixation between patients and controls were observed. Fixation and attention during audiovisual speech processing is a rather complex issue. In various listening conditions, the perceiver's eye gaze moves over the face, predominantly located at the mouth, nose and eyes (Vatikiotis-Bateson et al., 1998) with gaze patterns changing depending on the listening conditions (Buchan et al., 2007). It therefore would have been inappropriate to instruct participants to focus on a specific location on the face of the speaker. However, if gaze position were at the root of the deficit seen here, it should also have resulted in substantial deficits in speechreading, something that was not seen in the small sample tested here, and also not seen in a number of previous studies as discussed above. It should also be mentioned that a difference in gaze patterns might not necessarily explain AV gain deficits as AV gain is observed at varying gaze locations on the face. For example, Paré and colleagues (2003) have shown that oral foveation is not necessary for processing visual speech information. Lastly, if gaze were the cause of the deficits seen here, one would clearly expect to see deficits more evenly distributed across the various signal-to-noise ratios used rather than concentrated at the –12dB level.

5.2 The Optimal Integration Window

What remains to be discussed is the fact that patients with schizophrenia show the largest deficit at the “intermediate” SNR (–12dB). In an earlier study (Ross et al., 2007) we showed that healthy volunteers experience the largest gain from visual articulation at this SNR. We contended that the multisensory speech system is specially tuned for SNRs between extremes, extremes where the system relies on either the visual (speechreading) or the auditory modality alone, forming a window of maximal integration centered at intermediate SNRs.

This specially tuned integration window is very likely determined in part by properties of the speech stimulus. In spoken words, vowels are generally easier to identify whereas consonants, due to their lower intensity, are easier to mask with noise (Barnett, 1999; French and Steinberg, 1947). Of course, consonant sounds play a critical role in speech recognition, since many words, especially monosyllabic words, share the same vowel structure (e.g. game, shame, blame, tame, name). At lower SNR’s, often only vowels are intelligible and words can therefore remain ambiguous. It is possible that in our particular word identification task, critical consonant information became available to the listener at –12dB, which together with visual cues facilitated the identification of the whole word. That is, we posit that visual articulation becomes maximally effective when accompanied by a certain, critical amount of acoustic consonant information. At this point on the SNR function, the gain increases until word recognition in the AV condition becomes increasingly restricted by the performance ceiling (100%).

From then on, benefit decreases again, thereby bracketing what we have termed the window of maximal integration (Ross et al., 2007).

Development of this specially tuned window is likely a factor of repeated lifetime exposure to environmental conditions that are approximated by the -12dB condition in our study. In an ecological context we are often unable to eliminate the source of noise. We have, however, a variety of options to adjust or compensate: we can get closer to the sound source (i.e. move closer to the speaker), adjust the volume (ask the speaker to speak up) etc. Therefore, conditions where the acoustic stimulus is masked to the extent that we are forced to rely solely on speechreading are rather rare. Thus, the exposure to intermediate SNR's throughout development may have resulted in an adaptation of multisensory mechanisms in the brain to integrate at these SNR's more efficiently. Through repeated exposure to these specific environmental conditions, cells in multisensory regions integrating auditory and visual speech may have the capacity to be "tuned" or "sensitized" to audiovisual speech that is delivered at intermediate SNR's. In this view the integration window is therefore an emerging property of a plastic multisensory system. This notion is supported by evidence from electrophysiological experiments in the superior colliculus (SC) of cats. Wallace and Stein have shown that multisensory neurons are present in the SC of newborn cats (Wallace and Stein, 1997), but that these neurons show no capacity yet to integrate sensory inputs. These neurons, however, acquire the capacity to integrate information from different senses within the first few weeks after birth (Wallace and Stein, 1997). Critically, these integrative properties are

gated by feedback inputs from neocortex (Jiang et al., 2001; Wallace and Stein, 1997) and are dependent on the animal's specific sensory environment (Wallace, 2004). In a recent study by Wallace and Stein, the authors were able to show that when cats were raised in an altered sensory environment where auditory and visual inputs were temporally coupled but originated from different spatial locations multisensory cells in the SC developed adaptively by integrating auditory and visual input originating from different spatial locations (Wallace, 2006).

It also follows that the development of appropriately functioning multisensory networks is reliant on the integrity of the individual sensory systems themselves and is therefore vulnerable to abnormalities in basic sensory processing, as has been shown in schizophrenia. Consequently, we would predict differences in the attributes of this integration window where sensory processing is impacted by a disorder early in life, a notion that may well extend to other childhood clinical conditions such as autism (e.g. (Iarocci and McDonald, 2006b; Molholm and Foxe, 2005) and early hearing impairments (e.g. Schlumberger et al., 2004).

5.3 Why No Unisensory Auditory Deficits in Speech Recognition?

A somewhat surprising finding in our study was that patients with schizophrenia did not show any deficits in the condition where the words were presented without the aid of visual articulation. This is surprising, given the large number of reports of impairment in speech and language related functions (e.g. Baltaxe and

Simmons, 1995; Bull and Venables, 1974; Cannon et al., 2002; Condray, 2005; Lee et al., 2004; Leitman et al., 2005; Titone and Levy, 2004). It also appears reasonable to assume that lower level auditory deficits that have been shown in schizophrenia (e.g. Javitt et al., 1997, 2000a; Michie, 2001; Rosburg et al., 2004) would impact higher- order functions such as speech perception. Overall, however, receptive language does not seem to be as impaired as speech production (Weinstein, 2006). Thought disorder, the most prominent symptom of schizophrenia, observably manifests itself in language production. Functional brain imaging studies have found thought disorder to be associated with altered activation patterns in the left and right STS/G (McGuire, 1998a; McGuire, 1998b; Kircher, 2002) during speech production. It has been hypothesized that the relative preservation of receptive language in schizophrenia is due to a compensatory process (Weinstein, 2006) analogous to the mechanism proposed for the preservation of performance in tasks targeting working memory (Manoach et al., 1999; Manoach, 2003). Here, normal performance in low memory load conditions was associated with an increase in prefrontal activity whereas high demands produced low performance together with decreased activity in the prefrontal cortex. This particular pattern of performance and activation in schizophrenia is thought to be due to a memory system operating at higher intensity to maintain normal performance. This way, patients with schizophrenia are able to compensate for deficits in less demanding memory tasks whereas patients reach the limit of their capacity when higher demands are imposed. It is possible that language production represents a higher demand on the system

than receptive processes such as word recognition. Consequently, one might predict that more complex speech recognition may also reveal a receptive deficit in schizophrenia. For example, meaningful and syntactically correct sentences provide semantic and syntactic context that benefit word recognition. There is evidence that patients with schizophrenia do not experience that benefit to the same extent as controls (e.g. Kuperberg et al., 2006; Ruchsow et al., 2003).

In line with the found lack of impairment in the recognition of spoken speech in schizophrenia are studies that failed to find strong associations between psychoacoustic measures of auditory acuity and auditory speech recognition (CHABA Working Group 95, 1991). It has been argued elsewhere (Watson et al., 1996) that given the redundancy of the speech signal and the complexity of speech recognition as a cognitive task, strengths in auditory processing abilities may compensate for existing impairments. This argument is supported by evidence showing that the elimination of fine spectral detail has little impact on speech intelligibility (Dudley, 1939; Greenberg and Arai, 2004).

6. Conclusions

In conclusion, patients with schizophrenia showed deficits in their ability to derive benefit from visual articulatory motion while unisensory auditory speech perception remained fully intact. It is possible that dysfunction in audio-visual speech integration is related to a well-characterized dysfunction of the dorsal visual processing stream but this remains to be explicitly examined.

Table 1.

Demographic and clinical characteristics of schizophrenia patients.

Table 1

Variable	Mean	SD	Range
Patient Socioeconomic Status[1]	22	9.6	8-45
Parental Socioeconomic Status	40	22.6	8-99
Brief Psychiatric Rating Scale Score	40	20	28-55
Quick IQ[2]	98	11	73-116
Diagnostic Subtype	N		
Schizophrenia, no subtype	1		
Undifferentiated	7		
Paranoid	7		
Schizoaffective	3		
Antipsychotic Medications			
Typical	2		
Atypical[3]	22		
Chlorpromazine equivalent	1196	435	500-2000

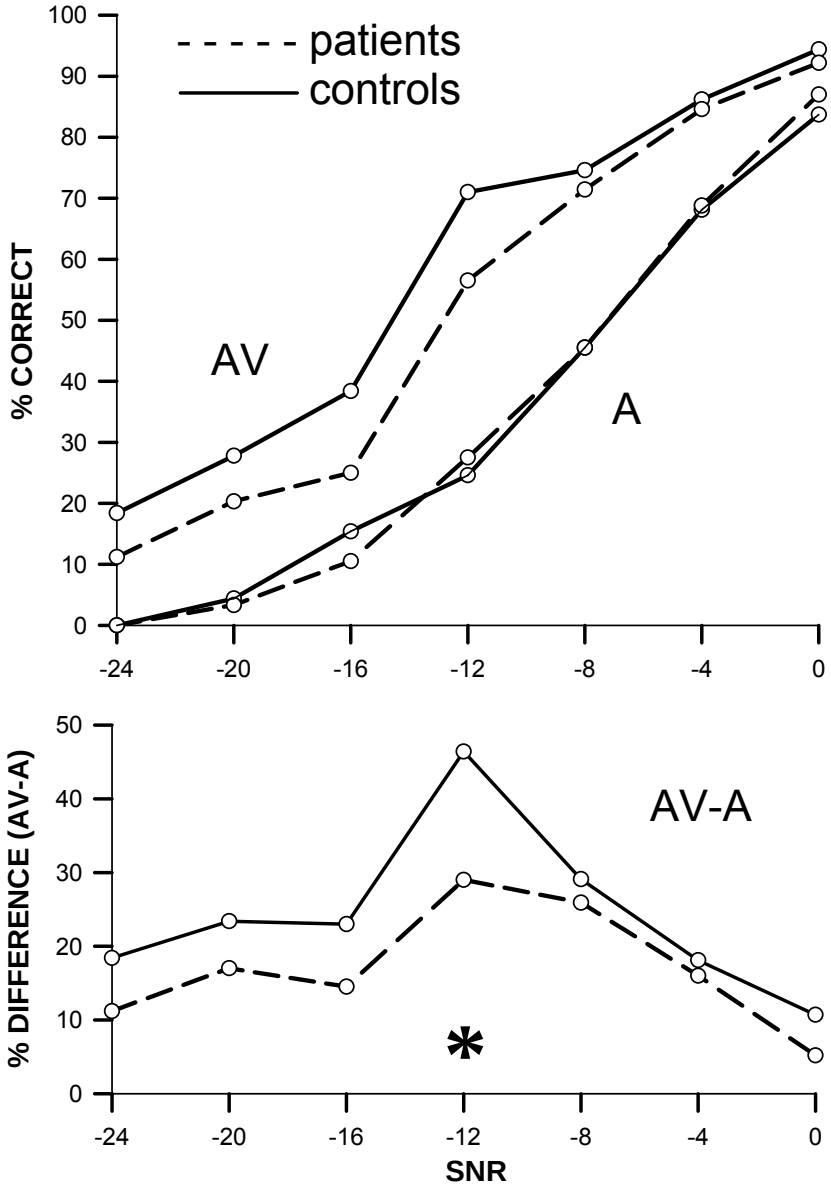
[1]Smaller numbers reflect higher socioeconomic status, per the Hollingshead scale
[2]Truncated version of the Peabody Picture Vocabulary (non-verbal)
[3]Note that 7 patients received 2 different antipsychotic medications

Figure 1.

Top panel: Percentage of correctly identified words (%correct) depending on the signal to noise ratio (SNR) for the auditory alone (A) and the audiovisual conditions. The dashed lines represent the performance of the patients with schizophrenia and the solid lines the performance of healthy controls. Bottom panel: Multisensory gain (AV-A) in speech- recognition accuracy as a function of SNR. Here, the significant difference between groups is indexed with an asterisk.

Figure 1

PATIENTS WITH SCHIZOPHRENIA (N=18) AND CONTROLS (N=18)



References

Alain, C., Bernstein, L. J., Cortese, F., Yu, H., & Zipursky, R. B. (2002). Deficits in automatically detecting changes in conjunction of auditory features in patients ith schizophrenia. *Psychophysiology, 39,* 599-606.

Allison, T. P. A., & McCarthy, G. (2000) Social perception from visual cues: Role of the STS region. *Trends Cogn Sci, 4,* 267-278.

Baltaxe, C. A., & Simmons, J. Q., 3rd (1995). Speech and language disorders in children and adolescents with schizophrenia. *Schizophr Bull, 21,* 677-692.

Barnett, H. (1999). Overview of speech intelligibility. *Proc IOA, 21,* 1-15.
Bernstein, L. J., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? Cambridge, MA Bradford: MIT Press.

Buchan, J. N., Pare, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2,* 1-13.

Bull, H. C., & Venables, P. H. (1974). Speech perception in schizophrenia. *Br J Psychiatry, 125,* 350-354.

Butler, P. D., Hoptman, M. J., Nierenberg, J., Foxe, J. J., Javitt, D. C., & Lim, K. O. (2006). Visual white matter integrity in schizophrenia. *American Journal of Psychiatry,163,* 2011-2013.

Butler, P. D., Martinez, A., Foxe, J. J., Kim, D., Silipo, G., Mahoney, J., Shpaner, M., Jalbrikowski, M., & Javitt, D. C. (2007). Subcortical visual dysfunction in schizophrenia drives secondary cortical impairments. *Brain, 130,* 417-430.

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

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

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *J Cogn Neurosci, 15,* 57-70.

Campbell, R., & MacSweeney, M. (2004). Neuroimaging studies of cross-modal plasticity and language processing in deaf people. Cambridge, MA Bradford: MIT Press.

Cannon, M, Caspi, A., & Moffitt, T. E. (2002). Evidence for early-childhood, pan-

developmental impairment specific to schizophreniform disorder: Results from a longitudinal birth cohort. *Arch Gen Psychiatry, 59,* 449-456.

CHABA, Working Group on Communication Aids for the Hearing- Impaired. (1991). Speech perception aids forth the hearing impaired people: Current status and needed research. *J Acoust Soc Am, 90,* 637-685.

Chen, Y., Bidwell, L. C., & Holzman, P. S. (2005). Visual motion integration in schizophrenia patients, their first-degree relatives, and patients with bipolar disorder. *Schizophr Res, 74,* 271-21.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear Hear, 23,* 439-449.

Condray, R. (2005). Language disorder in schizophrenia as a developmental learning disorder. *Schizophr Res, 73,* 5-20.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Percept Psychophys, 51,* 355-362.

De Gelder, B., Vroomen, J., Annen, L., Masthof, E., & Hodiamont, P. (2003). Audio-visual integration in schizophrenia. *Schizophr Res, 59,* 211-218.

Doniger, G. M., Foxe, J. J., Murray, M. M., Higgins, B. A., & Javitt, D. C. (2002). Impaired visual object recognition and dorsal/ ventral stream interaction in schizophrenia. *Arch Gen Psychiatry, 59,* 1011-1020.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hearing Res, 12,* 423-425.

First, M. B., Spitzer, R. L., Benjamin, L., Gibbon, M., & Williams, J. B. W. (1997). Structured Clinical Interview for DSM-IV: American Psychiatric Publishers INC.

Foxe, J. J., Doniger, G. M., & Javitt, D. C. (2001). Early visual processing deficits in schizophrenia: Impaired P1 generation revealed by high-density electrical mapping. *NeuroReport, 12,* 3815-3820.

Foxe, J. J., Murray, M. M., & Javitt, D. C. (2005). Filling-in in schizophrenia: A high-density electrical mapping and source-analysis investigation of illusory contour processing. *Cereb Cortex, 15,* 1914-1927.

Foxe, J. J., & Schroeder, C. E. (2005). The case for a feedforward component in multisensory integration mechanisms. *NeuroReport, 16,* 419-423.

French, N. R., Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The J Acoust Soc Am, 19,* 90-119.

Gagne, J. P., Querengesser, C., Folkeard, P., Munhall, K., & Masterson, V. M. (1995). Auditory, visual, and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech. *The Volta Review, 97,* 33-51.

Grant, K., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am, 108,* 1197-1208.

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Trans Inf & Syst, E87,* 1059-1070.

Hoffman, R. E., Rapaport, J., Mazure, C. M., & Quinlan, D. M. (1999). Selective speech perception alterations in schizophrenic patients reporting hallucinated "voices." *Am J Psychiatry, 156, 393*-399.

Hyman, S. E., Arana, G. W., & Rosenbaum, J. F. (1995). *Handbook of psychiatric drug therapy*. Boston, MA: Little, Brown and Company.

Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M. C., Mazziotta, J. C., & Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proc Natl Acad Sci U S A, 98,*13995-13999.

Iarocci, G., & McDonald, J. (2006a). Sensory integration and the perceptual experience of persons with autism. *J Autism Dev Disord, 36,* 77-90.

Javitt, D. C., Doneshka, P., Grochowski, S., & Ritter, W. (1995). Impaired mismatch negativity generation reflects widespread dysfunction of working memory in schizophrenia. *Arch Gen Psychiatry, 52,* 550-558.

Javitt, D. C., Doneshka, P., Zylberman, I., Ritter, W., & Vaughan, Jr., H. G. (1993). Impairment of early cortical processing in schizophrenia: An event-related potential confirmation study. *Biol Psychiatry, 33,* 513-519.

Javitt, D. C., Grochowski, S., Shelley, A. M., & Ritter, W. (1998). Impaired mismatch negativity (MMN) generation in schizophrenia as a function of stimulus deviance, probability, and interstimulus/ interdeviant interval. *Electroencephalogr Clin Neurophysiol, 108,* 143-153.

Javitt, D. C., Shelley, A., & Ritter, W. (2000a). Associated deficits in mismatch negativity generation and tone matching in schizophrenia. *Clin Neurophysiol, 111,* 1733-1737.

Javitt, D. C., Shelley, A. M., Silipo, G., & Lieberman, J. A. (2000b). Deficits in auditory and visual context-dependent processing in schizophrenia: Defining the pattern. *Arch Gen Psychiatry, 57,* 1131-1137.

Javitt, D. C., Strous, R. D., Grochowski, S., Ritter, W., & Cowan, N. (1997). Impaired precision, but normal retention, of auditory sensory ("echoic") memory information in schizophrenia. *J Abnorm Psychol, 106,* 315-324.

Jiang, W., Wallace, M., Jiang, H., Vaughan, J., & Stein, B. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. *J Neurophysiol, 85,* 506-2.

Jibson, M. D., & Tendon, R. (1998). New atypical antipsychotic medications. *L Psychiatr Res, 32,* 215-228.

Kern, J. K. (2002). The possible role of the cerebellum in autism/ PDD: Disruption of a multisensory feedback loop. *Med Hypotheses, 59,* 255-260.

Kim, D., Wylie, G., Pasternak, R., Butler, P. D., & Javitt, D. C. (2006). Magnocellular contributions to impaired motion processing in schizophrenia. *Schizophr Res, 82,* 1-8.

Kim, J., Doop, M. L., Blake, R., & Park, S. (2005). Impaired visual recognition of biological motion in schizophrenia. *Schizophr Res, 77,* 299-307.

Kircher, T., Liddle, P. F., Brammer, M. J., Williams, S. C., Murray, R. M., & McGuire, P. K. (2002). Reversed lateralization of temporal activation during speech production in thought disordered patients with schizophrenia. *Psychol Med, 32,* 439-449.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperberg, G. R., Sitnikova, T., Goff, D., & Holcomb, P. J. (2006). Making sense of sentences in schizophrenia: Electrophysiological evidence for abnormal interactions between semantic and syntactic processing. *J Abnorm Psychol, 115,* 251-265.

Lebib, R., Papo, D., De Bode, S., & Baudonniere, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci Lett, 341,* 185-188.

Lee, S. H., Chung, Y. C., Kim, Y. K., & Suh, K. Y. (2004). Abnormal speech perception in schizophrenia with auditory hallucinations. *Acta Neuropsychiatrica, 16,* 154-159.

Leitman, D. I., Foxe, J. J., Butler, P. D., Saperstein, A., Revheim, N., & Javitt, D. C. (2005). Sensory contributions to impaired prosodic processing in schizophrenia. *Biol Psychiatry, 58,* 56-61.

Manoach, D. S. (2003). Prefrontal cortex dysfunction during working memory performance in schizophrenia: Reconciling discrepant findings. *Schizophr Res, 60,* 285-
298.

Manoach, D. S., Press, D. Z., & Thangaraj, V. (1999). Schizophrenic subjects activate dorsolateral prefrontal cortex during a working memory task, as measured by fMRI. *Biol Psychiatry, 45,* 1128-1137.

Massaro, D. W. (1998). *Perceiving talking faces: Insights into auditory attention*. Cambridge, MA: MIT Press.

McGuire, P., Quested, D. J., Spence, S. A., Murray, R. M., Frith, C. D., & Liddle, P. F. (1998). Pathophysiology of "positive" thought disorder in schizophrenia. *Br J Psychiatry,173,* 231-235.

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

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol, 56,*640-662.

Michie, P. T. (2001). What has MMN revealed about the auditory system in schizophrenia? *Int J Psychophysiol, 42,* 177-194.

Michie, P. T., Innes-Brown, H., Todd, J., & Jablensky, A. V. (2002). Duration mismatch negativity in biological relatives of patients with schizophrenia spectrum disorders. *Biol Psychiatry, 52,* 749-758.

Molholm, S., & Foxe, J. J. (2005). Look "hear", primary auditory cortex is active during lip-reading. *NeuroReport, 16,* 123-124.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cereb Cortex, 14,* 452-465.

Molholm, S., Sehatpour, P., & Mehta, A. D. (2006). Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. *J Neurophysiol, 96,* 721-729.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E.

(2004a). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol Sci, 15,* 133-137.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004b). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol Sci,15,* 133-137.

Munhall, K. G., Servos, P., Santi, A, & Goodale, M. A. (2002). Dynamic visual speech perception in a patient with visual form agnosia. *NeuroReport, 13,* 1793-1796.

Myslobodsky, M. S., Goldberg, T., Johnson, F., Hicks, L., & Weinberger, D. R. (1992). Lipreading in patients with schizophrenia. *J Nerv Ment Dis, 180,* 168-171.

O'Neill, J. J. (1954). Contributions of the visual component of oral symbols to speech comprehension. *J Speech and Hearing Dis, 19,* 429.

Pare, M., Richler, R. C., Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. Percept Psychophys, 65, 553-567.

Peuskens, J., & Link, C. G. G. (1997). A comparison of quetiapine and chlorpromazine in the treatment of schizophrenia. *Acta Psychiatr Scand, 96,* 265-273.

Rosburg, T., Kreitschmann-Andermahr, I., & Sauer, H. (2004). Mismatch negativity in schizophrenia research: An indicator of early processing disorders of acoustic information. *Nervenarzt, 75,* 633-641.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex, 17,* 1147-1153.

Ruchsow, M., Trippel, N., Groen, G., Spitzer, M., & Kiefer, M. (2003). Semantic and syntactic processes during sentence comprehension in patients with schizophrenia:Evidence from event-related potentials. *Schizophr Res, 64,* 147-156.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2006). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia.*

Salisbury, D. F., Shenton, M. E., Griggs, C. B., Bonner-Jackson, A., McCarley, R. W. (2002). Mismatch negativity in chronic schizophrenia and first-episode schizophrenia. *Arch Gen Psychiatry, 59,* 686-694.

Schlumberger, E., Narbona, J., & Manrique, M. (2004). Non-verbal development of children with deafness with and without cochlear implants. *Dev Med Child Neurol,46,* 599-606.

Schonauer, K., Achtergarde, D., & Reker, T. (1998). Lipreading in prelingually deaf and hearing patients with schizophrenia. *J Nerv Ment Dis, 186,* 247-249.

Schroeder, C. E., & Foxe, J. J. (2005). Multisensory contributions to low-level, "Unisensory" processing. *Current Opinions in Neurobiology, 15,* 454-458.

Schwartz, B. D., Tomlin, H. R., Evans, W. J., & Ross, K. V. (2001). Neurophysiologic mechanisms of attention: A selective review of early information processing in schizophrenics. *Front Biosci, 6,* D120-134.

Shenton M. E., Dickey, C. C., Frumin, M., & McCarley, R. W. (2001). A review of MRI findings in schizophrenia. *Schizophr Res, 49,* 1-52.

Stein, B. E., Jiang, W., Wallace, M. T., & Stanford, T. R. (2001). Nonvisual influences on visual-information processing in the superior colliculus. *Prog Brain Res, 134,*143-156.

Stein, B. E., Wallace, M. W., Stanford, T. R., & Jiang, W. (2002). Cortex governs multisensory integration in the midbrain. *Neuroscientist, 8,* 306-314.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise.*The J Acoust Soc Am, 26,* 212-215.

Surguladze, S. A., Calvert, G. A., & Brammer, M. J. (2001). Audio-visual speech perception in schizophrenia: An fMRI study. *Psychiatry Res, 106,* 1-14.

Titone, D., & Levy, D. L. (2004). Lexical competition and spoken word identification in schizophrenia. *Schizophr Res, 68,* 75-85.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S. & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept Psychophys, 60,* 926-490.

Wallace, M., Perrault, Jr., T. J., Hairston, W. D., & Stein, B. E. (2004). Visual experience is necessary for the development of multisensory integration. *J Neurosci, 24,* 9580-9584.

Wallace, M., & Stein, B. E. (2006). Early experience determines how the senses will interact. *J Neurophysiol, 97*, 921-926.

Wallace, M. T., & Stein, B. E. (1997). Development of multisensory neurons and

multisensory integration in cat superior colliculus. *J Neurosci, 17,* 2429-2444.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *J Acoust. Soc Am, 100,* 1153-1162.

Weinstein, S., Werker, J. F., Vouloumanos, A., Woodward, T. S., & Ngan, E. T. (2006). Do you hear what I hear? Neural correlates of thought disorder during listening to speech in schizophrenia. *Schizophr Res, 86,* 130-137.

Woods, S. W. (2003). Chlorpromazine equivalent doses for the newer atypical antipsychotics. *J Clin Psychiatry, 64,* 663-667.

Yeap, S., Kelly, S. P., Sehatpour, P., Magno, E., Javitt, D. C., Garavan, H., Thakore, J. H., & Foxe, J. J. (2006). Early visual sensory deficits as endophenotypes for schizophrenia: High-density electrical mapping in clinically unaffected first-degree relatives. *Arch Gen Psychiatry, 63,* 1180-1188.

CHAPTER IV

MULTISENSORY INTEGRATION IN ILLUSORY MOTION PERCEPTION

1. Introduction

1.1 Background

The ability to perceive the motion of objects is a fundamental aspect of perception and a prerequisite for many forms of behavior. In order to perform in almost all everyday situations, organisms have to detect the trajectory and the speed of moving objects. Information about motion parameters can be transmitted through different sensory channels. Most obviously, the prominent modality to process motion is the visual system, but motion information can also be contained in changes in frequency and intensity of sounds and tactile stimulation. This multisensory aspect of motion provides more accurate information about the location and speed of a moving stimulus and supplemental information from one modality can disambiguate the perception of weak input from another (Sekuler, 1997). On the other hand, spatial disparity, or more generally, incongruency between dynamic auditory and visual events can have detrimental effects on perceptual judgments (see Soto-Faraco, 2004 for a review). An issue of current debate is the level of information processing at which visual and auditory dynamic information interact.

In a study by Meyer and Wuerger (2001), the authors investigated the potential influence of auditory motion on the discriminability of visual motion direction. Using a random dot kinematogram (RDK), participants were asked to discriminate the global direction of the dots at different coherence levels. The

distractor was a white noise sound source moving in a congruent or incongruent direction to the visual stimulus. In this study participants were biased towards the 'irrelevant' direction of the auditory noise in cases when the direction of the dots in the RDK was ambiguous. At higher coherence levels, no effect of the direction of the auditory sound on the direction of the visual stimulus was found. The authors interpreted these results as evidence for late decisional processes and that auditory directional information was unlikely to have affected visual motion perception at earlier processing stages. Visual and auditory information, the authors concluded, was to a large extent processed independently and the outputs of the two sensory modalities were combined following a simple probability summation rule (Wuerger, et al., 2003).

Later studies also supported the claim that audiovisual interactions in motion processing were confined to post-perceptual processes in the absence of sensory or perceptual effects (Alais and Burr, 2004; Wuerger et al., 2003). Contradictory evidence comes from a number of recent studies using motion adaptation aftereffects (Kitagawa and Ichihara, 2002, Vroomen and DeGelder, 2003) or a psychophysical staircase-method (Soto-Faraco et al., 2005) designed to eliminate possible confounds from later, decisional processes. Most recently Sanabria et al. (2007) investigated the relative contribution of perceptual and decisional processes using a signal detection paradigm. The authors were able to show a significant decrease in sensitivity to the direction of auditory targets in the presence of a moving visual stimulus, providing evidence for influence of dynamic visual information on the detection of auditory motion at a perceptual

level. Conversely, there were significant shifts in response criterion depending on the direction of the visual distractor motion. According to the authors, these results supported the view that there are both perceptual and decisional components involved in audiovisual interactions during motion processing which can coexist but are largely independent of one another.

Another aspect of these studies is that there seems to be an asymmetry regarding the direction of influence between visual and auditory motion. Often, visual motion influences auditory motion but not vice versa (e.g. Kitagawa and Ichiara, 2001; Soto-Faraco et al.; 2002). This is consistent with the so called modality appropriateness hypothesis (Howard and Templeton, 1966; Welch and Warren, 1986) which postulates that the modality that is most reliable with respect to a given task is the modality that dominates the perception in the context of that task. Vision, having higher spatial resolution would therefore dominate auditory perception in spatial tasks whereas audition, being more precise in the temporal domain, would dominate the visual percept in scenarios with a temporal emphasis. In the “illusory flash illusion” (Shams et al., 2000, 2002) sound radically changes the phenomenological quality of the percept of a nonambiguous visual stimulus. When a transient visual flash is accompanied by multiple auditory transients then the flash is perceived as multiple flashes. There are other studies that have shown an effect of static sounds on the perception of visual apparent motion (e.g. Gilbert, 1939; Hall and Earle, 1954; Hall et al., 1952; Maass, 1938; Zietz and Werner, 1927; Manabe and Riquimaroux, 2000),

although other studies have not always replicated these effects (Allen and Kohlers, 1981, Ohmura, 1987).

A more reliable example of the effect of sound on the perception of visual motion can be demonstrated by using the so-called bistable motion illusion. This illusion was first shown by the gestalt psychologist Meztger in 1937 and was more extensively investigated later by Michotte (1964). Here, two identical visual objects move toward each other, overlap, and then continue along their original motion paths. An observer may perceive the objects as either passing through each other (streaming or passing) or colliding and reversing their direction of motion (bouncing). In this ambiguous motion event, subjects are usually biased towards the percept of streaming (e.g. Metzger, 1934; Michotte, 1963). When a sharp-onset auditory transient is presented at the moment of the coincidence of the two objects, observers typically become strongly biased toward the perception of a collision and the impression of the two objects bouncing off one another and reversing their motion paths (Sekuler et al., 1997). In the study by Sekuler et al. (1997), participants were asked for their subjective report of their percept. This effect was recently replicated by Sanabria et al (2004) using more objective measures. Here, the scenario was disambiguated by presenting two disks of different color. Observers responded faster in the congruent conditions, when the objective repulsion was paired with a collision sound or when the disks streamed without the presentation of an auditory stimulus than in incongruent cases (streaming + sound; bouncing + no sound). This design, however does not rule out a decisional or cognitive bias. Also, it is reasonable to assume that these

biases play a stronger role when the visual display is truly ambiguous. It could therefore be questioned whether this design is really comparable to the original bi-stable motion illusion where observers are confronted with the ambiguity of the visual stimulus.

Taken together, it appears that a fairly extensive history of behavioral research supports the view that early perceptual, and late decisional and cognitive factors may be involved in the integration of dynamic audiovisual events. Surprisingly, the use of brain imaging technology like functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG) and high density electroencephalography (EEG) that are frequently used to unravel the spatial and temporal loci of integration effects have played only a small role in this endeavor and most of the research using these technologies has focused on the special topic of audiovisual integration in speech (Calvert, 2003, 2004; Campbell,1998).

An exception is a study by Bushara et al. (2003) using fMRI. Subjects were presented with the bi-stable motion illusion and participants were asked to indicate via button presses whether they perceived the two identical objects streaming or bouncing. A brief sound was presented at the moment of the overlap of the two objects. The authors contrasted blood oxygen level dependent (BOLD) responses on trials where streaming was perceived with trials that elicited the bounce percept. First, the authors identified regions that were activated during the presentation of sound and the movement of the visual objects, regardless of the perceived trial type. Within these regions they

postulated that areas in which bouncing (or colliding) evoked significantly higher response than pass trials are those that mediate multisensory interaction.

Their results showed that processing of simultaneously presented auditory and visual stimuli activated a distributed neural system of multisensory brain areas. The authors also found increased activations in the insula/ frontal operculum, dorsolateral and medial prefrontal cortex, posterior parietal cortex, posterior thalamus, superior colliculus and posterior cerebellar vermis in trials where a collision was perceived. This increased activity was associated with a relative decrease in the activity in both visual and auditory sensory specific areas in the occipital and temporal cortex. The authors interpreted their findings as evidence for a reciprocal and competitive interaction between multisensory and sensory specific brain regions and proposed that this interaction subserved the perceptual interpretation of simultaneous signals from multiple sensory modalities. The authors concluded that multisensory networks are sites for early sensory processing and work in parallel with sensory- specific areas, rather than representing a late, final station in a hierarchical processing stream from unisensory to multisensory networks.

The following section is in part redundant to the discussion of cortical MSI in the general introduction of this dissertation. I repeat certain issues here, however because they represent an important context for the rationale of this study. There are now numerous reports of MSI effects in lower-order sensory regions in the imaging literature (e.g. Calvert et al., 1999; Macaluso et al., 2000; King et al., 2001: Foxe et al., 2000; Foxe and Schroeder 2005). These findings

are consistent with the view that multisensory effects in early sensory regions represent feedback inputs from higher- order multisensory regions. Additional evidence from research in human event related potentials (Murray et al., 2002) and intracortical recordings in monkeys (Mehta et al., 2000; Bullier et al., 2001) support the view that feedback plays a critical role in multisensory integration.

The findings of feedback mechanisms in multisensory integration fit well with the traditional view of sensory processing (and information processing in general) being organized in a hierarchical fashion, from low-order early sensory processing to “late” multisensory and finally perceptual and cognitive (such as executive) processes. Early work suggested that multisensory regions are all located at higher-order stages of sensory processing and there were no known anatomical cross-connections between earlier stages of the sensory pathways (e.g. from V1/V2 to A1/A2), this model had long been dominating the scientific understanding on how sensory integration occurred in the brain. However, as mentioned in the general introduction, the evidence for feedback processes, does not preclude the existence of cross-talk between sensory modalities before higher- order multisensory regions are activated. In fact, studies using haemodynamic measures themselves are not suited to provide support for this model since the haemodynamic signal is an indirect measure of cell activity and reflects activation only many seconds after the underlying neural event occurs.

In this light, the conclusions of Bushara et al. about the relative timing of information processing between predominantly unisensory and multisensory sites are speculative since the fMRI method provides no measure of temporal

dynamics. In order to draw conclusions about the temporal architecture of the unisensory and multisensory processes associated with this multisensory effect, one needs to utilize a brain-imaging technology that provides high- resolution temporal information, like event- related potentials (ERP) or MEG. Indeed, as layed out in detail in the general introduction of this thesis, recent work in both humans and monkeys now strongly suggests that feedforward and lateral connections also contribute to multisensory integration in low level sensory processing and can evolve within the first 100ms of processing within early sensory cortices (Giard and Perronet, 1999 Molholm, et al., 2002; for reviews see Foxe and Schroeder, 2005; Schroeder and Foxe, 2005). These findings are supported by findings of direct projections from primary auditory regions to both V1 and V2 (Falchier, 2002; Rockland and Ojima, 2003).

1.2 Rationale

Behavioral and physiological research on (non-speech) audiovisual multisensory integration has used static, transient stimuli. Much of our experience consists of dynamic events and their qualities are frequently transmitted through multiple sensory channels. As Sekuler (1997) has shown using the bistable motion illusion, in dynamic events the information from different modalities is not processed independently. Here, sound disambiguates an ambiguous dynamic visual event and radically and changes the percept. Brain imaging has revealed a complex neural network that is thought to underlie the integration of sound and visual motion. A complete understanding of the underlying integrative operations

of the brain that give rise to this multisensory percept requires the investigation of the involvement of neural networks in time. This study was designed to provide this crucial temporal information and investigated the time-course of the integration of auditory and visual neural processes during dynamic visual events using high-density electrical mapping (128 scalp electrodes). We assessed the timeframe in which interactions between the auditory and the visual modality emerged as a means of inferring the processing stages at which these integrations are instantiated.

1.3 Possible outcomes

Based on previous studies on multisensory integration in general and the bistable motion illusion in particular, several outcomes are possible. Classical animal studies in MSI have shown that multisensory effects manifest themselves as vigorous increases in firing rates of MS neurons to spatially and temporally coincident (and attenuation when this precondition is not satisfied) stimulation from different modalities, at levels much exceeding that predicted by summing the unisensory responses [AV> (A+V)]. This enhancement has been found in multisensory and sensory- specific cortices (see Fort and Giard, 2004 and Calvert, 2001 for reviews). Parallels of multisensory integration at subcortical and cortical levels have been suggested to reflect a general operational mode of multisensory integration throughout the brain (Stein et al., 2004). Based on this we would expect multisensory effects to manifest themselves as significantly increased activity (as indexed by modified amplitudes of the ERP waveforms)

over scalp sites reflecting multisensory and sensory specific generator activity. If increased multisensory processes give rise to the percept of bouncing then MS enhancement should be larger in the condition where bouncing is perceived.

Alternatively, predictions would be different presuming an alternative explanation for the MS bistable motion effect proposed by Watanabe et al. (1998). In their series of experiments they were able to show that not only sounds but also other transient events like brief flashes or a brief tactile stimulus (Watanabe, 2001) could induce the perception of a collision. More importantly, the effectiveness of the transient to induce the illusion could be manipulated with attentional deployment to a second task designed to distract attention from the ongoing motion event. Watanabe et al. (1998) suggested that the effect of sound on the bistable motion illusion is due to an attentional shift away from the visual motion process caused by a transient during the coincidence of the visual objects. On the basis of this finding, one can speculate that attention to visual motion engages visual motion processing areas, probably MT/MT+. A brief sound may cause a temporary attentional disengagement from this process presumably toward the auditory stimulus. This process should result in attenuated processing in visual motion processing sites coupled with a relative enhancement of auditory processing in trials when bouncing is perceived. Support for this hypothesis comes from a study by Berman et al. (2002) where it was shown that auditory attentional demands can influence the perception of the motion aftereffect. This was associated with a decrease in activity in the human motion area MT+ as measured by fMRI.

Finally, and perhaps importantly the results of the fMRI study by Bushara et al. (2003), on the bistable motion illusion allow for the projection of yet another set of possible outcomes of our study. The authors hypothesized that multisensory processing is associated with an increase in activity in multisensory brain areas coupled with a decrease in sensory specific sites due to early, parallel competitive interaction. It is problematic to make predictions how and where we would see multisensory modulations due to increased activity in MSI sites in the ERP based on Bushara et al's findings because we don't precisely know how the neural generators in the MSI sites project to the scalp surface[11]. Possible candidates based on their findings are the posterior parietal cortex and the medial and dorsolateral prefrontal cortex. Since the projection sites for the primary auditory and visual cortex are known we can predict that if Bushara et al.'s hypothesis is correct, we would see an early decrease in activity reflected in an attenuation in the ERP waveforms over bilateral fronto- central and bilateral occipital scalp locations. Since the MSI processes are presumably enhanced when a collision or bouncing is perceived relative to the streaming percept, then this attenuation should be more pronounced when the perceptual impression is that of a collision of the two visual objects.

Conversely, early interactions may not necessarily be generated by a competitive interaction between multisensory and sensory specific sites but may

[11] The investigators identify the insula/frontal operculum, the dorsolateral medial and prefrontal cortex, posterior parietal cortex, posterior thalamus, superior colliculus and posterior cerebellar vermis as sites with enhanced activity in trials where subjects reported a collision percept

instead reflect interactions between sensory specific cortices in line with a feedforward convergence model (Foxe and Schroeder, 2005; Schroeder and Foxe, 2005). Under this view, the visual and the auditory processing streams may interact at early levels as shown in neurophysiological experiments using static stimuli (Giard and Peronnet, 1999; Molholm et al., 2002) and behavioral data using moving objects and sounds (Sanabria et al, 2004, 2007; Soto-Faraco et al., 2003, 2004). Further, multisensory interactions are possible in both directions: from the auditory to the visual modality and vice versa. Here, the transient sound may affect visual processing as observed by modulations of the visual evoked response over occipital cortices (<140ms). It is also possible that the moving visual stimulus affects the auditory evoked activity which could be observed in modulations of one or more of the components of the early auditory evoked potential (AEP) waveform such as P1, N1, P2, N2. It should be noted here that due to the "late" emergence of these components, a determination about whether the modulation occurred directly from the visual cortex or was mediated by higher order MS- regions is not within the scope of this experiment.

Interestingly, while evoked responses to multisensory stimuli tend to be larger in amplitude than the sum of the corresponding unisensory responses (superadditivity: AV>A+V) it has been reported that in some cases the auditory N1 is suppressed (Besle et al., 2004; van Wassenhove et al., 2005). It is possible, however, that this effect is due to a relative attentional shift from the auditory to the visual modality since the auditory N1 component is known to be modulated by attention (for a review see Näätänen and Teder, 1991).

As mentioned in the beginning, a critical aspect of the bistable motion illusion is that the presence of the sound radically affects the percept of the motion event in the absence of any real visual change. It was therefore interesting to see at what time this perceptual effect emerges and which cortical regions are implicated. Again, it is possible that these purely perceptual effects emerge late in the form of a convergence of unisensory information in multisensory cortices or as a feedback process from multisensory sites to primary sensory regions.

2. Materials and Methods

2.1 Subjects

Thirteen paid volunteers with no history of neurological disorders participated (7 females, mean age: 25.7, SD: 7.9; males: mean age: 32.1, SD: 4.6). All reported normal hearing and normal or corrected- to-normal vision. The experimental procedures were approved by the Institutional Review Board of the Nathan Kline Institute for Psychiatric Research, and written informed consent was provided by each participant.

2.2 Stimuli and conditions

Condition 1 (auditory alone: A)

A 1000 Hz pure tone (40 ms duration; 5 ms rise/ fall times; 75 dB SPL) was presented over headphones every 1.7 seconds. In all other conditions the visual objects also coincided every 1.7 seconds.

Condition 2 and 3 (visual alone: $V_{aligned}$, $V_{misaligned}$)

In these two conditions no sound was presented. Two identical red squares (0.8° visual angle, luminance: 85 cd/m^2) were presented against a dark background, originating from either side of a computer CRT monitor (17", Iiyama) at 6.12° visual angle distance from one another presented at 150 cm viewing distance. Both disks started traveling with constant speed of 3.6° of visual angle per second towards one another, coincided at 0.85 seconds and continued along their motion paths until they ended at the place on the screen where they originated. In the first of the visual conditions where no sound was played ($V_{aligned}$) the squares completely overlapped at the time of coincidence. In pilot work participants reported almost exclusively streaming. In the $V_{misaligned}$ condition, both squares were slightly misaligned by approximately 0.2° in the vertical dimension and squares were universally perceived as passing by one another (i.e. there was no perceptual ambiguity). Both conditions were control conditions introduced for the purpose of subtraction from the corresponding multisensory conditions.

Condition 4 ($AV_{aligned}$)

In this audiovisual condition subjects were presented with the $V_{aligned}$- stimulus as described above. The above-specified tone (from the A condition) was presented at precisely the moment when the two squares overlapped in the center of the monitor screen

Condition 5 ($AV_{misaligned}$)

In the second audiovisual condition the $V_{misaligned}$ stimulus was paired with the auditory stimulus at the moment when both squares overlapped in the center of the screen, except that at that moment, one square was vertically displaced by 0.2° visual angle.

Conditions 6 and 7 ($AV_{aligned/lowvol/bounce}$, $AV_{aligned/lowvol/stream}$)

These two audiovisual conditions were identical to the $AV_{aligned}$ condition with the exception that the loudness of the auditory stimulus was adjusted downwards on a subject-by-subject basis to the point of subjective equality where bouncing and streaming were perceived with equal probability. This was on average at approximately 45dB SPL. When the amplitude of the tone is reduced in this fashion, subjects reported less bounce percepts and were able to willfully adjust their percept according to instructions. That is, in a pilot study, we found that a point of subjective equality could be approximated for subjects such that when instructed to perceive bouncing, they could do so, and when instructed to perceive streaming, they could also do so. Thus, having established the tone loudness at which this occurred for each subject, they were then instructed during alternating blocks of stimuli to “make” the disks bounce off each other or to willfully generate a streaming percept. After each of these blocks, subjects were asked if they had successfully induced the instructed percept and the block was only included in further analysis if the desired percept was successfully

generated. The advantage of this pair of conditions is that the percept of bouncing or streaming was induced by the identical stimulus configuration, precluding any stimulus-based account for effects seen.

Condition 8 ($V_{reverse}$)

In this visual condition no sound was presented. The two disks were misaligned from one another by 0.9° visual angle and make a "real" motion reversal when they reach the center of the screen. Any other parameters are identical to the other visual conditions. This condition was introduced to have an assessment of "real" instead of illusory motion reversal. We reasoned that if we could index an ERP marker of "real" motion reversal, then we might well be able to find that marker expressed in the ERP of the illusory bouncing condition ($AV_{aligned}$) and not in the multisensory condition that induced a streaming percept ($AV_{misaligned}$). As we will further describe in the section on our analysis strategy, this condition served as a guide for the analysis and interpretation of differences of our main multisensory conditions.

2.3 Design and task

The experimental conditions described above are presented in blocked form in counterbalanced order. Subjects fixated on a central cross and were required to make a speeded button push upon the occurrence of a green square at the center of the screen occurring randomly in 10% of the cases. That is, at the

moment of central overlap of the leftward and rightward moving squares, the red squares became green for a single screen refresh (17ms). This task was presented in all conditions of the experiment and was introduced to maintain an equal level of alertness between conditions and to ensure central fixation on the disks. Target trials were excluded from the analysis. It is important to note that participants were not required to indicate their percept of either streaming or bouncing. This avoided any possible response-related confounds. The experimental session includes 3.5 hours of data collection and took approximately 6.5 hours overall with breaks. Conditions were delivered in varying orders (see Table 2).

2.4 Data acquisition

Continuous EEG was recorded in AC-coupled mode (Neuroscan Synamp I amplifier system) from 128 scalp electrodes (impedances < 5 kΩ), referenced to the nose, band-pass filtered from 0.05 to 100Hz, and digitized at 500 Hz. The continuous EEG was divided into epochs (-100ms pre- to 500ms post- stimulus onset) and baseline corrected over the full 600 ms. Trials with blinks and eye movements were rejected off-line on the basis of vertical and horizontal electro-occulogram (VEOG and HEOG). An artifact criterion of ± 60 μV was used at all other scalp sites to reject trials with excessive EMG or other noise transients.

EEG epochs were sorted according to the 7 stimulus conditions for the non-target trials (target trials, where the central squares turned green, were not analyzed), and averaged from each subject to compute the event-related

potential (ERP). The baseline was then redefined as the pre-stimulus portion of the epoch (-100 to 0 ms).

2.5 Data Analysis Strategy

2.5.1 General description of waveforms

The first step was to compute the grand averages over each stimulus condition over all 13 subjects. We produced a general description of the ERP's of our basic conditions to see whether they produced an event-related waveform with appropriate componentry, timing and scalp topography. For that, we selected a number of representative electrodes over the scalp and displayed the ERP's of the unisensory auditory, visual and multisensory conditions.

The next step was to compare the ERP's of our unisensory and multisensory conditions. We inspected the auditory evoked potentials at the electrode site with the maximum amplitude in the A-condition. We also compared them with the corresponding audiovisual evoked potential to have a preliminary view of the multisensory effect without performing any subtractions. Accordingly, we compared the visual-alone responses at lateral-occipital scalp sites to the audiovisual (AV) responses.

2.5.2 Multisensory effects

Here, we inspected the timing and topography of cortical auditory-visual interactions in the multisensory audiovisual ERP's in comparison to the elicited brain-electrical activity in the unisensory conditions. A typical approach to study

multisensory interactions is to compare the response to a bisensory stimulus (say an AV stimulus) with the sum of the component unisensory cues presented in isolation (i.e. A+V; Fort et al., 2002a,b; Foxe et al., 2000; Giard and Peronnet, 1999; Schroeger and Widmann, 1998; Raij et al., 2000 al). The rationale is that, at early stages of stimulus analysis, the neural activities induced by the audiovisual stimulus are equal to the sum of activities induced separately by the component A and V stimuli plus the putative activities specifically elicited by the bisensory stimulation, or in other words, the multisensory interactions (Fort and Giard, 2004).

ERP (AV) = ERP (A) + ERP (V) + ERP (A x V interaction)

This comparison relies on the law of superposition of electric fields and applies whatever the number, configuration or geometry of the intracerebral generators are. The difference between the sum of the unisensory responses and the AV-response contains, according to this model, any multisensory effects.

The validity of this method, however, relies on a number of conditions. First, any activity that is common to all conditions, unisensory and multisensory, may confound the comparison since this activation is contained twice in the A + V waveform but only once in the AV waveform (see Teder-Salejaervi et al., 2002). A computation of the difference waveform [AV – (A+V)] in order to arrive at the multisensory effect would therefore subtract this activity twice. In our experiment, task-related trials were excluded from the analysis and this therefore did not

represent a possible confound. There is the possibility, however, that the task evoked a continuous tonic (rather than phasic) increase in activation. Subadditive effects in the multisensory condition may therefore be contaminated by this nonspecific activity. A possible confounding influence of tonic activation could not be completely ruled out and our results will be discussed in light of this possibility.

A critical consideration is that the audiovisual waveforms will not only reflect activation from the auditory stimulus and any multisensory effects, but also from the visual stimulus, despite the fact that we compare the waveforms over central scalp regions. The head is a conductive medium and current from the generators in visual cortex (or any other generators) will volume conduct widely across the scalp. In order to compare the auditory ERP with the auditory ERP that contains multisensory effects without any visual contributions, one first needs to subtract the visual activation from the audiovisual ERP (AV-V vs. A). Accordingly, we subtracted the auditory ERP's from the audiovisual ERP's when assessing multisensory effects over visual cortices (AV-A vs. V).

Since there is no precedence in the literature for ERP recordings to moving visual stimuli of this nature, we had no way to predict what the morphology of the visual waveforms would be. We therefore did not generate any predictions about the multisensory modulation of particular componentry of the visual ERP.

The following approach was taken to constrain the analyses performed. We collapsed the data for the $V_{aligned}$ and the $V_{misaligned}$ conditions into a single

average waveform. On the basis of our hypothesis that visual activity is modulated by auditory input, we investigated the componentry of the ERPs over occipital scalp sites. The grand-average waveforms from these locations were visually inspected, and their componentry broadly defined. Defining the componentry in this way—for exploratory studies—has the advantage that the overall waveform can be used to constrain the number of analyses undertaken, without reference to the dependent variables. Further, while the use of broadly defined component peaks is a good means of limiting the number of statistical tests that are conducted, it is important to bear in mind that these components clearly represent the activity of many simultaneously active brain generators at any given moment (e.g. Foxe and Simpson, 2002). As such, effects may not necessarily be coincident with a given component peak, especially in the scenario that only certain brain generators are effected by a given experimental condition. Thus, limiting the analysis to a set of discrete component peaks represents a very conservative approach to the analysis of high-density ERP data and raises the likelihood of missed effects (so-called Type II errors). Therefore, a second exploratory analysis phase was also undertaken with this dataset (described below).

2.5.3 Statistical cluster plots

To provide a more general description of the spatio-temporal properties of putative multisensory effects over visual and auditory cortices, we used the statistical cluster plot method (see Molholm et al., 2002). These are made up of

pointwise paired comparisons between conditions for all electrodes at each data point at our digitization rate of 500 Hz. Each individual p-value that falls below an alpha-level of 0.05% and is preceded by at least 10 data points that also fall below this value will be color-coded and reported on a two-dimensional map representing time and electrode location on the scalp. This method is regarded as being sufficiently conservative in avoiding Type I errors without loosing statistical power to an extent that we would risk Type II errors (e.g. Murray et al., 2002).

It is important to note, however, that it is ill-advised to use this method as a hypothesis generating tool for the same study or for "data- snooping" in the absence of an anticipated effect. We believe that this method is particularly suitable here because unlike the auditory components in our audiovisual ERP's and auditory ERP's we didn't anticipate our visual evoked response to have the typical well-defined componentry of a VEP to a static transient or an ERP to the onset of motion. That is, since there is no precedence in the literature for ERP recordings to moving visual stimuli of this nature, we had no way to predict what the morphology of the visual waveforms would be. We therefore did not generate any predictions about the multisensory modulation of a particular componentry of the visual ERP. We were interested, rather, in finding out in which timeframe these interactions occur and whether they include modulations of visual and/or auditory processes. In this sense, the question that we ask in this study requires a more exploratory analysis technique. Additionally, this type of paradigm has

been shown to involve a fairly extensive and distributed network of cortical regions (Bushara, 2003).

In order to create a comprehensive description of the timing and the scalp distribution of the multisensory effects we created a cluster plot by comparing the respective audiovisual conditions ($AV_{aligned}$, $AV_{misaligned}$) with the sum of their unisensory counterparts ($A_{alone}+V_{aligned}$, $A_{alone}+V_{misaligned}$) with running pairwise t-tests. Further, we compared both plots to examine whether the condition that created the bouncing illusion (aligned) displayed a different pattern of multisensory effects than the audiovisual condition where the squares were perceived as passing through one another (streaming).

2.5.4 Differences in multisensory integration between bouncing and streaming percepts.

In this part of our analysis we investigated whether the bouncing and streaming conditions differed in the spatio-temporal pattern of multisensory effects they evoked.

We employed several strategies to evaluate this effect. The first strategy was to directly compare the $AV_{aligned}$ and the $AV_{misaligned}$ conditions ($AV_{aligned}$ vs $AV_{misaligned}$). In order to assess this effect, we had to ensure that physical differences between conditions were negligible. In order to ensure that responses to the visual stimulation between these conditions is similar we first compared both V- conditions ($V_{alinged}$ vs $V_{misaligned}$). For this exploratory comparison we used a statistical cluster plot.

As part of the inspection of differences between our two main multisensory conditions we used the $V_{reverse}$ condition as a guide. We first compared the ERP that this condition evoked with the $V_{misaligned}$ condition. We reasoned that this comparison may reveal componentry in the visual evoked potential that was associated with “real” instead of illusory motion reversal. We further reasoned that these components may be present in the multisensory condition that produced the illusion of a reversal ($AV_{aligned}$) but not in the condition where the objects were perceived as streaming ($AV_{misaligned}$).

Our second step was to see whether the multisensory effects in the statistical cluster plots differed between the bounce and the streaming condition. If the statistical cluster plots showed differences in multisensory effects between the aligned condition where bouncing was perceived and the misaligned condition where streaming percepts dominated, then we deemed it reasonable to assume that at least part of the effects (probably the later ones) are a representation of perception of the event. The advantage of this method is that possible visual differences do not play a role here, because the aligned and misaligned conditions with their potentially different visual stimulation are not directly compared. If differences in effects are robust, then they should manifest themselves as a significant interaction between multisensory aligned and misaligned conditions. For that, we submitted the areas under the difference waveforms ($AV_{aligned}$-A_{alone}; $AV_{misaligned}$-A_{alone}; $AV_{aligned}$-$V_{aligned}$ and $AV_{misaligned}$-$V_{misaligned}$) to an RM-ANOVA. Area measures were computed for a chosen

latency window (see results section) for a selection of electrodes where the amplitude of the difference waveforms were maximal.

The final strategy was to compare the activity generated by the conditions where the intensity of the tone was individually titrated so that there was a subjective equality between bouncing and streaming percepts. Subjects were instructed to willfully evoke the percept of bouncing and streaming on alternating blocks of stimuli. In this pair of conditions the stimulation is exactly identical and does not evoke the illusory percept by itself. We therefore hypothesized that the multisensory effects that are evoked by the concurrent presentation of the tone and the moving squares are reduced in comparison to the illusory condition where the stimulus evoked the bouncing percept automatically. We further hypothesized that if we found differences between the elicited waveforms of the bounce and the streaming condition, they would be entirely due to the percept and possibly differences in cognitive strategies to evoke the percept since stimulation parameters are identical between conditions.

3. Results

3.1 Characterization of the unisensory conditions A_{alone}, $V_{aligned}$ and $V_{misaligned}$

3.1.1 The auditory evoked potential (A_{alone}) over the fronto-central cortex.

The purpose of this assessment of the A_{alone}-condition was to ensure that our stimulation parameters resulted in a brain response that had the componentry that is traditionally found when stimulating with an auditory transient. Indeed, the auditory evoked response in the A_{alone}-condition displayed the characteristic timing and topography of an auditory evoked potential (see Picton et al., 1974; Leavitt et al., 2007). As can be seen in Figure 2., the clearly defined N1-P2 component complex was maximal over fronto-central scalp sites with N1 peaking at approximately 100 ms (~ -6.0 μV) and P2 peaking at 182 ms (~ 6.0 μV). The earlier auditory P1 component is relatively smaller in amplitude (~1.0 μV) but clearly present, peaking at 56 ms. Over visual cortices (i.e. inferior posterior scalp regions) the auditory components (N1 and P2) were inverted and may be slightly more pronounced over the left hemisphere.

3.1.2 The unisensory visual evoked potentials ($V_{aligned}$, $V_{misaligned}$, $V_{reverse}$) over the occipital cortex.

As outlined in the methods section, there were three visual-alone conditions in this study- One where the two disks were slightly misaligned ($V_{misaligned}$), one where the disks were fully aligned ($V_{aligned}$), and one where the disks were misaligned and reversed their motion direction at the center of the screen ($V_{reverse}$).

The first two were control conditions, implemented for subtraction from the audiovisual ERP in the corresponding multisensory conditions. Their average waveform also served as a tool for the identification of components that may be

modulated in our multisensory conditions. $V_{reverse}$ served as a tool in our exploratory analysis of multisensory effects to identify possible components in the waveform that are associated with the perception of motion reversal. For that, we compared $V_{misaligned}$ and $V_{reverse}$ since the essential difference between both conditions is the reversal of motion direction in the $V_{reverse}$ condition.

The chosen onset for the visual ERP's at 0 ms is coincident with the complete overlap of both disks at the midpoint of the screen, precisely at the time of the onset of the auditory stimulus in the A_{alone} and the AV-conditions. Of course, at this timepoint, both disks were already in motion, converging towards the center from the left and right edges of the monitor, respectively. We therefore did not expect to find the distinct componentry that one would expect from a standard visual transient (i.e. the C1-P1-N1 component complex of the VEP) and the baseline period was also expected to contain ongoing active motion processing activity.

Figure 3. contains a general description of the evoked responses of the three visual-alone conditions over representative scalp locations. In the $V_{misaligned}$ condition (see Figure 3) we found a slow positive-going waveform with peak maxima over bilateral parieto-occipital sites at approximately 320 ms. The waveform in the $V_{aligned}$ condition also showed a slow-onset ramp that is negative-going over lateral occipital sites of both hemispheres and then shifted into the positive direction with a maximum, again, at around 320 ms. Responses to both motion stimuli were extremely similar, as one might have expected. However, they may have been different in the percept they evoked. While it was

virtually if not completely impossible to perceive the disks bouncing off of one another in the $V_{misaligned}$ condition, there were occasional if rare reports of bouncing percepts during the $V_{aligned}$ condition. In bistable motion displays like this participants perceive a bouncing motion in approximately 65% of the cases (Sekuler et al., 1997, Watanabe & Shimojo, 1998; Ross et al., 2004). A morphological similarity of the $V_{reverse}$ with the $V_{aligned}$ condition was particularly pronounced over right lateral-occipital and parieto-occipital sites. The $V_{reverse}$ condition showed an earlier onset of the ramp to the maximum peak (by approximately 50 ms).

3.1.2.1 Differences between $V_{aligned}$ and $V_{misaligned}$

We predicted that the very small alignment difference between the disks in these two visual-alone motion conditions would be very unlikely to produce any measurable difference in responses. This is essentially what was observed and the differences shown in the cluster plot (Figure 3.) are small and distributed in an unsystematic manner over the scalp in both space and time. No systematic pattern is evident across the entire data matrix. What differences are observed are seen to onset at around 50-60ms over occipital sites, but these only appear to occur in only two scalp sites. Beyond 250ms we find only small, isolated differences spreading further anterior until reaching frontal scalp locations at around 360 ms. Again, there is no obvious systematic pattern or none of the typical clusters that are observed when real physiological processing differences are present and we therefore considered these relatively isolated spots as

negligible. As such, we felt comfortable with the comparison of our primary multisensory conditions ($AV_{aligned}$ and $AV_{misaligned}$).

3.1.2.2 Identification of components in the visual evoked potential

As detailed in the section on our analysis strategy, we identified "components" in the ERP obtained to visual-alone motion stimuli as a means to constrain the number of comparisons between our multisensory conditions in an effort to control Type I errors.

Figure 5. shows the grand-average waveform of the $V_{aligned}$ and $V_{misaligned}$ conditions over a right occipito-parietal scalp location. At this scalp location the amplitude of the average visual ERP was maximal. Through visual inspection we identified four phases that we considered to have a "component-like" morphology. We selected a time window of +/-10 ms around the respective peaks for analysis. Table 3. lists peak and time windows of the selected components. We used these time-windows for the statistical comparison between our multisensory conditions.

3.1.2.3 Identification of components associated with motion reversal ($V_{reverse}$ vs. $V_{misaligned}$)

As mentioned in the methods section we used the $V_{reverse}$ condition to guide us in the selection of time windows for the comparison of both multisensory conditions (i.e. $AV_{aligned}$ vs. $AV_{misaligned}$). Figure 3. reveals that the waveform shows a distinct

componentry only over the lateral occipital and the parieto-occipital scalp over the right hemisphere. We therefore chose a scalp location over the right lateral occipital scalp to identify time windows for subsequent analysis. Figure 6. shows the $V_{reverse}$ and the $V_{misaligned}$ waveform and the difference waveform. The difference waveform shows several deflections. We chose time windows (+/- 10 ms) around the two prominent deflections for our analysis, which were located at 275 ms and 460 ms (see also Table 4.).

3.2 Comparison of the multisensory audiovisual evoked potentials with the unisensory evoked potentials

3.2.1 The audiovisual evoked potentials ($AV_{aligned}$, $AV_{misaligned}$) in reference to the auditory evoked potential (A_{alone}) over the fronto-central scalp.

Contrasting the A_{alone}- response to the AV condition allowed us to make an initial assessment of whether stimulation in two sensory modalities resulted in a modification of the ERP. We forbore the statistical analysis of this contrast at this point because the proper comparison, which involved the subtraction of the visual component ($V_{aligned}$) will follow in the next section.

3.2.1.1 Bounce condition: $AV_{aligned}$ vs. A_{alone}

Comparing the waveforms of the A_{alone} and the $AV_{aligned}$ conditions we found similar morphology of the auditory components (P1, N1 and P2) in both conditions (Figure 8., graphs A and B). Timing as well as scalp distribution were also similar with maxima over central and fronto-central sites at the same

latencies. Visual inspection revealed an apparent suppression/diminution of the N1 and the P2 amplitudes in the $AV_{aligned}$ condition over both hemispheres relative to the auditory alone condition.

3.2.1.2 Pass condition: $AV_{misaligned}$ vs. A_{alone}

As in the comparison above, we see a similar suppression of the N1 component in the multisensory condition, although somewhat smaller in magnitude (Figure 7., graphs C and D). The P2–suppression is absent in this condition.

3.2.2 The audiovisual evoked potentials ($AV_{aligned}$, $AV_{misaligned}$) in reference to the visual evoked potential (V) over the lateral occipital cortex.

3.2.2.1 $AV_{aligned}$ vs. $V_{aligned}$

Not surprisingly, the $AV_{aligned}$ waveform over the bilateral LOC (Figure 7., graphs E and F) displayed a similar componentry to the visual only condition ($V_{aligned}$). We found a negative slope over both hemispheres with a maximum at approximately 80 ms, followed by a rising ramp with a small bump at approximately 240 ms. The waveform continued to rise until it reached its maximum peak at about 320 ms.

The multisensory condition differed from the $V_{aligned}$ waveform in several respects. First, visual inspection revealed an early (starting at 0 ms) oscillatory activity in the 30 Hz frequency range. We also found this activity in the A_{alone} waveform. This is possibly entrained by the predictable nature of the auditory

stimulus. After three cycles the waveform showed a positive going ramp peaking, like the $V_{aligned}$ waveform, at 310 ms.

3.2.2.2 $AV_{misaligned}$ vs. $V_{misaligned}$

Overall, the $V_{misaligned}$ and the $AV_{misaligned}$ (Figure 7., graphs G and H) waveforms were morphologically similar to the $V_{aligned}$ and the $AV_{aligned}$ waveforms (graphs E and F). The components in the $V_{misaligned}$ waveform, however, were not as distinct as in $V_{aligned}$.

3.3. Multisensory effects

We assessed the spatio-temporal characteristics of the multisensory bistable motion illusion by comparing the AV-evoked responses to the evoked responses of both unisensory conditions, respectively. Following the wide-spread convention in multisensory research to characterize multisensory effects by comparing multisensory activity with the sum of the unisensory (here auditory and visual) activities, we first compared both multisensory conditions (AV) to the sum of the constituent unisensory conditions (A_{alone} + V) to provide a general overview of the multisensory effects. We then proceeded to examine the unbiased (see methods section) multisensory modulations of waveforms over central, fronto-central (i.e. the auditory cortical projection to the scalp), and occipital (i.e. the visual scalp projection) sites.

3.3.1 The comparison of the multisensory evoked responses (AV) with the sum of the constituent unisensory responses (A_{alone} + V).

Figure 9. shows response to the $AV_{aligned}$ condition and the sum of the unisensory evoked responses (A_{alone} + $V_{aligned}$) recorded from a set of representative electrode locations distributed over the scalp. It is clear in both comparisons that the multisensory responses in both conditions are subadditive in comparison to the sum of the unisensory responses. This pattern is present over scalp locations, representing auditory and visual generator activity. By visual inspection, the earliest differences over fronto-central locations were in the timeframe of the P1 component. An attenuation in the timeframe of the auditory N1 component was apparent in the bounce condition but appeared to be absent in the passing condition. In both comparisons, a continuous attenuation started in the timeframe of the P2 component and continued until the end of the epoch. Differences over lateral occipital and parieto-occipital scalp locations start before 100ms and continue as a long-lasting attenuation until the end of the epoch. The attenuation appears to be larger in the passing condition than on the bouncing condition.

3.3.2 Multisensory effects over the central scalp

3.3.2.1 Bounce condition: ($AV_{aligned}$ - $V_{aligned}$) vs. A_{alone}

Here, we were interested in assessing any multisensory effects in the ERP evoked by the audiovisual stimulus that produced a bouncing percept. We assessed these effects by subtracting the activation evoked by visual stimulus

($V_{aligned}$) from the audiovisual evoked potential ($AV_{aligned}$). By subtracting the visual contribution ($V_{aligned}$) to the multisensory evoked response (AV) we were left with the unbiased contribution from the auditory input and any possible multisensory effects. We then compared this derived response (i.e. $AV_{aligned}$ - $V_{aligned}$) with the waveforms of the A_{alone} condition over scalp locations where the amplitude of the auditory evoked potential was maximal.

The Graph B. in Figure 9. displays the ERP's evoked by the unisensory auditory (A_{alone}) and the unbiased multisensory condition ($AV_{aligned}$) together with the difference waveform. The inspection of the differences revealed a strong suppression of the audiovisual ERP in the timeframe of the N1 (~1.7 μV) and P2 (~3 μV). These effects were maximal over centro-parietal scalp locations which is somewhat posterior to scalp locations that commonly reflect N1 and P2 generator activity. This can be observed in the scalp voltage maps of the difference waveforms of the unbiased multisensory and auditory conditions displayed in Figure 10. Further, the two waveforms maintain a divergence for the entire epoch.

3.3.2.2 Pass condition: ($AV_{misaligned}$ -$V_{misaligned}$) vs. A_{alone}

Analogous to the analysis in the bounce condition, we computed the unbiased audiovisual ERP in the streaming condition by subtracting the visual contribution from the evoked response in the multisensory condition that produced a streaming percept ($AV_{misaligned}$ - $V_{misaligned}$).

The comparison of the audiovisual ERP with the unbiased audiovisual ERP revealed that the multisensory suppression effects in the N1 and the P2 timeframe that we found in the bounce condition were also present in the streaming condition. The maximal difference was ~1.7 μV for the N1 and ~2 μV for the P2 component over the centro-parietal scalp (Figure 9. graph A). Note that the continuous divergence of the two waveforms past 200 ms in the bouncing condition is absent in the passing condition.

Multisensory effects were analyzed statistically with two separate omnibus RM-ANOVA's for the areas under the curve of the N1 (80 ms – 100 ms) and P2 (166 ms – 186 ms) timeframe, respectively. Both 3x3 RM-ANOVA's were computed involving the factors stimulus type (A_{alone}, ($AV_{aligned}$ - $V_{aligned}$), ($AV_{misaligned}$ - $V_{misaligned}$) x electrode. The three electrodes were chosen at scalp locations where the amplitude of the auditory evoked potential was maximal. This was the case for central scalp locations. For the N1-timeframe factor stimulus type was significant with $F(2, 24) = 3.64$ ($p = 0.04$; eta-squared = 0.43). Planned T-tests were performed between the unisensory and both multisensory conditions, respectively. Differences in amplitude were only significant between A_{alone} and $AV_{aligned}$ with $t(12) = -2.99$ ($p = 0.01$). The T-test for differences between the unisensory auditory and the multisensory ($AV_{misaligned}$ - $V_{misaligned}$) conditions that evoked the passing percept showed no significant effect. We also tested the difference between ($AV_{aligned}$ - $V_{aligned}$) and ($AV_{misaligned}$ - $V_{misaligned}$) with an additional T-test which did not lead to a rejection of the Null- hypothesis.

For the P2 timeframe factor stimulus-type was significant with $F(2,24) = 4.9$ ($p = 0.02$; eta-squared = 0.29). We therefore followed up with planned comparisons (T- tests) between A_{alone} and ($AV_{aligned}$ - $V_{aligned}$) and ($AV_{misaligned}$ - $V_{misaligned}$), respectively. Again, the difference in amplitude between A_{alone} and ($AV_{aligned}$ - $V_{aligned}$) was significant with $t(12) = 2.94$ ($p=0.01$). This effect was somewhat weaker in the comparison between A_{alone} and $AV_{misaligned}$ and could therefore not be confirmed on a significant level $t(12) = 1.95$ ($p = 0.08$). Although this pattern suggested a difference between both multisensory conditions, the statistical analysis only revealed a trend suggesting a small effect that is most likely masked by the high variability in our data.

3.3.3 Multisensory effects over the occipital scalp

Here, we investigated multisensory effects over lateral occipital scalp locations. Analogous to our assessment of the effects over the central scalp, we subtracted the activation that was due to the auditory-alone stimulus from the multisensory evoked response (AV- A_{alone}) in order to make a comparison of the unisensory visual condition (V) that was unbiased by volume conduction caused by activation from auditory stimulation.

We chose representative electrodes over the lateral occipital scalp of both hemispheres where we previously identified components evoked by the average of both visual- alone stimuli (avg ($V_{aligned}$, $V_{misaligned}$)).

3.3.3.1 Bounce condition: ($AV_{aligned} – A_{alone}$) vs. $V_{aligned}$

Graphs C and D in Figure 9. show the elicited ERP's ($AV_{aligned}$ - A_{alone} vs. $V_{aligned}$) together with the difference waveform bilaterally over the lateral occipital cortex. Over both hemispheres the difference waveform revealed several peaks. This first was located at approximately 50 ms (~0.9 μV) over the left and right hemisphere and a second at about 150 ms to 160ms (~1.15 μV). A third positive divergence around 250ms was only seen over the left hemisphere. After 300ms both curve diverged again, this time $V_{aligned}$ having a more positive amplitude. This difference became most extensive at around 300ms over the left hemisphere and approximately 450 ms (~0.75 μV) over the right hemisphere. Figure 11. displays the voltage maps of the difference waveforms between the unbiased multisensory ($AV_{aligned}$ – A_{alone}) and the unisensory visual condition ($V_{aligned}$) at selected early time points (55 ms and 160 ms).

We tested differences between the unisensory and the multisensory conditions using the time windows that we had previously identified in the waveform of the visual alone conditions (avg($V_{aligned}$, $V_{misaligned}$) using a 2x2x3 RM-ANOVA with factors stimulus type (($AV_{aligned}$ -A_{alone}), $V_{aligned}$), hemisphere and electrode (3 electrodes over each hemisphere of the lateral occipital scalp).

The RM-ANOVA's for the first two time windows (60ms -70ms and 230ms – 250ms) yielded no significant effects. However, there was a trend for a main effect of factor stimulus type for the difference in the late time window (310ms – 330ms) with $F(1,12) = 4.41$; $p = 0.058$; eta squared = 0.27.

3.3.3.2 Pass condition: ($AV_{misaligned}$ -A_{alone}) vs. $V_{misaligned}$

Here, we compared the unbiased audiovisual condition producing a streaming percept with the corresponding unisensory visual condition ($AV_{misaligned}$ – A_{alone}) vs. $V_{misaligned}$). The waveforms over the bilateral lateral occipital cortex are displayed in Figure 9., graphs E and F. Over the right hemisphere the unisensory and the multisensory waveforms diverged early (around 40ms) and diplayed a maximum divergence at around 220 ms (~1.6 μV). Over the left hemisphere the difference waveforms showed a maximum divergence at around 220 ms (2.7 μV) and showed a sustained divergence until the end of the epoch.

As in the bounce condition, we followed our strategy to test differences between waveforms using components that we identified in our avg($V_{aligned}$; $V_{misaligned}$) waveform. None of the RM-ANOVA's for the three time windows, respectively, revealed any significant main effects or interactions.

3.3.4. The perception of the motion event: bouncing vs. streaming

3.3.4.1 Direct comparison of the multisensory conditions ($AV_{aligned}$ vs. $AV_{misaligned}$*)*

In looking for manifestations of the difference in multisensory interactions leading to the subjective impression of bouncing and streaming, our first strategy was to compare both multisensory conditions directly ($AV_{aligned}$ vs. $AV_{misaligned}$). We assumed the difference in visual stimulation between both cases to be minimal, since the misalignment was negligible. To ensure this, we compared the two visual control conditions ($V_{aligned}$ and $V_{misaligned}$) in a statistical cluster plot. As

described above, there were no systematic differences between both visual conditions. Figure 12. shows the waveforms of both conditions across the scalp.

Figure 12. reveals remarkable similarities between the evoked responses over frontal and fronto-central scalp locations. The $AV_{aligned}$ condition showed a small reduction in amplitude in the N1, and to a lesser degree in the P2 timeframe. Larger differences can be found over lateral-occipital scalp locations.

We tested differences between both conditions in the timeframes of the auditory N1 and P2 components. To also test for hemispheric differences, we chose electrode locations over "auditory" scalp projection sites of both hemispheres. The 2x2x3 RM-ANOVA on the areas under the curve for factors condition ($AV_{aligned}$, $AV_{misaligned}$), hemisphere (left, right) and electrode (3 electrodes over each hemisphere) revealed no significant main effects for the factor condition nor hemisphere, and no interactions.

For the analysis of effects over the occipital scalp, we used two timeframes that we had previously identified by inspection of the $V_{reverse}$ condition. We used the same electrode locations as in our analysis of multisensory effects over visual cortices. Again, the 2x2x3 RM-ANOVA for both timeframes respectively remained without significant main effects for factors condition, hemisphere, and there were no significant interactions.

Statistical cluster plot

As delineated in the section on our data analysis strategy, we employed statistical cluster plots for an exploratory analysis. This seemed especially

appropriate here in light of our failure to find multisensory effects over occipital cortices and between our main multisensory conditions. We employed a rather conservative data analysis strategy especially since we did not have any specific hypothesis about the location and timeframe of multisensory effects. Therefore, a more explorative data analysis strategy was warranted here since without further analysis, a Type II error was likely. As Figure 13. reveals, we failed to find any systematic differences between both multisensory conditions when we compared them directly with one another.

3.3.4.2 Comparison of the unbiased multisensory response (AV- A_{alone}) with the unisensory control condition ($V_{aligned}$, $V_{misaligned}$).

As described in the methods section, we employed another strategy to assess possible differences in multisensory effects between conditions where bouncing and streaming was perceived. For that we compared the unbiased multisensory response in both conditions ($AV_{aligned}$ - A_{alone}; $AV_{misaligned}$ - A_{alone}) with their unisensory visual control condition ($V_{aligned}$; $V_{misaligned}$) in a statistical cluster plot. These cluster plots show effects over the entire epoch and over all scalp locations for each of the conditions, respectively. We subsequently compared both cluster plots for differences in the timing and scalp distribution of multisensory effects. Figure 14. shows the comparisons of the multisensory with the unisensory conditions. The top plot represents multisensory effects in the bouncng condition ($AV_{aligned}$ - A_{alone} vs. $V_{aligned}$), whereas the bottom plot shows the streaming condition ($AV_{misaligned}$ – A_{alone} vs. $V_{misaligned}$).

Effects over central scalp regions

The cluster plots show remarkable differences between the bouncing and the streaming conditions. The N1 effect was small in the bouncing condition but absent in the streaming condition. The P2 effect was more pronounced in the bouncing condition than in the streaming condition, and the distribution of this effect was more anterior in the bouncing condition than in the streaming condition. This is most likely the reason why the RM-ANOVA in our analysis of multisensory effects over central scalp locations only revealed a trend in the ($AV_{misaligned}$ - $V_{misaligned}$) vs. A_{alone} comparison. Please note, however, that the pattern of effects correspond with the results of our statistical analysis of multisensory effects in the bouncing and streaming conditions, respectively.

Effects over the occipital scalp

This analysis was focused on differences in multisensory effects over visual cortices between the condition evoking the bouncing percept and the condition where a streaming percept dominated. Here, the statistical cluster plots showed an early effect between 50-60 ms in the bouncing condition over the occipital and parieto-occipital cortex that was not present in the streaming condition. We can see the location of this effect in the waveforms C and D in Figure 9. Note that the two waveforms in the bouncing condition diverge remarkably early in comparison to the passing condition (E, F) and display a peak at around 50 ms, respectively. The amplitude of this peak is larger in the bouncing condition. A second effect was distributed in a timeframe between 200 and 230 ms with a maximum located

bilaterally over occipital and parieto-occipital sites. This effect can also be seen in the difference waveform between the unbiased multisensory ($AV_{misaligned}$ - A_{alone}) and the unisensory visual ($V_{misaligned}$) condition in Figure 9. (E and F). This effect was absent in the illusion (bouncing) condition.

The statistical cluster plot of the illusion condition revealed two later phases of effects that were absent in the streaming condition. The first one was located in the timeframe between 220 and approximately 300 ms over the frontal cortex. Further, we observed a large distributed pattern of late effects between 400 and 500 ms spreading from occipital over parietal to central scalp sites.

3.5 The percept of bouncing and streaming (perceptual ambiguity condition)

In the last two conditions, subjects were instructed to willfully bias their percept towards bouncing ($AV_{aligned/lowvol}$) or streaming ($AV_{aligned/lowvol}$) after the sound intensity level was individually titrated to allow for perceptual ambiguity. Therefore, the physical stimulation in both conditions was identical. We had no prior hypothesis about the nature of possible differences between these conditions. We therefore proceeded to analyze differences in a descriptive manner using a statistical cluster plot (Figure 15.).

As in the other multisensory conditions, the evoked responses showed the distinct componentry that is typically found in response to an auditory stimulus. The lowering of sound pressure (from 75 dB to approximately 45 dB) to induce perceptual ambiguity resulted in greatly reduced amplitudes of the N1/P2

complex. Over the fronto-central scalp sites the amplitudes show a deflection with the magnitude of 3-4 μV whereas the other multisensory conditions showed amplitudes between 6 and 7 μV. What is interesting here is that, analogous to the comparison between the $AV_{aligned}$ and $AV_{misaligned}$ conditions, we found a suppression of the N1 component over bilateral fronto-central sites in the condition where bouncing was perceived. We found the maximal magnitude of this suppression to be located at more anterior scalp locations (central scalp), whereas the difference between N1 amplitudes was maximal at a further posterior location (centro-parietal scalp) in the $AV_{aligned}$ vs. $AV_{misaligned}$ comparison. Further, the waveform in the condition where participants were instructed to induce a bouncing percept appeared to be slightly accelerated. Finally, inspection of the statistical cluster plot also revealed a cluster of late effects (~400 ms) distributed over frontal and fronto-polar scalp regions. This cluster was also present in the ($AV_{aligned}$ - A_{alone}) vs. $V_{aligned}$ comparison but absent in the ($AV_{misaligned}$ -A_{alone}) vs. $V_{misaligned}$ comparison (refer to cluster plots above).

3.6. Control Analysis

One major concern about the validity of our results was that they may have been confounded with effects introduced by lack of counterbalancing of the order in which the conditions were presented. However, the order of presentation was varied, which allowed for an analysis of the influence that the order may have played. The foremost concern was that habituation to the stimulation may have

resulted in a reduction in the brain response, which in turn may have resulted in a relative decrease in evoked responses in conditions presented later.

To investigate whether such confounding factor may have played a role, we compared the auditory evoked responses in the A_{alone} and the $AV_{misaligned}$ condition in cases where: a) the A_{alone} condition was presented first, and b) when the $AV_{misaligned}$ condition was presented first. If the order of presentation was the major source for differences in the evoked responses then we predict that the ERP to the condition that was presented earlier will be larger in amplitude.

The graphs in Figure 16. show the evoked responses to the A_{alone} and the ($AV_{misaligned}$ – $V_{misaligned}$) conditions over central scalp locations (left and right hemisphere) in cases where the A_{alone} condition was presented first (top two) or vice versa (bottom two). As can be seen by visual inspection, the order of presentation does not appear to have any impact on the difference between the waveforms. The ERP to the A_{alone} condition is larger in amplitude when presented in a later block than the $AV_{misaligned}$ condition. I therefore consider it save to conclude that the observed effects over the auditory projection areas are not confounded by the order of the presentation of the conditions.

4. Discussion

4.1 Summary of rationale

In the past, it has been shown that the presentation of a transient sound can alter the percept of an otherwise ambiguous motion display (Sekuler et al., 1997).

While most participants perceive two identical objects as streaming past one another when moving toward one another and then coincide, a transient sound induces a bouncing perception. While a recent fMRI study (Bushara et al., 2003) investigated which neuronal structures are implicated in this illusion, the spatio-temporal aspects of the integration of sound and motion are still unknown.

In this study we investigated the spatio-temporal characteristics of this illusion using high-density EEG. We were interested in discovering where and at what time multisensory effects emerge in the cortex. We hypothesized that the sound may modulate visual cortical activity as indexed by a multisensory effect in the evoked response to the audiovisual stimulus. On the other hand, it is possible that auditory activity is affected by the concurrent visual stimulation. We expected these modulations to manifest themselves in the waveforms of the multisensory evoked responses over scalp locations, reflecting auditory and visual generator activity. Finally, we hypothesized that there may be a difference in multisensory effects (in timing, magnitude and topography) between the condition that gave rise to the illusion and the condition where streaming was perceived, since the subjective perceptions in these conditions were qualitatively distinct from one another.

4.2 Summary of Findings

We assessed multisensory effects for each $AV_{aligned}$ and $AV_{misaligned}$ condition, respectively. We found that in both multisensory conditions, the responses were mainly attenuated in comparison to the unisensory responses. This attenuation

was found widespread over the scalp and reached significance over a scalp location, reflecting auditory generator activity in the form of a suppression of the N1/P2 complex. This effect was stronger in the condition that evoked the bouncing percept.

The analysis of multisensory effects over visual cortices suggested a “late” multisensory effect (also an attenuation) over lateral occipital sites in a time window between 310ms and 330ms as indexed by a near-significant trend. No multisensory effects over visual cortices were found in the streaming condition despite large differences in the waveforms.

In order to isolate multisensory effects that were related to differences in the perception of the stimulation, we also compared both multisensory conditions directly. Visual inspection revealed an attenuation of the N1/P2 complex, which could, however, not be confirmed statistically. Evoked responses over the scalp were remarkably similar, and we found no statistical differences between both conditions in the waveforms over the auditory and visual scalp. A more exploratory analysis of the data using statistical cluster plots confirmed this and showed no systematic differences between both main multisensory conditions.

We also used an alternative exploratory strategy that was more sensitive to potential effects. Here, we used statistical cluster plots to compare both multisensory (unbiased) conditions with their unisensory visual control conditions, respectively. By doing so, we could confirm the N1/P2 attenuation effects over the auditory scalp. This effect appeared larger in the illusory condition. Interestingly, this exploratory analysis also revealed an early effect between

50ms and 60ms emerging bilaterally, with a right hemispheric dominance over parieto- occipital regions. This early difference was absent in cases where passing was perceived. We also observed two later differences that were absent in the streaming condition. One was distributed over frontal cortices in the timeframe between 220 ms and 300 ms and the other was a late (400ms-500ms), widespread effect spanning from occipital over parietal to central scalp sites. A difference around 200 ms-230 ms over occipital and parieto-occipital sites observed in the pass condition, was absent in the bounce condition.

In our last two conditions, the visual and auditory stimulation was identical and the auditory levels were titrated to allow for perceptual ambiguity. The critical difference between the conditions was that the observer was instructed to willfully bias the percept towards either bouncing or streaming, respectively. Interestingly we found a smaller amplitude of the N1-component in the condition with instructions to perceive bouncing. In addition, we found a cluster of late differences around 400ms over frontal and fronto-polar scalp regions that was absent in the passing condition. Note, that this anterior distribution resembles, although with slightly different timing, the anterior activation found in the multisensory condition eliciting the bouncing percept.

4.3 Discussion of findings

4.3.1 Multisensory attenuation in the auditory cortex:

As mentioned in the introduction, Bushara et al. found that an increase in activity in multisensory areas was coupled with a relative decrease in activity in visual

and auditory sensory specific areas. The authors speculated that this effect was due to competitive interactions between multisensory and predominantly unisensory processing networks. Here, multisensory networks down-regulate activity in primary sensory sites. This modulation was present in both multisensory conditions but was larger in cases where participants perceived a collision between the two objects. Our electrophysiological data also showed reduced activity over scalp sites reflecting auditory and visual generator activity when multisensory conditions were compared with the combined unisensory auditory visual conditions. The attenuation was larger over auditory scalp projection sites in the bouncing condition. However, we were unable to confirm an increase in activity in multisensory cortices.

The findings of multisensory attenuation in the study by Bushara et al. (2003) and in particular our present study, are not in line with the classical multisensory enhancement found in most multisensory experiments. It could be argued that instead of representing a multisensory effect, the attenuation is a result of an attentional mechanism. Watanabe et al. (1998) proposed that the effect of sound on the bistable motion illusion is due to an attentional shift away from the visual motion process caused by a transient stimulus (auditory, visual or tactile) during the coincidence of the visual objects. Further, functional imaging studies have repeatedly shown a decrease in activation in sensory-specific cortices in paradigms where subjects were exposed to a concurrent stimulus in another modality (Kawashima, et al., 1995; Lewis et al., 2000; Bense et al., 2001; Laurienti et al., 2002; Wright et al., 2003). It is also well known that attention

modulates auditory N1 and P2 components (see Näätänen, 1991 for a review). Bushara et al. (2003) argued that if attention–based top-down modulation of sensory-specific cortices caused this effect, then trial-by-trial attentional shifts should have led to an increase of activation in at least one modality. Decreased activity was found in both auditory and visual sensory-specific cortices suggesting an alternative mechanism at play.

In our experiment visual attentional deployment was controlled by the presence of a visual task in the A_{alone}-condition[12]. Further, a general attentional effect does not explain differences between multisensory conditions with an increased attenuation of N1/P2 in the condition where bouncing was perceived, unless we assume stronger attentional attenuation in the bouncing condition (this being hard to believe when the sound is what gives rise to the illusion and there are otherwise no significant differences in the visual ERP's). Further, this effect was replicated for N1 when stimulation was identical and bouncing and passing percepts were induced voluntarily. Our results, therefore, do not support the hypothesis proposed by Watanabe et al. (1998) that the illusion is induced by a brief distraction from the visual motion caused by the auditory transient. If this was the case, there would have been no or little attenuation of the auditory activity in the bouncing condition and a stronger attenuation in the passing condition.

The amplitude reduction of the N1/P2 responses is not a novel finding. Besle et al. (2003) recorded ERP's while subjects performed an auditory

[12] In fact the A-condition in our experiment is not a pure auditory–alone condition. We labeled it A-alone to emphasize the absence of the moving squares

recognition task among four different syllables presented in an auditory, visual, and audiovisual condition. The investigators found that audiovisual stimulation led to behavioral facilitation and multisensory ERP effects were found to be expressed mainly as a decrease of the N1 generator activity. The authors interpreted their findings as suppressive speech-specific audiovisual integration mechanisms likely to operate on a pre-representational stage of stimulus analysis, probably via feedback projections from visual and/or polymodal areas. The authors speculate that lip movements have facilitated feature analysis of the syllables by a depression mechanism in the auditory cortex similar to that found in the visual cortex for object processing.

Van Wassenhove et al. (2005) also reported reduced N1/P2 activity over the centro-parietal scalp to an audiovisual (McGurk) speech stimulus in comparison to an auditory-alone condition. Unfortunately, the authors failed to subtract the activity generated by the visual stimulus from the AV-waveform and their data are therefore hard to interpret. However, we can estimate the visual contribution to the AV-waveform from the V-waveform over the centro-parietal scalp location, and it looks as if visual contribution is rather weak, which allows the assumption of a genuine N1/P2 effect. The authors attributed the sub-additivity in the N1/P2 response to a deactivation mechanism that minimizes the processing of redundant information across senses to extract novel information. In their analysis-by-synthesis model van Wassenhove et al. propose that the precedence of the visual information allows the system to build an abstract speech representation that provides the context in which auditory inputs are

being evaluated. Within this framework, the sub-additivity in neural populations in the auditory cortex is caused by projections from networks containing an abstract representation of the speech stimulus. This input aids the auditory cortex in extracting relevant frequency ranges from incoming auditory stimuli. If this is true, we would predict that the attenuation of auditory activity would vary depending on the visual input, especially between cases where the auditory signal is mainly redundant to the visual signal, such as in cases where articulation provided clear place of articulation (more N1/ P2 attenuation according to the model) and cases where the articulation underspecifies the AV-stimulus. The authors failed to show a visual modulation of the ERP amplitude varying with stimulus identity, which does, in fact, contradict the proposed model that was designed to explain their data. However, it is not possible at this time to analyze and evaluate the model and data of the van Wassenhove et al. publication in detail, but it should be expressed that the general notion that the visual signal, in some form or another, constrains the processing of the incoming auditory input is, nevertheless, reasonable and a likely mechanism in AV-speech integration as was discussed in the general introduction of this thesis. Their model, however, makes strong assumptions (anatomical and functional), that by themselves, are not supported by the empirical data of the study.

In a more recent paper by Reale et al. (2007) the auditory-visual response recorded from intracranial electrodes in humans to a spoken syllable was greatly reduced in comparison to the syllable presented in the auditory modality alone. A large attenuation could be seen at deflections of the waveforms, similar or

equivalent to the auditory P1, N1 and (to a lesser degree) P2 component. Surprisingly, the authors refrained from explaining or even mentioning this attenuation despite its large magnitude that reached a reduction in amplitude of 20% to 30% for the first and second deflections. Consequently, these data support the validity of our finding but do not aid in its explanation.

The question remains of how our findings fit with those reported by Besle et al. (2003) and van Wassenhove et al. (2005). Both authors suggest that the attenuation of auditory generator activity in the N1/P2 range is speech-specific, although the mere fact that these experiments use speech stimuli cannot be considered to be sufficient support for this notion. Our finding of N1/P2 depression suggests that a more basic mechanism is at work, since our stimuli lack complexity and no task was involved. Despite the strong differences between the stimulation used in these two studies and those in our experiment, all four (including Reale et al., 2007) studies have one major aspect in common. The visual stimulus is dynamic and precedes the onset of the auditory signal. If the dynamic rather then the phonetic nature of the visual stimuli was the cause for the observed effects then we would expect the auditory attenuation to be absent in cases when auditory speech is presented with still images of faces. This is indeed the case, as a study by Miki et al. (2004) has shown. In this MEG-study, the M100 response was unaffected by the concurrent presentation of vowel sounds and still images of open and closed mouths pronouncing the vowels.

The hypothesis that the audiovisual attenuation is due to the presence of anticipatory motion of the stimulus (speech or non-speech) preceding and predicting the occurrence of the sound was most recently supported by Stekelenburg and Vroomen (2007). They found that the auditory evoked N1 and P2 potentials were smaller when preceded by articulatory visual motion but also (and actually more so) by non- speech motion like clapping hands. No topographical differences were observed between the auditory and the audiovisual N1 led the authors to suggest that the audiovisual integration modulated the activity of the neural generators of the N1. The authors conclude that the neural correlates of audiovisual integration of speech are not speech-specific because they are also found in non- speech audiovisual events.

Visual articulation is a biological motion stimulus and also engages motion processing areas MT and MST as well as parts of STS/G (Allison et al., 2000; Puce et al., 2003). The STS/G is an important cortical site for the processing of different kinds of biological motion, and it is possible that both kinds of motion stimuli, the one used in our paradigm and facial articulation, involve motion networks that engage the processing of sound in the auditory cortex, possibly with the involvement of superior temporal polysensory areas (STP). It is indeed possible that the scalp waveforms reflect STS/G generator activity which roughly has the same orientation as the supratemporal plane. Recall, that the maximum effect was observed posterior to the scalp locations where auditory generator activity would commonly be observed.

In addition to our finding of a general multisensory attenuation of the auditory evoked responses we found that this effect appeared to be slightly, but nevertheless larger, in the condition that evoked the bouncing percept. In the study by Bushara et al. (2003), the perception of bouncing and passing happened spontaneously during trials of identical stimulation. The authors suggested that the percept was evoked because of greater activity in multisensory networks but don't explain why there was a shift between both percepts in the absence of differences in stimulation. In the situation of spontaneous change in percept and without accurate timing information it is difficult to determine whether the observed differences in neural activation represent the cause of the percept, its consequence or the neural correlate of the percept itself. In our study, the percept and the assumed activity in neural networks underlying the percept, was evoked by a minor difference in the alignment of the visual stimuli. In the multisensory condition where the objects were aligned, they were reliably perceived as bouncing off one another. We hypothesize that under these visual conditions the sound directly causes the percept of an impact between the objects. It has been suggested elsewhere that the percept of two objects in motion with intercepting trajectories as passing through or by one another, is the default condition of the perceptual system (Sekuler et al., 1997). Indeed, in the physical world objects in (uniform) motion don't change their trajectories unless acted upon by an outside force (Newton's first law of motion). The impact of force releases energy that reaches us through different sensory modalities most commonly involving vision and sound. Sound is

therefore a reliable indicator of dynamic changes in the environment involving motion, and it is therefore reasonable to assume that our brains are set up to utilize such environmental regularities to respond fast and effectively.

The question of what process this auditory attenuation reflects, can not conclusively be answered, given the current information. Nonetheless, there is a god deal of room for speculation. Multisensory convergence does not necessarily have to result in an amplification of neuronal processes, as Fort and Giard (2004) have pointed out. Depending on task demands, neural indices of integration may take on the form of a depression, which may also signal multisensory facilitation. However, it is difficult to interpret our results in light of this scenario. It is unclear why the particular stimulus configuration that evoked the bouncing percept would result in a facilitation of the processing of the auditory stimulus.

Another speculation on the mechanism responsible for this effect is motivated by the similarities in studies where this effect has been found. As mentioned, our experiment, as well as other studies that found an attenuation of the auditory evoked potential, used a visual motion stimulus that preceded and predicted the upcoming auditory stimulus. The auditory stimulus was delivered in a regular pattern, but the visual stimulus in the multisensory condition could have served as a warning cue for the upcoming auditory stimulus. In our study the visual stimulus is unrelated to the task, is delivered many times, and is presented in a predictable fashion. This highly redundant and predictable stimulus may, therefore, receive a low amount of attentional resources and is therefore “filtered out”. The difference between the multisensory and the unisensory condition may

lie in the efficiency of that filtering process. In the multisensory conditions, the auditory stimulus may be filtered out more efficiently because of the visual stimulus serving as a warning for the upcoming auditory stimulus.

A further related speculation may be that the brain deploys a small amount of resources to the auditory stimulus not only because of its redundancy and irrelevance to the task, but because of its intensity. Our auditory stimulus was delivered with a fairly high intensity (75dB) in order to induce the visual illusion. It is possible that this filter represents a protective function of the auditory system. It is known that the outer hair cells in the cochlea receive efferent input from the medial olivary nucleus. Olivocochlear efferent neurons permit the central nervous system to control the way sounds are processed in the auditory periphery, with the potential to improve the detection of signals in background noise in order to selectively attend to signals or to protect the periphery from damage caused by excessive sound pressure (see Guinan, 1996 for a review). Protective attenuation in the periphery could have caused the reduced activation in the auditory cortex. However, this mechanism does not explain why this attenuation appeared to be stronger in the bouncing condition.

4.3.2 Multisensory attenuation in the visual cortex

In the multisensory bistable motion illusion, the transient sound induces a change in the percept of the moving stimuli. Subjectively, this effect is visual in nature and it was therefore reasonable to predict that presence of the sound, when evoking a bouncing percept, would affect processing in visual motion processing

areas such as MT+. We were surprised about the absence of such effects when directly comparing our multisensory conditions. Instead, we found a mainly subadditive effect in both multisensory conditions, with a larger attenuation in the passing condition. Overall, these are consistent with the findings of the study of Bushara et al. (2003). In their study, sensory-specific attenuation was accompanied by an increased activation in multisensory convergence sites. The fact that the majority of these sites are in subcortical structures may be the reason for the predominace of subaddictive effects on the scalp surface.

Our exploratory analysis revealed a pattern of multisensory effects. We found the earliest effect over the occipital scalp at around 60ms, possibly reflecting an early modulation of visual processing by the auditory stimulus. Such early effects have been found for static stimuli (Giard and Perronet, 1999; Molholm et al., 2002) and may, therefore, not be specific to the dynamic nature of our visual stimulus. This early modulation may be related to the percept of the collision of the two objects, since this effect is missing in the condition where streaming is perceived, but such conclusions should remain cautionary. One may speculate that the sound only modulates processes in the visual cortex when the visual stimulation is truly ambiguous- a condition that is only given in the $AV_{aligned}$ condition. This early sensory modulation may then facilitate the later perceptual impression of the repulsion of the two objects as a consequence of a collision.

Also speculative are interpretations of the late anterior differences found in the conditions where participants willingly generated the bouncing and passing percepts. This relative attenuation in the bouncing condition could be neural

activity associated with the perception of the event or, on the other hand, a reflection of the effort to generate the percept in question. It should be noted here that Bushara et al. found increased frontal (dorsolateral and medial prefrontal) activity in the bouncing condition.

4.4 Discussion of problems related to the design and analysis:

A direct comparison between both multisensory conditions, one eliciting a bouncing percept and the other eliciting a streaming percept, revealed no significant differences. Recall that participants were asked after each block to retrospectively give an estimation of the percentage of bouncing percepts. Although participants reported the appropriate percept in either condition, respectively (>80%), the stimuli may have remained ambiguous, which in turn may have introduced error variance resulting in the failure to find a robust effect between conditions. A possible reason for this may be a lack of general attentional deployment to the stimuli. As described in the introduction, a dependency of the perception of the bistable motion illusion on attention has been suggested by Watanabe and Shimojo (1998). In their study, a second task was introduced to detract attention from the bistable motion illusion. They found that when subjects were engaged in a second task, bouncing percepts were significantly reduced. In our experiment, a similar distraction may have been introduced by the task that was used to assure similar attentional deployment between conditions where participants had to respond to a target. However, no response was required to the audiovisual stimuli in both conditions which could

have contributed to the lack of attentional deployment. On the other hand, the blocked design, being a repetitive and monotonous stimulation may have (also) led to boredom and thus reduced attention to the percept of the event. An improved design would involve the pseudorandom presentation of conditions with a report of the percept after every trial. Inter-trial intervals would be randomized to avoid anticipatory effects.

Another issue to be discussed is related to the exploratory nature of the study. As mentioned in the method section, the visual stimuli that are usually used in ERP studies are of transient nature. Therefore, the morphology of the evoked response to our (or similar) dynamic visual stimuli was not known to us prior to the study. We used the average of the visual evoked responses to derive “components” or time windows of the ERP that we used subsequently to test for differences between conditions. These “components” had no resemblance to known visual evoked responses, nor was anything known about their functional properties. The choice of time windows was, therefore, somewhat arbitrary, which raises doubts about the usefulness of this analysis strategy in this context

Table 1.

All eight conditions with modality of stimulation, alignment and percept (streaming or bouncing) that the stimulation in each condition evoked.

Table 1: Summary of conditions

	Modality		Alignment		Percept	
Condition	Visual	Auditory	Aligned	Misaligned	Streaming	Bouncing
1. A_{alone}		X				
2. $V_{aligned}$	X		X		X	
3. $V_{misaligned}$	X			X	X	
4. $AV_{aligned}$	X	X	X			X
5. $AV_{misaligned}$	X	X		X	X	
6. $AV_{aligned/lowvol}$	X	Low vol.		X	X	
7. $AV_{aligned/lowvol}$	X	Low vol.	X			X
8. $V_{reverse}$	X			X		X

Table 2.

Presentation order of conditions

Table 2.

Subject	Order of conditions
1	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$,
2	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$,
3	$V_{aligned}$, $AV_{aligned}$, $AV_{misaligned}$, A_{alone}, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
4	$V_{aligned}$, $AV_{aligned}$, $AV_{misaligned}$, A_{alone}, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
5	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$,
6	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
7	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
8	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$, $V_{reverse}$
9	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$,
10	$V_{aligned}$, $AV_{aligned}$, $AV_{misaligned}$, A_{alone}, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
11	$V_{aligned}$, $AV_{aligned}$, $AV_{misaligned}$, A_{alone}, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
12	A_{alone}, $V_{aligned}$, $AV_{misaligned}$, $AV_{aligned}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$
13	$V_{aligned}$, $AV_{aligned}$, $AV_{misaligned}$, A_{alone}, $V_{reverse}$, $V_{misaligned}$, $AV_{misaligned/lowvol}$, $AV_{aligned/lowvol}$

Table 3.

Lists the peaks of the components that were identified by visual inspection in Figure 5.

Table 3: Components in the average visual ERP

Component	Peak (ms)	Time window (ms)
1	70	60-80
2	240	230-250
3	320	310-330

Table 4.

Peak latencies of maximal differences between $V_{reverse}$ and $V_{misaligned}$.

Table 4: Components associated with motion reversal

Component	Peak (ms)	Time window (ms)
1	275	265-285
2	460	450-470

Figure 1.

Timing and arrangement of visual stimuli for $AV_{aligned}$ and the $AV_{misaligned}$ conditions**.**

Figure 1: Stimulus schema.

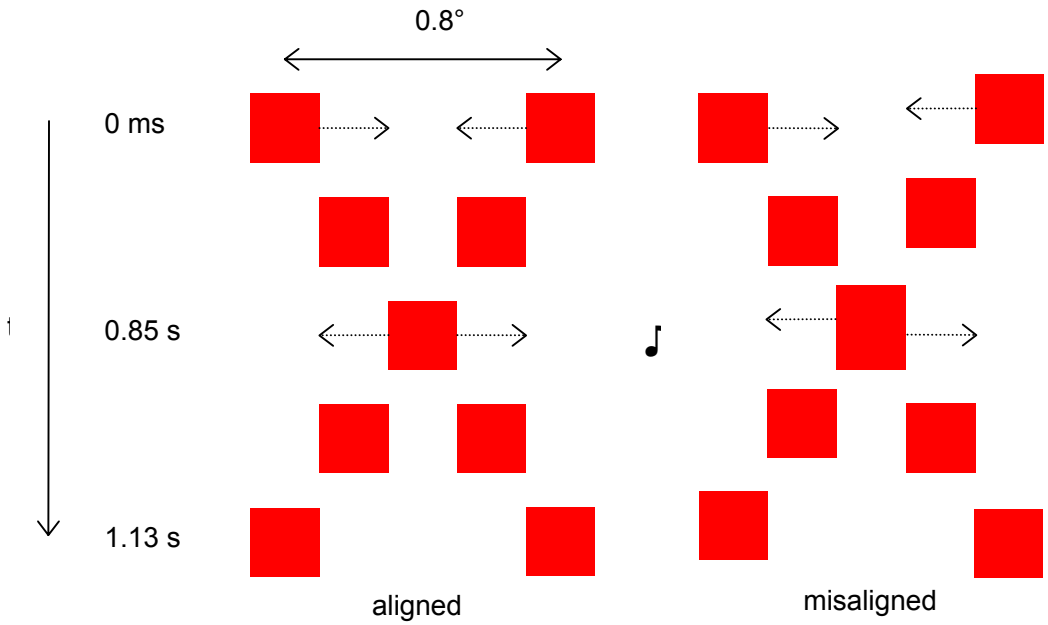


Figure 2.

The auditory-alone evoked response (A_{alone}: dashed trace) and the averaged visual evoked response to the unisensory motion stimuli (derived by averaging the $V_{aligned}$ and $V_{misaligned}$ conditions - solid trace) over bilateral frontal, fronto-central, parieto-occipital and lateral-occipital scalp locations. The white dots on the centrally presented scalp reconstructions display the locations of the electrode sites for which waveforms are plotted.

Figure 2: Unisensory conditions: Auditory alone (A_{alone}) and the visual-alone motion response.

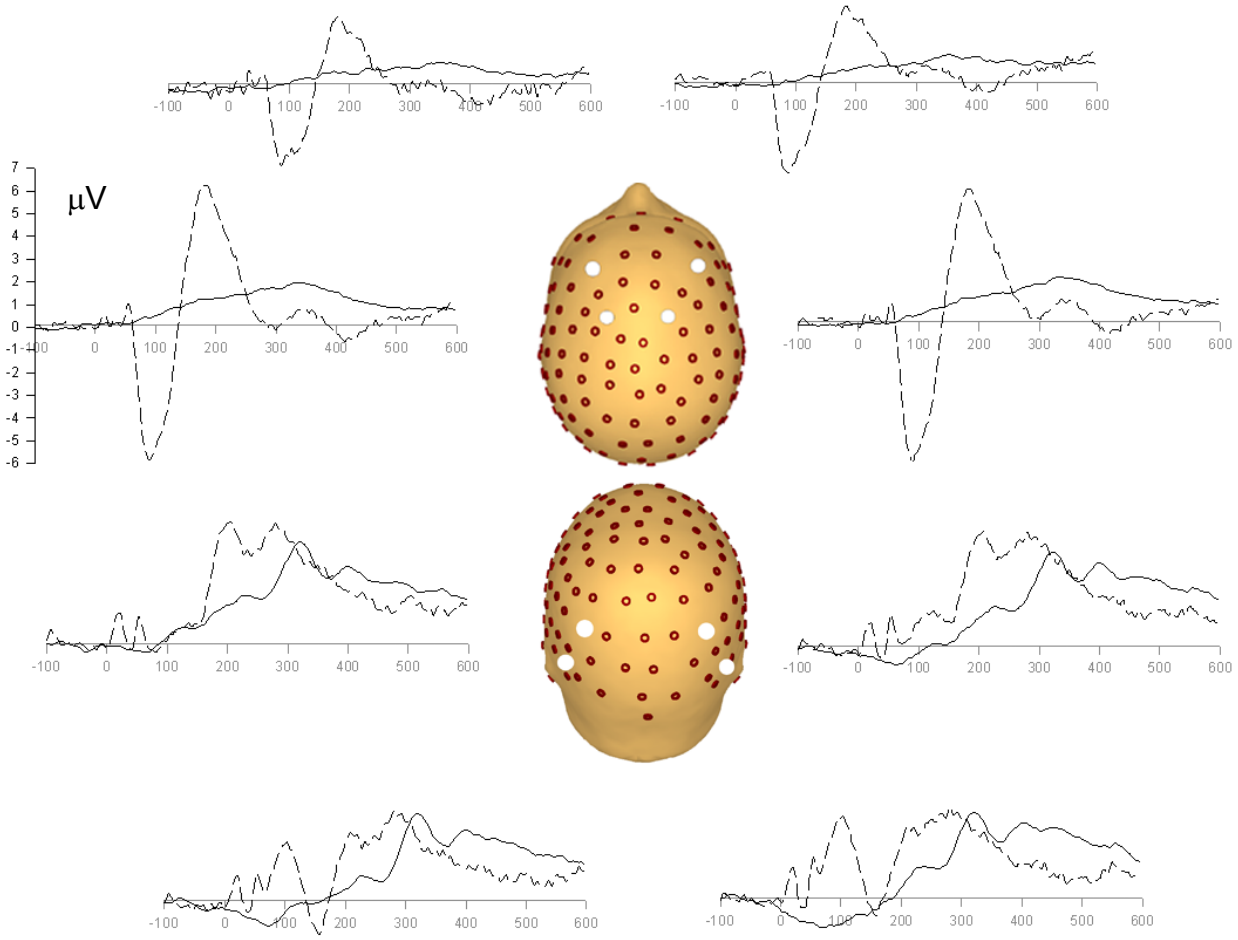


Figure 3.

Waveforms of the three unisensory visual responses ($V_{aligned}$: dotted; $V_{misaligned}$: solid; $V_{reverse}$: dashed) over bilateral frontal, fronto-central, parieto-occipital and lateral-occipital scalp locations.

Figure 3: Visual alone conditions ($V_{aligned}$, $V_{misaligned}$, $V_{reverse}$).

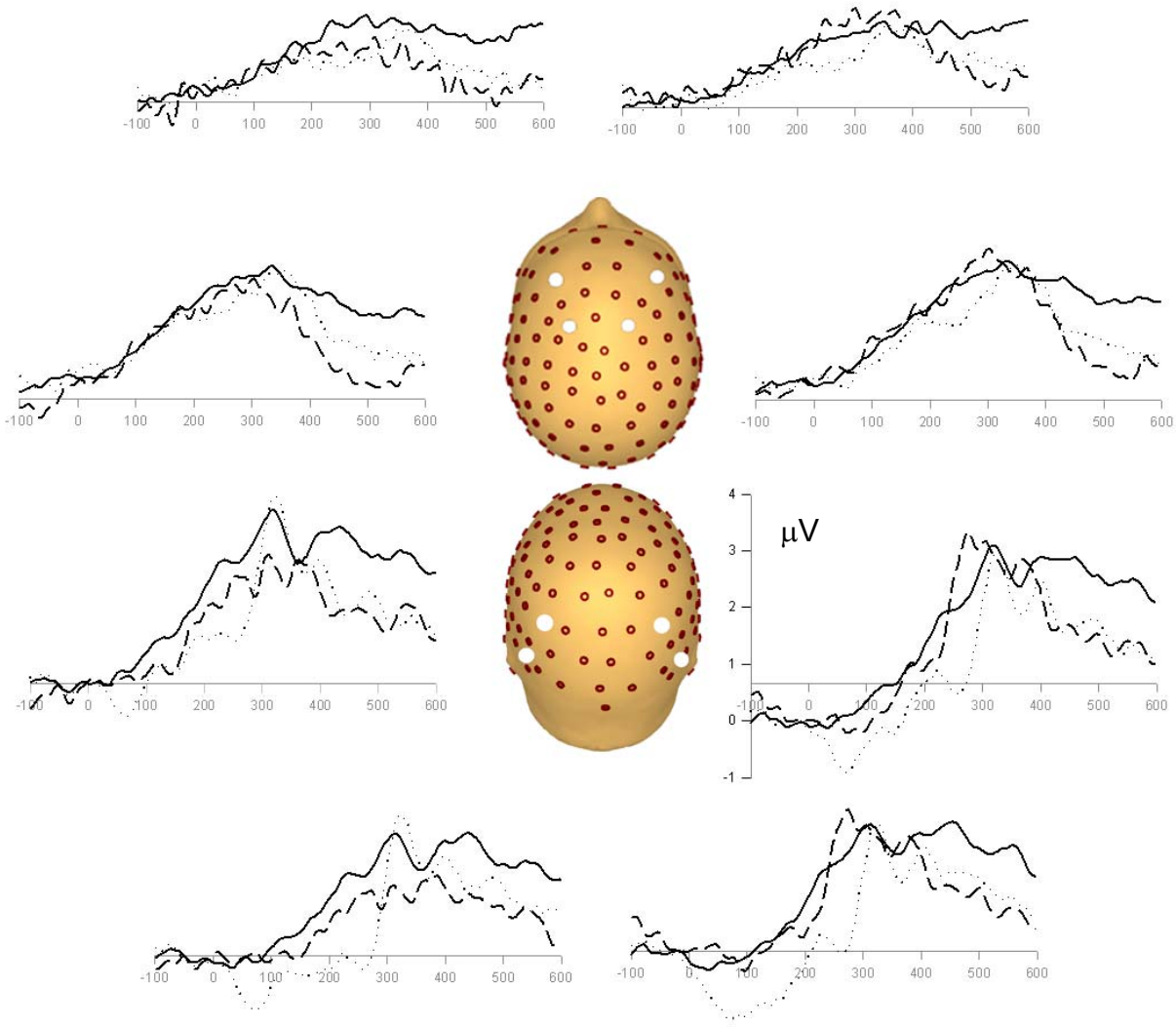


Figure 4.

The scale on the right indicates the color code for the p-value. The map on the left is a map of p-values in time (x-axis) over all 128 electrode locations (y-axis). Scalp locations are ordered from posterior (occipital scalp) to the anterior (fronto-polar) scalp and from left to right from each individual scalp region, respectively. The legend on the right contains the color code for the p- values.

Figure 4: Statistical Cluster Plot for the comparison of the unisensory visual conditions ($V_{aligned}$ vs. $V_{misaligned}$)

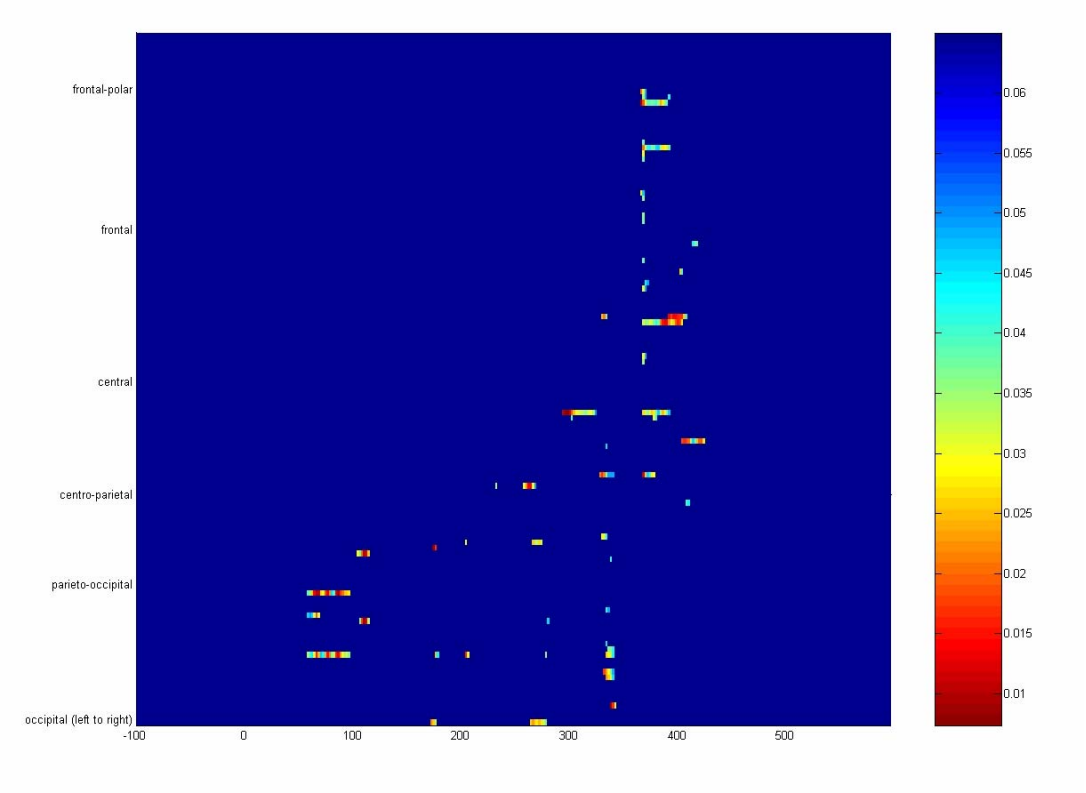


Figure 5.

The plotted waveform represents the average response of the conditions $V_{aligned}$ and $V_{misaligned}$ over an occipito-parietal scalp location. Arrows indicate components that were chosen to constrain the analysis of differences between multisensory conditions.

Figure 5: Components in the Average visual-alone ERP.

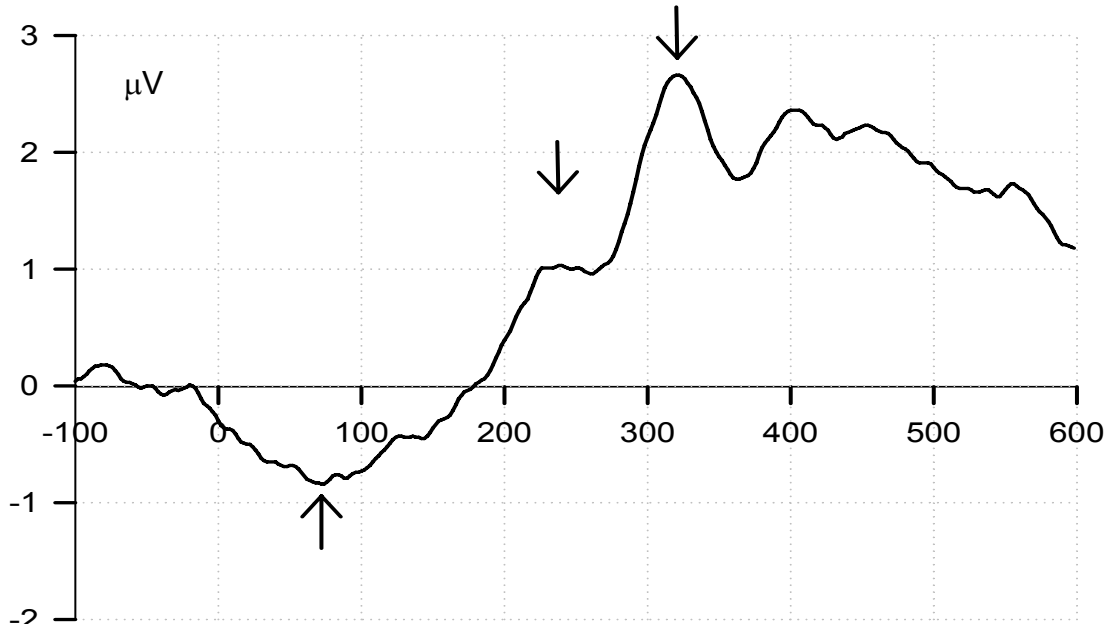


Figure 6.

$V_{reverse}$ (dashed), $V_{misaligned}$ (solid black), difference waveform (solid gray).

Figure 6: Components associated with motion reversal in the $V_{reverse}$ condition.

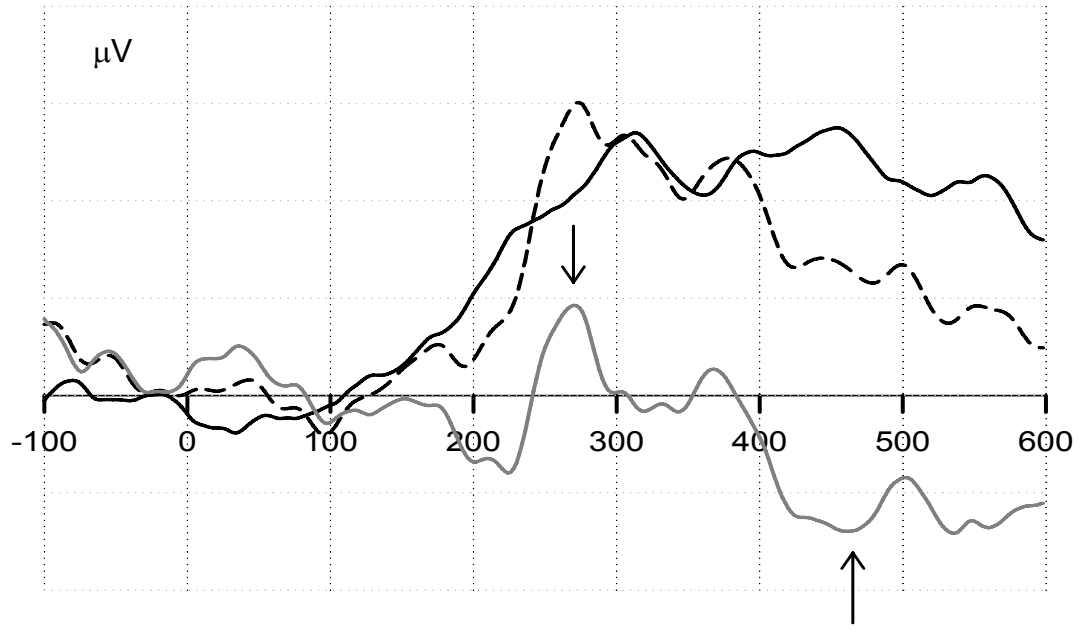


Figure 7.

Multisensory waveforms ($AV_{aligned}$, $AV_{misaligned}$: both solid trace) and unisensory evoked potentials (A: dashed trace, $V_{aligned}$: gray, $V_{misaligned}$: gray) over fronto-central and lateral occipital scalp locations. Multisensory waveforms of the bounce-condition (aligned) are in A, B, E and F, whereas C, D, G, H contain waveforms evoked in the passing condition.

Figure 7: Multisensory audiovisual evoked potentials ($AV_{aligned}$, $AV_{misaligned}$) and unisensory evoked potentials (A_{alone}, $V_{aligned}$, $V_{misaligned}$)

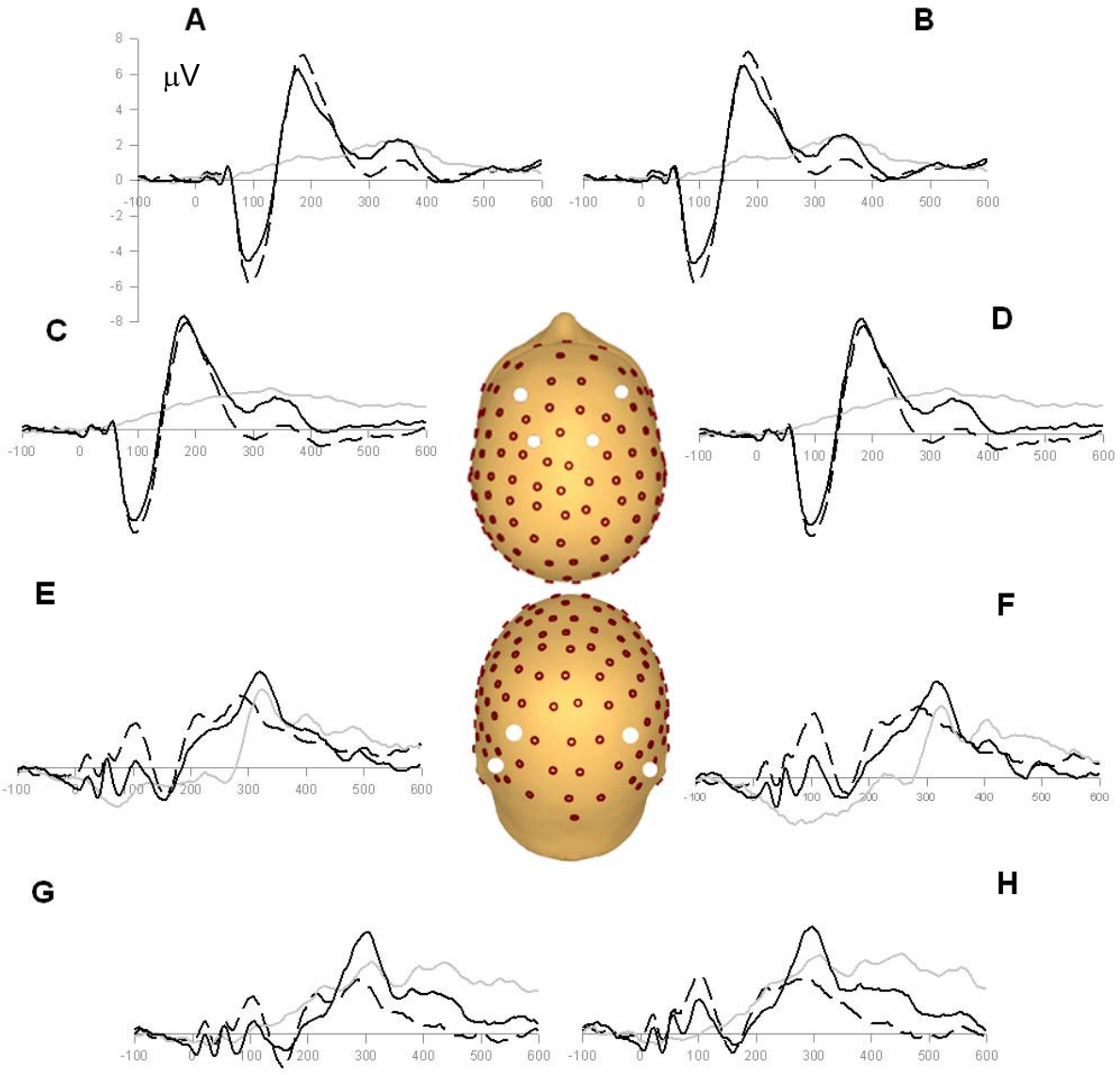


Figure 8 A and B.

Multisensory (AV: solid trace) and the sum of the unisensory evoked responses (A_{alone} + V: dashed trace) over bilateral frontal, fronto-central, occipito-parietal and lateral occipital scalp locations. A.: $AV_{aligned}$ vs. (A_{alone} + $V_{aligned}$). B.: $AV_{misaligned}$ vs. (A_{alone} + $V_{misaligned}$).

Figure 8 A: $AV_{aligned}$ vs. (A_{alone} + $V_{aligned}$)

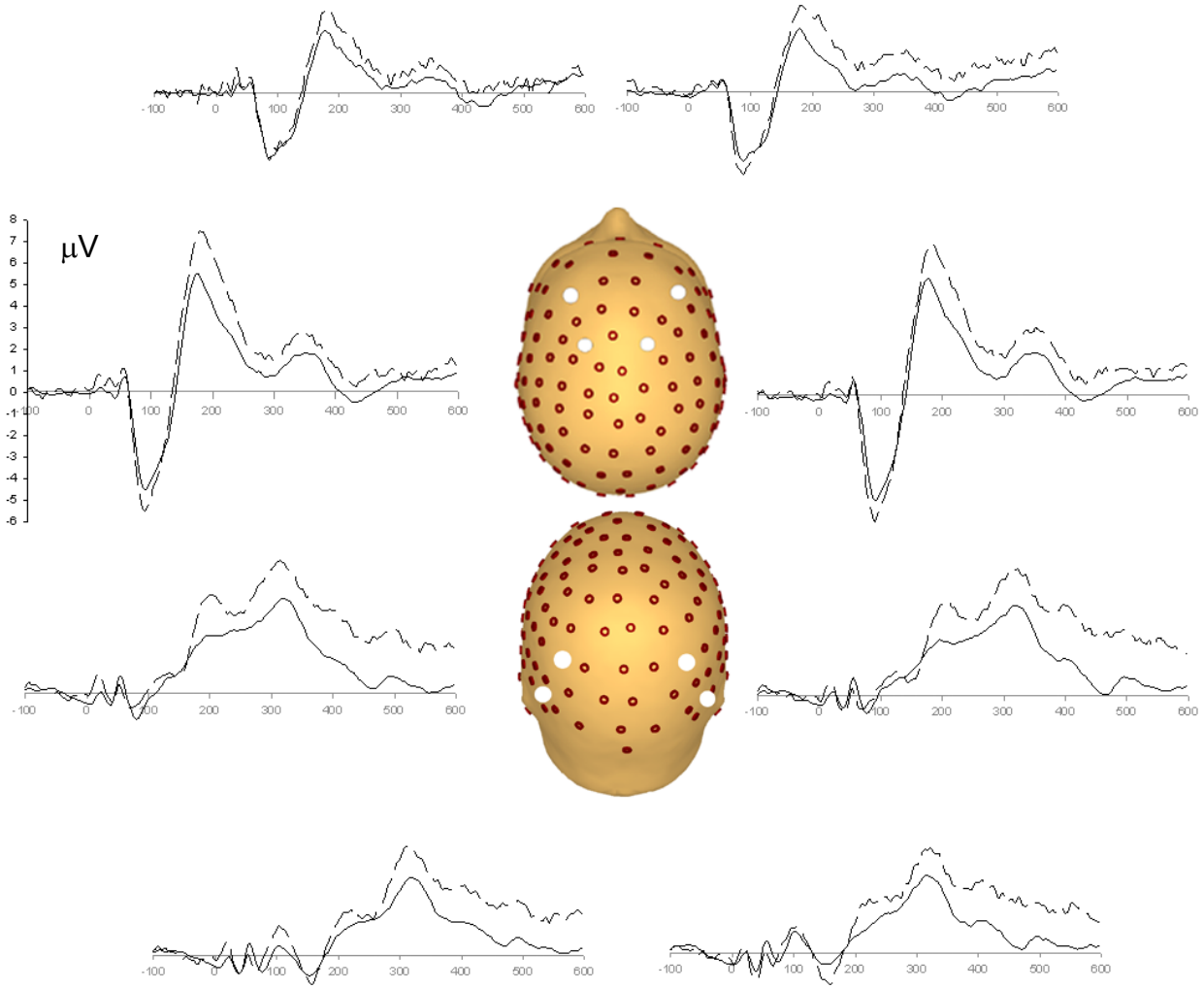


Figure 8 B: $AV_{misaligned}$ vs. (A_{alone} + $V_{misaligned}$)

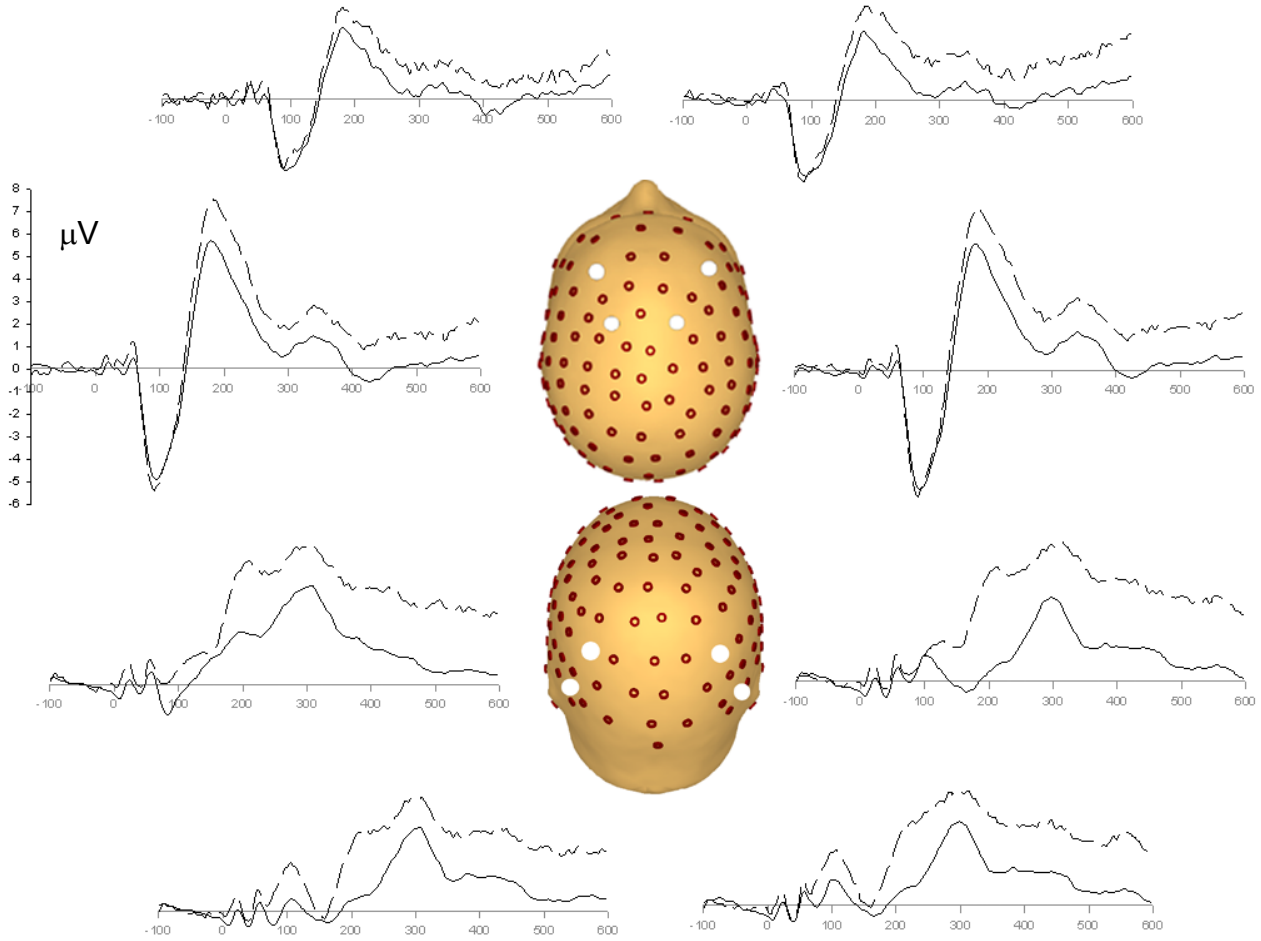


Figure 9.

Unbiased multisensory ($AV_{aligned}$ – $V_{aligned}$; $AV_{misaligned}$ - $V_{misaligned}$: both solid traces) and unisensory ERP's (dashed traces) over the central (A, B) and lateral occipital scalp (C, D, E, F). Graphs A, E and F contain the audiovisual ERP's associated with the streaming percept and graphs B, C, and D represent audiovisual ERP's in the conditions that produced a bouncing percept. The gray traces represent difference waveforms between the multisensory and the unisensory waveforms.

Figure 9. Unbiased Multisensory and Unisensory ERP's

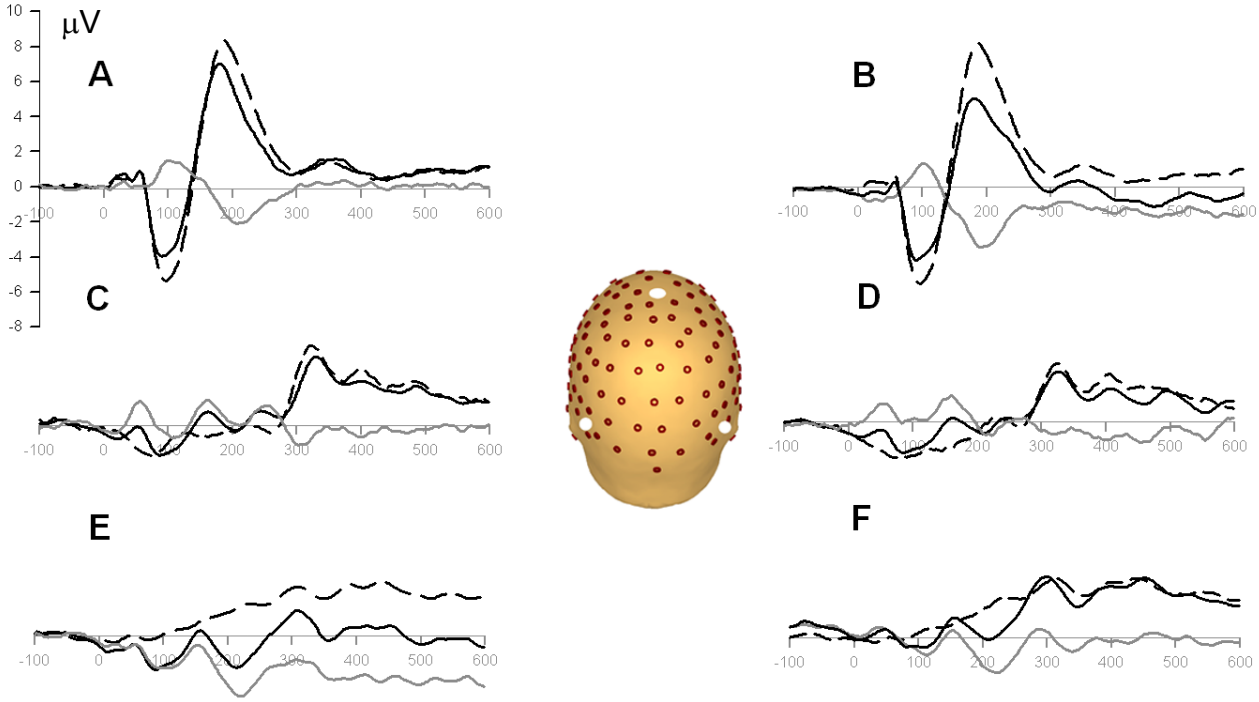


Figure 10.

Voltage maps of the difference waveforms between the unbiased multisensory and the auditory unisensory condition at the peak amplitudes of the N1 and P2 components. Bouncing condition: $(AV_{aligned} - V_{aligned}) – A_{alone}$; Passing condition: $(AV_{misaligned} – V_{misaligned}) – A_{alone}$.

Figure 10: Voltage maps of the difference waveforms between the unbiased multisensory and the unisensory auditory conditions

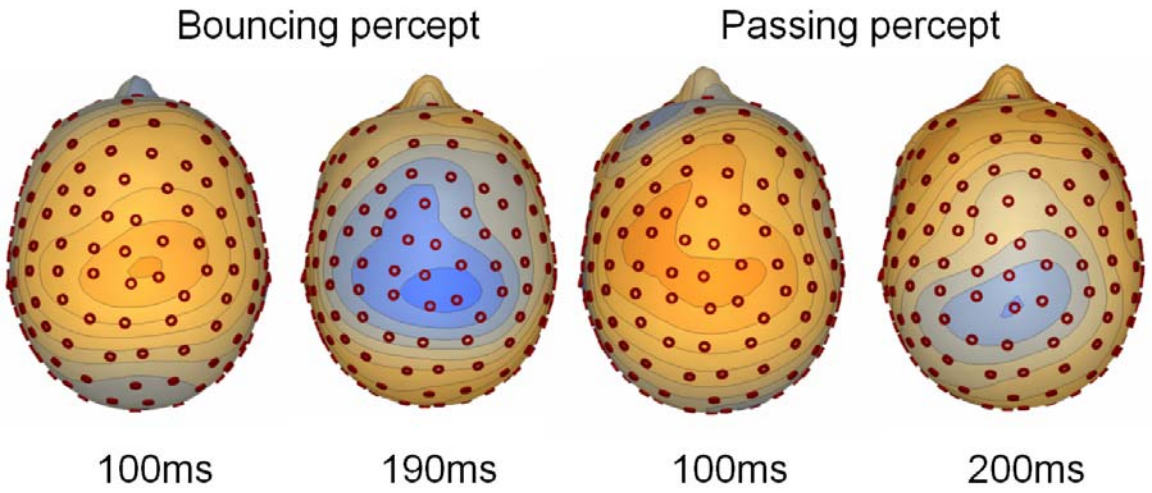


Figure 11.

Voltage maps for the difference waveforms ($AV_{aligned} – A_{alone}$) vs $V_{aligned}$ at 55 ms and 160 ms.

Figure 11: Voltage maps for the difference waveforms between the unbiased multisensory and unisensory visual condition.

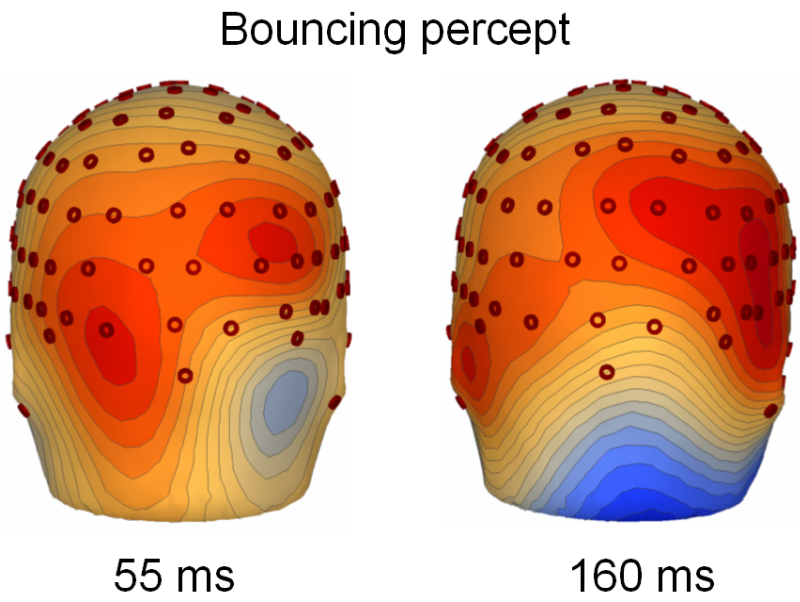


Figure 12.

Waveforms of the $AV_{aligned}$ (solid) and $AV_{misaligned}$ (dashed) condition over frontal (A, B), fronto-central (C, D), occipito-parietal (E, F), and lateral- occipital (G, H) scalp locations.

Figure 12. $AV_{aligned}$ and $AV_{misaligned}$ across the scalp

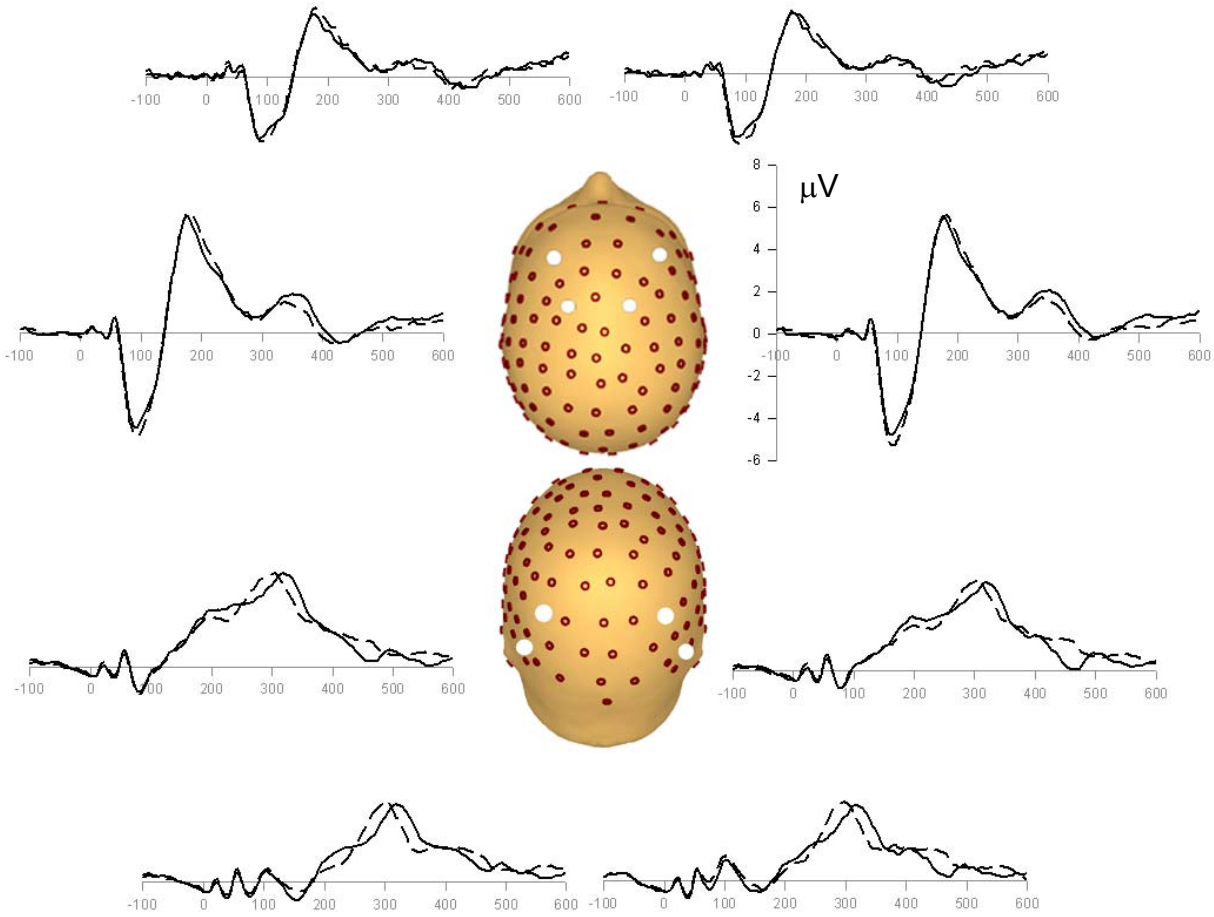


Figure 13.

Statistical cluster plot $AV_{aligned}$ vs. $AV_{misaligned}$. The scale on the right indicates the color code for the p-value. The map on the left is a map of p-values in time (x-axis) over all 128 electrode locations (y-axis). Scalp locations are ordered from posterior (occipital scalp) to the anterior (fronto- polar) scalp and from left to right fro each individual scalp region respectively. The legend on the right contains the color code for the p- values.

Figure 13: Statistical cluster plot for the comparison between $AV_{aligned}$ and $AV_{misaligned}$

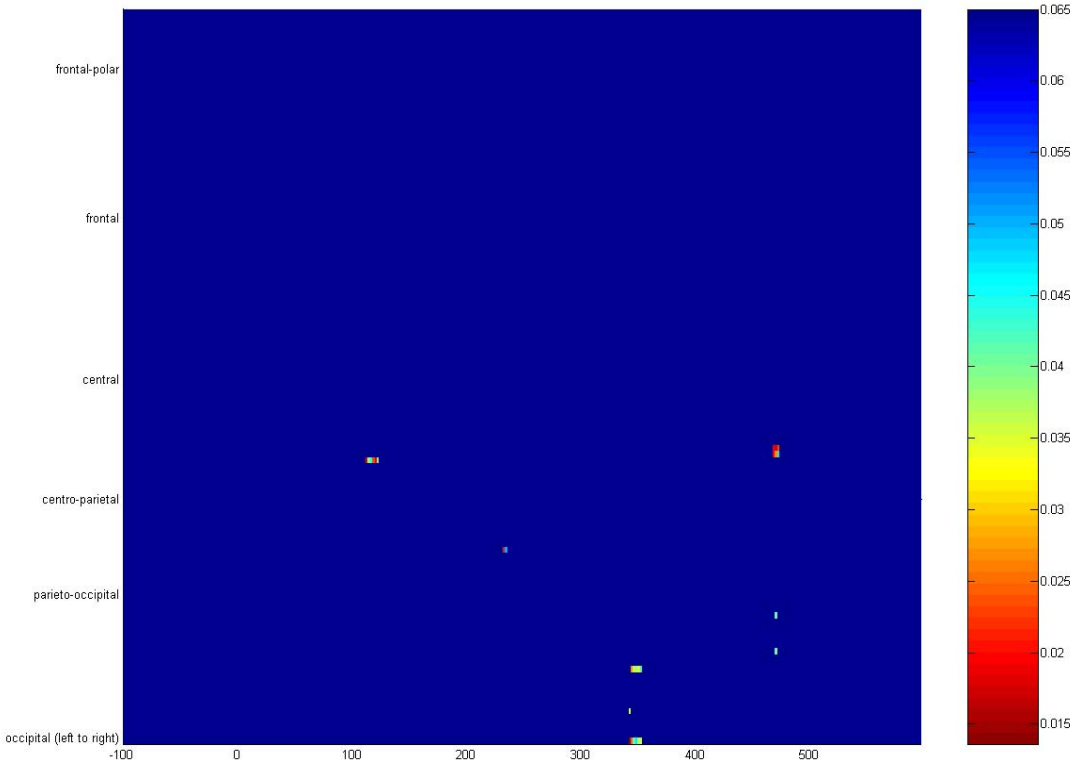


Figure 14.

Top: ($AV_{aligned}$ - A_{alone} vs. $V_{aligned}$); Bottom: ($AV_{misaligned}$ – A_{alone} vs. $V_{misaligned}$). The plots represent maps of p-values in time (x-axis) over all 128 electrode locations (y-axis). Scalp locations are ordered from posterior (occipital scalp) to the anterior (fronto-polar) scalp and from left to right for each individual scalp region, respectively. The legend on the right contains the color code for the p- values.

Figure 14: Statistical Cluster Plot: Multisensory Effects.

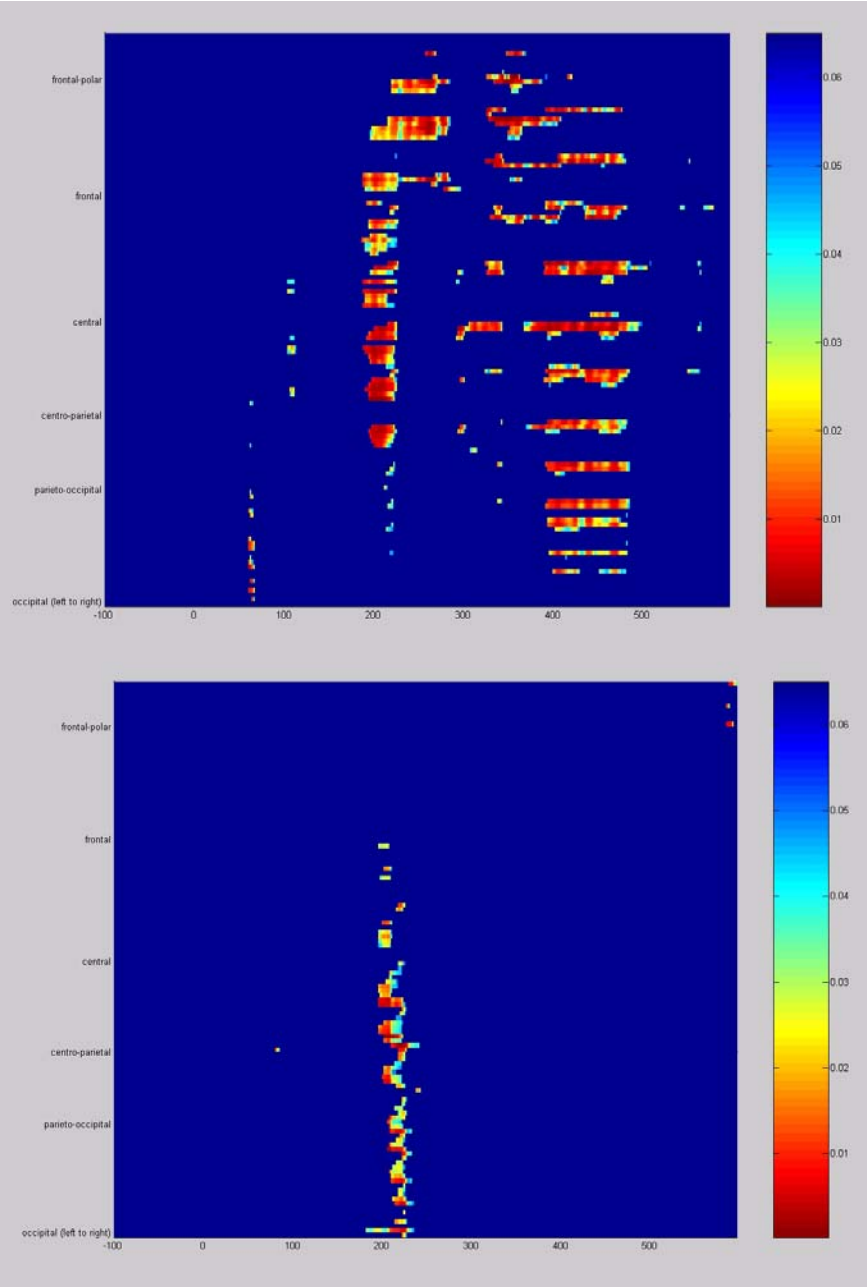


Figure 15.

Statistical cluster plot with waveforms. AV_{lowvol} with the instruction to perceive the squares as bouncing is represented by the solid trace whereas the waveforms corresponding to a passing percept are represented by a dashed trace. Graphs A and B show evoked activity over the bilateral fronto-polar scalp, whereas graphs C and D show waveforms over the fronto-central scalp. In the lower panel, the map on the left is a map of p-values in time (x-axis) over all 128 electrode locations (y-axis). Scalp locations are ordered from posterior (occipital scalp) to the anterior (fronto- polar) scalp and from left to right for each individual scalp region, respectively. The legend on the right contains the color code for the p-values.

Figure 15: Comparison of the evoked responses of the multisensory perceptual ambiguity (lowvol) conditions.

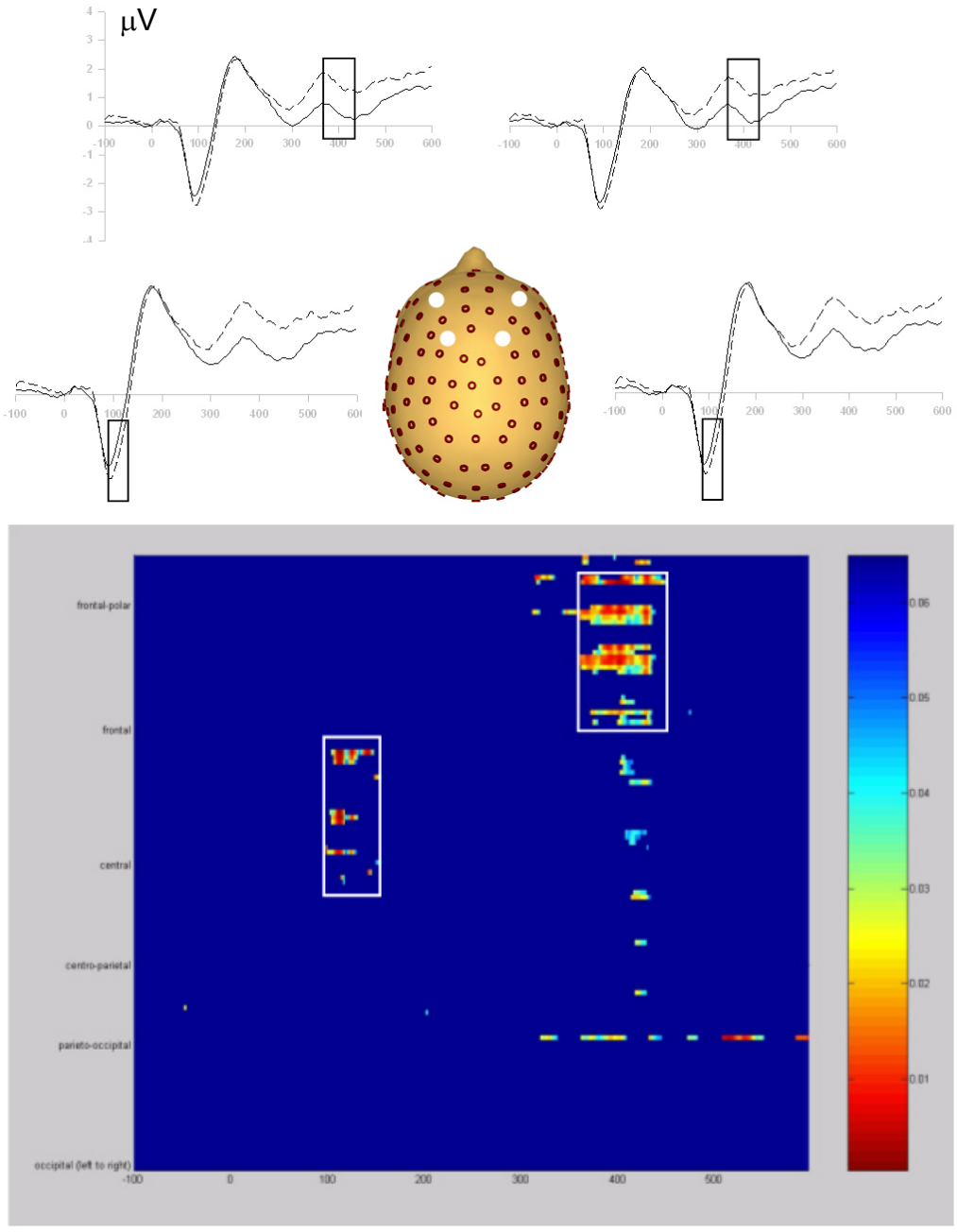
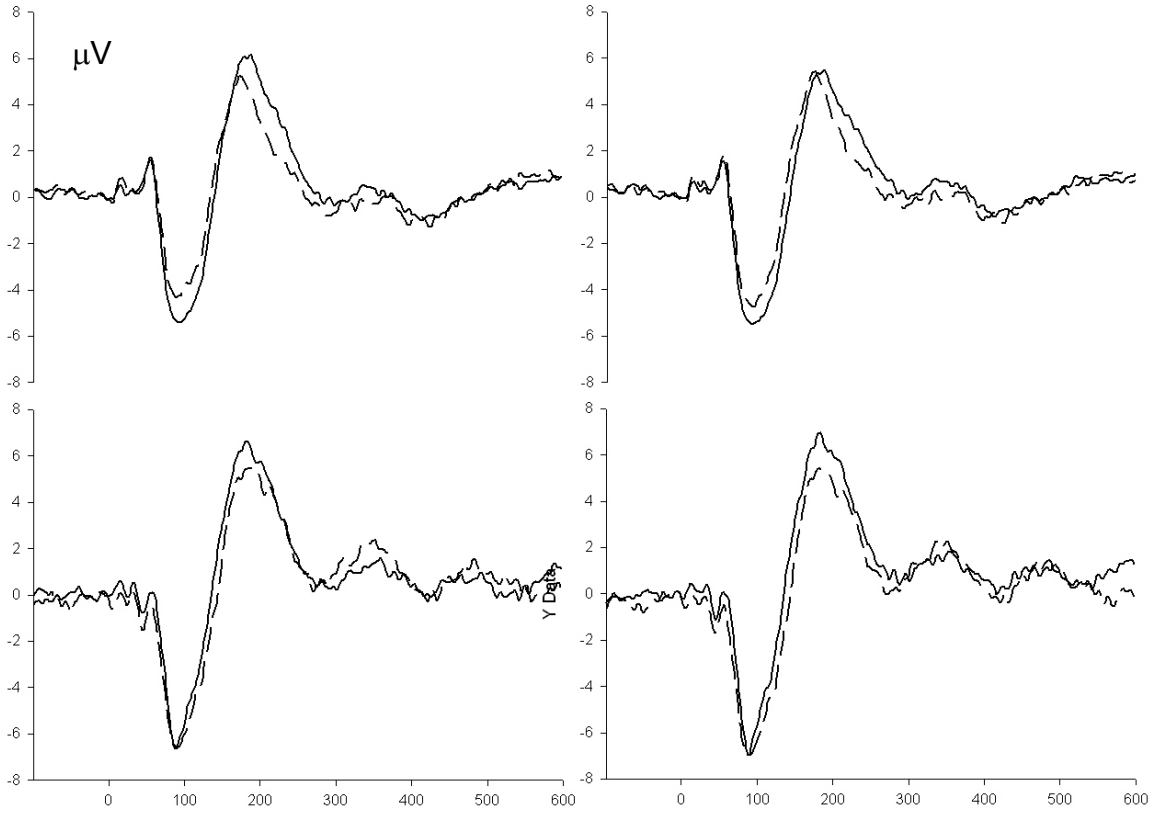


Figure 16.

Top: ERP's of the A_{alone} and $AV_{misaligned}$ (dashed) conditions when A_{alone} (solid) was presented first (N = 8). Bottom: ERP's of the same conditions when the $AV_{misaligned}$ condition was presented first (N = 5).

Figure 16: Comparison of waveforms by order of presentation



References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol, 14*(3), 257-262.

Allen, P. G., & Kohlers, P. A. (1981). Sensory specificity of apparent motion. *Journal of Experimental Psychology: Human Perception and Performace, 7,* 1318-1326.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends Cogn Sci, 4*(7), 267-278.

Bense, S., Stephan, T., Yousry, T. A., Brandt, T., & Dieterich, M. (2001, Feb). Multisensory cortical signal increases and decreases during vestibular galvanic stimulation (fMRI). *J Neurophysiol, 85*(2), 886-899.

Berman, R. A., & Colby, C. L. (2002). Auditory and visual attention modulate motion processing in area MT+. *Brain Res Cogn Brain Res, 14*(1), 64-74.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *Eur J Neurosci, 20*(8), 2225-2234.

Bullier, J., Hupé, J. M, James, A. C., & Girard, P. (2001). The role of feedback connections in shaping the responses of visual cortical neurons. *Prog Brain Res, 134,* 193-204.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nat Neurosci, 6*(2), 190-195.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. J Cogn Neurosci, 1, 151, 57-70.

Calvert, G. A., & Lewis, J. W. (2004). Hemodynamic studies of audiovisual interactions. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Campbell, R., Dodd, B., & Burnham, D. (1998). Hearing by eye: Part II. The Psychology of speech-reading and audiovisual speech (pp. 85-108). East Sussex, England: Psychology Press.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence for multimodal integration in the primate striate cortex. *J Neuroscience, 22,* 5749-5759.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Cereb Cortex, 12*(10), 1031-1039.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002, June). Early auditory-visual interactions in human cortex during nonredundant target identification. *Brain Res Cogn Brain Res, 14*(1), 20-30.

Fort, A. & Giard, M. H. (2004). Multiple electrophysiological mechanisms of audiovisual integration in human perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes.* Cambridge, MA: MIT Press.

Foxe, J. J., Morocz, I. A., Higgins, B. A., Murray, M. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing. *Brain Res Cogn Brain Res, 10,* 77-83.

Foxe, J. J., & Schroeder, C. E. (2005). Multisensory contributions to low-level, “unisensory” processing. *Curr Opin Neurobiol, 15*(4), 454-458.

Giard, M., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *J Cogn Neurosci, 11,* 437-490.

Gilbert, G. M. (1939). Dynamic psychophysics and the phi phenomenon. *Archives of Psychology, 237,* 5-43.

Guinan Jr., J. J. (1996). Physiology of olivocochlear efferents. In P. Dallos, A. Popper, & R. Fay (Eds.), *The cochlea (pp. 435-502)*. Vol. 8:, Springer, New York.

Hall, K. R. L., Earle, A. E., & Crookes, T. G. (1952). A pendulum phenomenon in the visual perception of apparent movement. *Quarterly Journal of Experimental Psychology, 4,* 109-120.

Hall, K. R. L., & Earle, A. E. (1954). A further study of the pendulum phenomenon. *Quarterly Journal of Experimental Psychology, 6,* 112-124.

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. London: Wiley.

Kawashima, R., O'Sullivan, B. T., & Roland, P. E. (1995, June 20). Positron-emission tomography studies of cross-modality inhibition in selective attentional tasks: Closing the "mind's eye." *Proc Natl Acad Sci U S A, 92*(13), 5969-5972.

King, A. J., & Calvert, G. A. (2001). Multisensory integration: Perceptual grouping by eye and ear. *Current Biology, 11,* 322-325.

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. *Nature, 416*(6877), 172-174.

Leavitt, V. M., Molholm, S., Ritter, W., Shpaner, M., & Foxe, J. J. (2007). Auditory processing in schizophrenia during the middle latency period (10-50 ms): High-density electrical mapping and source analysis reveal subcortical antecedents to early cortical deficits. *J Psychiatry Neurosci, 32*(5), 339-353.

Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y. F., Field, A. S., & Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. *J Cogn Neurosci, 14*(3), 420-429.

Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000, September). A comparison of visual and auditory motion processing in human cerebral cortex. *Cereb Cortex, 10*(9), 873-888.

Maass, H. (1938). Ueber den einfluss akustischer rhythmen auf optische bewegungsgestaltungen. In Sander, F.: Ganzheit und Gestalt. Psychol. Untersuch. VIII. Archiv fuer die Gesamte Psychologie, 100, 424-464.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science, 289,* 1206-1208.

Manabe, K., & Riquimaroux, H. (2000). Sound controls velocity perception of visual apparent motion. *Journal of the Acoustical Society of Japan, 21,* 171-174.

Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2000, April). Intermodal selective attention in monkeys. II: Physiological mechanisms of modulation. *Cereb Cortex, 10*(4), 359-370.

Metzger, W. (1934). Beobachtungen ueber phenomenale Identitaet. *Psychologische Forschung, 19,* 1-49.

Meyer, G. F., & Wuerger, S. M. (2001, August 8). Cross-modal integration of auditory and visual motion signals. *Neuroreport, 12*(11), 2557-2560.

Michotte, A. (1963). *The perception of causality*. London: Methuen (English translation of Michotte, 1954).

Miki, K., Watanabe, S., Kakigi, R., & Puce, A. (2004, July). Magnetoencephalographic study of occipitotemporal activity elicited by viewing mouth movements. *Clin Neurophysiol, 115*(7), 1559-1574.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Brain Res Cogn Brain Res, 14,* 115-128.

Murray, M. M., Wylie, G. R., Higgins, B. A., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). The spatiotemporal dynamics of illusory contour processing: Combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging. *J Neurosci, 22*(12), 5055-5073.

Näätänen, R., & Teder, W. (1991). Attention effects on the auditory event-related potential. *Acta Otolaryngol Suppl, 491,* 161-166.

Ohmura, H. (1987). Intersensory influences on the perception of apparent motion. *Japanese Psychological Research, 29,* 1-19.

Picton, T. W., Hillyard, S. A., Krausz, H. I., & Galambos, R. (1974). Human auditory evoked potentials: Evaluation of components. *Electroencephalogr Clin Neurophysiol, 36*(2), 179-190.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Phil Trans R Soc, 358,* 435-445.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron, 28,* 617-625.

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., Howard, M. A., & Brugge, J. F. (2007, March). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience, 145*(1), 162-184.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkeys. *Int Journal of Psycholophysiology, 50,* 19-26.

Ross, L. A., Molholm, S., Lewkowicz, D., Javitt, D. C., & Foxe, J. J. (2004, April 18). *"Swoosh": How the quality of sounds affects the perception of a visual motion illusion*. Poster session presented at the annual meeting of the Cognitive Neuroscience Society, San Francisco, USA.

Sanabria, D., Spence, C., & Soto-Faraco, S. (2006). Perceptual and decisional contributions to audiovisual interactions in the perception of apparent motion: A signal detection study. *Cognition, 102*(2), 299-310.

Sanabria, D., Correa, A., Lupiáñez, J., & Spence, C. (2004, August). Bouncing or streaming? Exploring the influence of auditory cues on the interpretation of ambiguous visual motion. *Exp Brain Res, 157*(4), 537-541.

Sanabria, D., Lupiáñez, J., & Spence, C. (2007). Auditory motion affects visual motion perception in a speeded discrimination task. *Exp Brain Res, 178*(3), 415-421.

Schroeger, E., & Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology, 35*, 755-759.

Sekuler, R. A., Sekuler, B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308.

Schroeder, C., & Foxe, J. J. (2005). The case for feedforward multisensory convergence during early cortical processing. *NeuroReport, 16*(5), 419-423.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature, 408,* 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14,* 147-152.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Curr Opin Neurobiol,11*(4), 505-509.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C, & Kingstone, A. (2002, June). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Brain Res Cogn Brain Res, 14*(1), 139-146.

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003) Multisensory contributions to the perception of motion. *Neuropsychologia, 41*(13), 1847-1862.

Soto-Faraco, S., & Kingstone, A. (2004). Multisensory integration of dynamic information. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Soto-Faraco, S., Spence, C., & Kingstone, A. (2005, January-February). Assessing automaticity in the audiovisual integration of motion. *Acta Psychol (Amst), 118*(1-2), 71-92.

Stein, B. E., Jiang, W., & Stanford, T. R. (2004). Multisensory integration in single neurons in the midbrain. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Brain Res Cogn Brain Res, 14*(1), 106-114.

Watanabe, K., & Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception, 27*(9), 1041-1054.

Watanabe, K. (2001). *Crossmodal interactions in humans*. (PhD thesis). Pasadena: California Institute of Technology.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005, January 25) Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A, 102*(4), 1181-1186.

Vroomen, J., & De Gelder, B. (2003, July). Visual motion influences the contingent auditory motion aftereffect. *Psychol Sci, 14*(4), 357-361.

Welch, R., & Warren, D. (1986). Intersensory interactions. In K. Boff, L. Kaufmann, & J. Thomas (Eds.), *Handbook of perception and human performance:* Vol. 1 Sensory Processes and Human Performance. New York: Wiley.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003, October). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex, 13*(10), 1034-1043.

Wuerger, S. M., Hofbauer, M., & Meyer, G. F. (2003). The integration of auditory and visual motion signals. *Percept Psychophys, 65*(8), 1188-1196.

Zietz, K., & Werner, H. (1927). Uber die Dynamische Struktur der Bewegung. Werner's Studien ueber Strukturgesetze, VIII. *Zeitschrif fuer Psychologie, 105,* 226-249.

CHAPTER V

OUTLOOK

1. Introduction

The findings of the experiments described in this thesis, in particular the experiments on speech perception, inspired a number of new projects that are presently at various stages of development. In this chapter I will summarize a selection of these projects and will provide, if present, pilot data. This overview is meant to be a discussion that seeks to provide a meaningful context for the work described in this thesis by linking selected open questions discussed in the previous chapters to ongoing research that attempts to answer them.

One important question that will be elaborated on in the early part of this chapter is why a disparity exists between integration processes on a subcortical level and AV-speech recognition. In the first study conducted in collaboration with the Bioengineering Department at the City College of the City University of New York (CCNY), we asked whether one cause for this disagreement might be the fact that the nature of the speech stimulus is much more complex than the stimuli used in studies that found inverse effectiveness. Dr. Wej Ji Ma from the University of Rochester developed a mathematical model that predicts AV-performance that takes the complexity of the underlying stimuli into account. Further, we asked whether timing cues delivered in the visual modality by the articulation of the speaker are a significant cue for the recovery of the auditory speech signal.

A second focus of ongoing research concerns the timing and locus of integration of AV-speech signals in the brain. As discussed in the general

introduction of this thesis, research on the McGurk-effect (McGurk and MacDonald, 1976) has provided the major empirical support for common format models of AV-speech perception. We have also seen how electrophysiological approaches, due to their ability to index brain processes with high temporal acuity, are able to constrain existing models of AV- speech perception. In this ongoing project we assess multisensory processes directly from the brain surface by recording from implanted electrodes in patients with epilepsy, before surgery.

A third focus of new research is currently being conducted by the Child Research Unit at CCNY and was motivated by the idea that there may be parallels between certain information processing deficits in schizophrenia and autism (e.g. Kelemen et al., 2005). In this line of experiments we investigate the integrity of processes relying on the integration of information from the speakers face in speech perception and speaker identification.

2. Audiovisual Speech Recognition is Consistent with Bayesian Optimal Cue Combination

Wei Ji Ma[1]*, Xiang Zhou[2]*, Lars A. Ross[3,4], John J. Foxe[3,4,5], Lucas C. Parra[2]

[1] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14620, USA
[2] Department of Biomedical Engineering, The City College of New York, New York, NY 10031, USA
[3] Program in Cognitive Neuroscience, Department of Psychology, The City College of New York, New York, NY 10031, USA
[4] The Cognitive Neuroscience Laboratory, Nathan S. Kline Institute for Psychiatric Research, Program in Cognitive Neuroscience and Schizophrenia, Orangeburg, NY 10962, USA
[5] Program in Neuropsychology, Department of Psychology, Queens College of the City University of New York, Flushing, NY 11367, USA
* These authors contributed equally to this work.

This ongoing collaborative project with Dr. Lucas Parra and his student Xiang Zhou (Department of Bioengineering at CCNY) and Dr. Wei Ji Ma at the Department of Brain and Cognitive Sciences at the University of Rochester, NY spans a number of experiments that are at various stages of progress. The first and second studies within this project are completed and have been submitted for publication. The first experiment investigated the role of temporal aspects of visual information in AV- speech in the absence of phonemic cues. As mentioned in the general introduction, temporal aspects of the visual stimulus may play a role in early, low level modulation of the acoustic signal by the visual stimulus.

The second study involved the proposal of a mathematical model that predicts AV- performance in word recognition. As discussed in chapters II and III, one of the questions that remain is why we failed to find inverse effectiveness in AV-word recognition. This model reconciles the “principle of inverse effectiveness” with the findings of AV- benefit in word recognition by showing that

the locus of maximal gain moves to higher SNRs as the number of features of the stimuli increases.

2.1 The role of timing of the visual signal in the recognition of AV speech

2.1.1 Brief summary of the methods

The first study used a methodology similar the one described in chapter II. In addition, a second AV-condition was added in order to investigate the contribution of timing information derived from the visual cue. To achieve this, we used a video-synthesis program that could generate a face that was similar in appearance to a given natural moving face (Lehn-Schioler, 2005). The method uses features that were extracted from the audio-signal (acoustic waveforms of word utterances) to generate the visual articulations of the mouth, eyes, brows, and outline of the face. Using this material, realistic video frames of a speaker were generated. Critically, only the power of the audio- signal (in 40ms time frames) was used to generate the articulation of the face. Hence, the V-stimulus exclusively represented visual information associated with the overall intensity fluctuation of the audio-signal in time. Therefore, the animated face did not reflect any information associated with the spectral content of the original speech signal, and was therefore unlikely to convey any phonemic information.

2.1.2 Brief summary of results

Figure 1 summarizes the percentages of correct identification of the A-, AV- and modified AV-condition with only temporal information as well as the resulting AV-gain curves.

The gain in performance due to exposure to visual articulation confirms our findings described in Chapters II and III, i.e. the largest gain is not found at the lowest SNRs as the principle of inverse effectiveness would predict but instead at the intermediate SNR (-12dB). The modified AV-stimulus also produced significant but much smaller gain. Interestingly, contrary to what would be expected if the principle of inverse effectiveness applied, the AV-gain increases with increasing SNR.

2.1.3 Brief discussion

It has been shown that seeing a speaker's head motion alone can improve recognition of speech (Thomas and Jordan, 2004). Schwarz et al., (2004) demonstrated that syllable identification can be improved when different syllables are paired with the same visual stimulus. These findings suggest that synchronous visual information alone can enhance the speech perception. A possible explanation for this effect was suggested by Grant et al. (2001). The synchronous visual signal provides information about whether an auditory signal belongs to the speech signal rather than to the background noise. In other words, the visual signal facilitates the grouping of auditory information with a speech signal rather than with the noise. This co-modulation is known to work in the

auditory modality (Buus, 1985) and may also work across modalities as a so-called bi-modal coherence masking protection (Grant, 2001). Apparently, this co-modulation requires that a certain minimum of auditory information is available. The maximal gain due to this modulation is at higher SNRs than with regular visual articulation. Another important observation is that the contribution of timing information is relatively small in comparison to the benefit of regular speech articulation. This illustrates nicely that the phonetic content in cues from natural visual articulation are significant and constitute the main contribution from the visual signal (especially place of articulation). The relative contribution of visual timing information in comparison to visual phonetic information has never, to my knowledge, been quantified before. This brings us back to the earlier discussion about models of AV-speech perception in the general introduction. There, I speculated that the temporal (and spatial) coincidence of A and V signals serves as a “tagging” mechanism signaling that stimuli in different modalities belong to the same external event. This information is available through the earliest convergence of the signals in different modalities on subcortical and early cortical levels. However, these data suggest that temporal coincidence information from the visual stimulus plays only a limited role in the recovery of the auditory stimulus in speech recognition. As outlined in the introduction, higher-level phonetic information is likely to require more elaborate analysis by higher-order motion processing networks in the STS/G.

2.2 Bayesian optimality in speech recognition and inverse effectiveness

This Bayesian model was developed by Dr. Wei Ji Ma from the University of Rochester and was motivated by the findings of Ross et al. (2007) and the newest replication shown above. While a detailed description of this model is beyond the scope of this thesis, a short introduction will be given here.

Based on the simple probabilistic principle originally formulated by Reverend Thomas Bayes in 1761, we can predict the likelihood of the detection of an audiovisual target on the basis of the probabilities of the detection of the unisensory targets (A and V) and overall likelihood of detecting the stimulus (e.g. the detectability of a word dependent on its frequency of use). Empirical data that follow the prediction of this model are called “Bayes optimal.”

The present model is an extension of this classic model. It is based on the assumption that the representation (source) of a word in the speaker and listener is amodal in nature. The listener derives the identity of the word from the signals in the auditory and visual modality that are assumed to be (and independently so) embedded in noise. The probability of detecting each unisensory stimulus is represented by a probability distribution characterized by the a priori likelihood of detecting the stimulus in the noise it is embedded in. In the case where the stimulus is delivered in both modalities, the likelihood of detecting the stimulus is the product of both probability distributions. As detailed in Figure 2 (a), this new, bisensory distribution yields a higher likelihood of detection.

The critical property of this model is that it makes different predictions depending on the dimensionality of the source stimulus (or word that is to be detected). Dimensions are auditory and visual features of the AV stimulus, such as the specific phonemic constellation of the word, its articulatory features, and its frequency of usage. In short, the dimensions represent all features that determine the probability of its detection. The specific features of the words themselves remain unspecified (although feature systems for words have been suggested, e.g., Chomsky and Halle, 1968; Wang and Bilger, 1973; Wang, Reed and Bilger, 1978), but the influence of the number of features on the detectability of the word can be modeled within this proposed Bayesian framework. As can be seen in graph a) of Figure 2, the gain clearly shows an inverse effectiveness pattern.

This changes when performance in the A- and AV- conditions is modeled on the basis of a source-stimulus with more than one dimension (as is assumed to be the case in complex speech signals such as words). Figure 2. shows that the prediction for maximal AV-gain shifts in relation to of the number of assumed dimensions. This reconciles the rather ubiquitous findings of inverse effectiveness at neuronal levels (for reviews see Stein and Meredith, 1993; Stein et al., 2004) and behavior in cats (Stein, et al., 1988) and humans (Corneil et al., 2002; Diederich and Colonius, 2004; Bolognini, et al., 2005; Rach and Diederich, 2006; Gillmeister and Eimer, 2007; Serino, et al., 2007) and the findings of our group. The predictions of this model suggest that multisensory gain patterns may be dependent on the complexity of the stimuli that are integrated. When stimuli

are simple (unidimensional) then the gain from concurrent stimulation follows inverse effectiveness as has been shown in studies where the underlying stimulus feature is unidimensional (Alais and Burr, 2004; Rowland et al., 2007). In these examples, participants were asked to estimate the position of an A-, V-, or AV- event on a horizontal line with varying degrees of visual and auditory reliability. Gain in speech recognition has also been found to conform to inverse effectiveness when simple speech stimuli (syllables) had to be categorized, as in Massaro's paradigms, (Massaro, 1987).

3. An intracranial investigation of the mismatch negativity evoked by the McGurk effect

In the introduction to this thesis, we saw how the electrophysiological investigation of AV-integration of speech, due its high temporal acuity, can be a particularly useful tool to constrain existing models of AV-speech perception. In this study we extended the investigations of the mismatch negativity evoked by the McGurk illusion published by our group (Saint-Amour et al., 2007) by using intracranial recordings. Although this study is in progress, initial data have been collected and the evoked responses from an exemplary subject will be presented here.

As shown in the introduction to this thesis, the incongruent visual articulation of the speaker can evoke a mismatch negativity (MMN), a physiological component of the auditory ERP that signals change in the auditory

modality. If the change in activity evoked by the McGurk stimulus is similar to the activity evoked by the real acoustic change in time and topography, we can consider this as strong evidence for a modulation of auditory activity by visual articulation at early stages of processing. It would contradict the notion that auditory and visual information is combined at a late stage in multisensory cortices with a preservation of sensory specific processing (Berstein et al., 2004).

3.1 Brief description of methods

McGurk-MMN was generated using an oddball paradigm in two conditions: auditory alone (/ba/ as “standard” and /va/ as “deviant” and vice versa) and audiovisual (congruent audio /ba/ and visual /ba/ and audio /va/ and visual /va/ as “standard” and incongruent audio /ba/ and visual /va/ as “deviant” (and vice versa)). Hence, in the critical audiovisual condition, the phoneme /ba/ (or /va/) was presented auditorily on every trial but was perceived as /va/ (or /ba/) when presented with an incongruent visual articulation. Deviants were presented with 15% probability. Stimulus duration was 280ms in all conditions and stimuli were presented with an inter-trial interval of 720ms. The waveforms were generated by averaging over all auditory and audiovisual standards and deviants.

3.2 Brief description of results

The data of one subject are presented here. As can be seen in Figure 3, (bottom panel) the acoustic deviant evoked a proper MMN. The difference in waveforms onsets around 150ms and peaks at about 340ms. The incongruent visual

stimulus also evoked a MMN, although it onsets somewhat later at 175ms and is not associated with the same distinct increase in amplitude. In addition, when comparing the waveforms to the standards at the N2 component, it can be seen that the additional visual stimulation resulted in an approximately 50% modulation in amplitude. This activity was maximal over the primary auditory cortex.

In this experiment we have replicated the finding of a MMN to an incongruent visual deviant in the absence of an acoustic change. Timing and locus of this effect suggest that visual articulation modulates the acoustic processing of the auditory stimulus. As discussed in the introduction to this thesis, it is likely that the phonemic information from the visual stimulus arrives at auditory cortices from "higher order" multisensory cortices because the visual motion stimulus has to be subject to fairly elaborate analysis in order to deliver the phonemic information that gives rise to the McGurk effect.

4. Research in autism

It has repeatedly been suggested that there are parallels between certain information processing deficits in schizophrenia and autism (e.g. Kelemen et al., 2005). This has led to the extension of my (our) work on MSI in Schizophrenia to the investigation of MS processes in children with autism. I enjoy the formidable opportunity to participate in the Children Research Unit (CRU) at the City College of New York, led by Dr.'s John Foxe, Hillary Gomes and Sophie Molholm, that

specializes in developmental disabilities, in particular autism. The following is a description of a research plan that I have developed with the CRU. The study of AV- speech processing is already implemented and the first sample of children have participated in the experiment.

4.1 Research plan

Specific aims:

This research will focus on whether deficits in sensory integration might be a contributing factor to higher-order cognitive and perceptual deficits typical of autism. We specifically focus on processes that are of high importance for social interaction since such deficits are the hallmark symptoms of autism.

Aim 1: The first aim is to assess whether deficits in sensory integration might be a contributing factor to deficits in speech processing abilities under noisy environmental conditions, as is commonly seen in individuals with autism. The perception of audiovisual speech is reliant on the integrity of processes underlying the perception of biological motion (BM). There is recent behavioral evidence that this process may be impaired in people with autism (Blake et al., 2003). The understanding of the impairment of BM processing is, in our view, of high importance since it underlies a variety of other higher–order processes that are essential for social interaction.

Aim 2: The second aim is, therefore, to investigate the integrity of neurophysiological mechanisms underlying BM in autism. We will see whether electrophysiological correlates of BM that have been identified in the past

(Jokisch et al., 2005; and in our laboratory: see above) differ between typically developing children and those with autism.

Aim 3: The third aim is to assess whether children with autism develop multisensory representations of people. The perception of the identity of people in our environment is fundamental for social interaction and to a large extent a multisensory process since we (and specifically developing children) perceive visual attributes of people in unison with their voices. While there is an extensive documentation of impaired face processing in autism, it remains to be shown whether autistic children respond to combined faces and voices in their social environment in the same way as typically developing children. Again, our goal will be to investigate whether the neurophysiological correlates of multisensory speaker processing are impacted in autism.

4.2 Brief description of general methods

Subjects:

In each experiment, 20 typically developing children and 20 children with autism will serve as participants. Where possible, we will recruit participants that have participated in the previously conducted experiments. Children's ages will vary between 4 and 15 years. In each experiment the children in both groups will be carefully age-matched. Since there is an often observed developmental change in evoked responses (see Taylor et al., 2004), we will break up our subject pool into age bands of 4-6, 7-10, and 11-15 years of age. This is also important in the

first experiment because there may be general performance differences due to age-related differences in word lexicon- size.

Electrophysiological recordings:

Continuous electroencephalographic (EEG) data, digitized at 512 Hz, will be acquired through the ActiveTwo Biosemi electrode system from 168 scalp electrodes. A major advantage of this high-impedance recording system is that scalp abrasion is avoided and electrode application is quick and causes no, or minor, discomfort. We have been successfully using this system for control populations, clinical populations and recordings developmental disorders and ageing for more than two years now. A repeated measured analysis of variance (RM-ANOVA) with an alpha level of 0.05 will be used to test for statistical differences.

4.3 Experiment 1: Audiovisual speech perception in noise

4.3.1 Background:

Children with autism have well-characterized deficits in speech perception under noisy environmental conditions (Alcantara et al., 2004). One prediction is that deficits in basic auditory-visual integration in autism might be a significant precursor to higher-order deficits in speech recognition that are found in everyday environments often full of distracting noise sources. I showed earlier that that audiovisual integration in background noise is impaired in patients with schizophrenia. The hypothesis that schizophrenia and autism share abnormalities has been formulated in the past (e.g. Kelemen et al., 2005).

Evidence that gives rise to this idea is the fact that processes underlying multisensory integration in speech processing has been shown to be located in the superior temporal gyrus/sulcus (STS/G) of the human cortex, an area that was shown to have abnormalities in both schizophrenia and autism (Shenton, 2001).

4.3.2 Methods:

Stimuli, design, and procedure:

A slightly modified version of the methods described in the experiments in chapter 2 will be used. For children, a smaller stimulus set of 300 selected, high frequency words that are likely to be part of the child's lexicon is used. The stimulus delivery is now on a portable laptop computer and the auditory stimuli are delivered through headphones.

4.4 Experiment 2: biological motion processing

4.4.1 Background:

Recent evidence suggests that biological motion processing is impaired in children with autism (Blake et al., 2003). It has been suggested that this deficit may be related to impaired social skills and communication, characteristics of autism. It has also been found that performance in the perception of biological motion is impaired in patients with schizophrenia (Kim et al., 2005) which led the authors to suggest that the same neural substrate, namely the STS/G, may be

compromised in schizophrenia and autism,. Recent functional magnetic resonance studies support this view (Boddaert et al., 2004; Shenton 2001; Zilbovicius et al., 2006). For us, biological motion perception is of particular interest here this is because the perception of visual articulation is essentially a biological motion process and therefore underlies audiovisual integration in speech perception as was tested in Experiment 1. In this experiment we will test the neurophysiological correlates of this deficit. In previous work (see Appendix) we identified the componentry that is associated with the perception of BM in healthy adults. We found early modulations (P1, N1) that are likely to be associated with the well-known “attentional capture” qualities of BM (Blake et al., 2003), whereas later activity is likely to be associated with the perception of the BM process. In Experiment 2 we will test whether children with autism show abnormalities in these components. Differences between children with autism and typically developing children may index impairments that are at the source of higher- order processes like the attentional orientation to socially relevant stimuli, audiovisual integration, communication, etc.

4.4.2 Methods:

Stimuli, design and procedure:

Participants will be presented with 10 animated point-light displays (1 sec. duration/ 29 frames per sec.) of an adult engaged in various activities (walking, throwing, jumping, etc.) and phase-scrambled sequences will be generated from the same animations. The scrambled motion displays will be perturbed in terms

of the hierarchical, pendular motions characteristic of biological motion, while maintaining all other aspects of motion such as velocity, coherence, etc. For all animations, the black dots will appear against a white background. In order to maintain attentional deployment on the clips on the screen, one of the points will briefly turn red. Participants will be instructed to press a mouse-button whenever this target occurs. In a second condition participants have a simple task which they are asked to indicate with a mouse button press, after the presentation of each clip whether the motion sequence contained “human-like” activity (“forced choice”). Participants will be presented with seven 5-minute blocks, each containing 100 clips of which 50 will be biological motion and 50 will be scrambled motion. Each inter-stimulus interval will be either 500, 1000 or 1500 ms long and will vary pseudo-randomly.

4.5 Experiment 3: multisensory speaker processing in autism

4.5.1 Background:

The ability to use facial information, such as gaze monitoring, is considered one of the critical cues in the early diagnosis of autism. It has been suggested that the failure to process faces in a normal manner might be one of the earliest measurable autism symptoms (Dawson et al., 2005). Facial information is an essential prerequisite for the identification of peers in our social environment and is necessary for the formation of social bonds, social interest and adequate social interaction. Person identification (PI) in general is a special form of object classification. It consists of the binding of concurrent information about attributes

such as facial physiognomy, voice, body, and name with a specific person (Amedi et al., 2005) and is therefore, in essence, a multisensory process. In the environment of the developing child, visual information (especially faces) does not occur in isolation and is usually accompanied by the voice of the speaker. The voice is an important social cue and directs attention to the speaker.

While it has been shown extensively that several aspects of human face perception (for a selection of reviews see Behrmann et al., 2006; Dawson et al., 2005; Grelotti et al., 2002; Sasson, 2006), including the recognition of familiar faces and voice recognition (Gervais et al., 2004), are impaired in autism, no study that investigates multisensory face/voice perception in autism exists to date. With this experiment we would like to eliminate this shortcoming by investigating multisensory aspects of face/voice processing using high-density EEG and to gain insight into the underlying neurophysiological processes. Since in a normal environment faces and voices appear together, multisensory stimulation will provide more ecologically valid means to assess speaker processing. The electrophysiology of face processing over the course of childhood has been extensively studied (see Taylor et al., 2004 for a review) and electrophysiological abnormalities have already been found in autism (Dawson et al., 2003, 2005a, 2005b; Kylliainen et al., 2006; McPartland, 2004; Valdizan, 2005). These findings assist us in making a relatively good source for predictions about possible abnormalities in autism.

The N170 component of the visual ERP preferentially activates to faces (Bentin et al., 1996; Eimer, 1998, 2000a, 2000b; George et al., 1996 and peaks

around 170 ms post stimulus onset over the posterior temporal lobe with a generator source in the fusiform face area (FFA). Facial inversion alters both latency and amplitude of the N170 (Eimer, 2000b). Whether the N170 is affected by the familiarity of the faces remains controversial (see Bentin et. al., 2000; Eimer, 2000a; and also Marzi & Viggiano, 2007). The N170 has a prolonged period of development (Taylor et al., 2001). The onset is significantly later in younger children and accelerates until adolescence. While in typically developing children the N170 shows a shorter latency to faces than to objects, this effect is missing in children with autism. Children with autism also fail to display a latency slowing for inverted faces (Valdizan, 2005). Two later components are affected by the familiarity of faces, one peaking at 250ms (N250) (Schweinberger et al., 2002) and another peaking 400ms (P400) post-stimulus onset (Eimer, 2000b; Henson, et al., 2003; Mnatsakanian & Tarkka, 2003) in adults and typically developing children (e.g. Carver et al., 2003; deHaan & Nelson, 1999). However, children with autism fail to show differential ERP's to their mother's face versus an unfamiliar face at either component but did show differential ERP's to a favorite versus an unfamiliar toy (Dawson et al., 2002).

In Experiment 3 we will investigate the integrity of neurophysiological processes underlying multisensory speaker perception in autism. In a study conducted in our laboratory (see Molholm et al., 2005) we found multisensory modulations in the latency range of the N1 when pictures of animals were paired with their appropriate vocalizations. The N1 is associated with structural

encoding in the ventral stream and is believed to be a homolog of the N170 component for face processing. We hypothesize that if multisensory integration for highly socially relevant stimuli is impaired in autism, then the N170 component should not be modulated by multisensory input to the same extent as in typically developing children. Further, we will test whether a later component (N250, P400) that has been shown to be related to facial familiarity is modulated by multisensory stimulation. In the above mentioned study conducted by laboratory, we found a possibly homologuous multisensory object recognition effect in the ERP around 270ms and a late, semantic congruency effect around 400ms (N400). We hypothesize that if multisensory integration is impaired for socially relevant stimuli in autism, no (or significantly reduced) multisensory modulations should be found for familiar faces/ voice pairings. In addition, we predict that familiar faces that are presented with incongruent voices will elicit a late, semantic effect around 400ms (N400). This semantic effect requires intact multisensory integration capacities and is predicted to be absent or attenuated if these are impaired in autism.

4.5.2 Methods:

Stimuli, design and procedure:

In this experiment we will use unisensory visual (V), unisensory visual inverted (V_{inv}), unisensory auditory (A), multisensory (AV), and multisensory inverted (AV_{inv}) stimuli. Any of these stimuli can be from an either familiar or unfamiliar speaker. V- stimuli will consist of standardized and carefully controlled digital

color images of 10 still faces (5 male, 5 female) that will be presented on a black background. These will be taken from a previously established database with pictures and voices of speakers of different genders, ages and ethnicities. An additional set of 5 visual stimuli will consist of faces that are very familiar to the subject. These can be parents, other family members, guardians, caretakers or friends of the subject. In order to obtain these images, we will consult with the guardian(s) to determine a set of individuals of the close social environment of the subject. We will invite these individuals to have their pictures taken and their voices recorded. V-stimuli of familiar and unfamiliar faces will be matched in terms of gender, age and ethnicity. A-stimuli will consist of the utterance of the above 10 speakers (5 male, 5 female). Utterances are 20 simple monosyllabic words from the MRC Psycholinguistic database that are likely to be part of the lexicon of a child of our youngest age band (4-5 years). Words will be presented with 50 dB SPL from speakers on each side of the screen in a sound-controlled environment. AV-stimuli will consist of the simultaneous presentation of V- and A-stimuli. The onset of the presentation of the pictures will coincide with sound onset. In the multisensory condition where unfamiliar faces are presented the voicing is mapped to the face in a pseudorandom fashion. In this way the subjects will not be able to build a multisensory representation of the individuals (face-voice pairing) in the stimulus set. In contrast, the face of a familiar person will always be paired with its natural voice. This experiment will have 10 stimulus types with the A-stimulus being a familiar or unfamiliar voice, and V- and AV-stimuli, that can be presented upright or inverted, containing familiar or unfamiliar

faces. Stimuli will be presented in short 10 blocks containing 180 trials each, giving subjects the opportunity for many breaks. Each block will be approximately 4 minutes long, and it is estimated that the total duration of the experiment will be approximately 50 minutes.

Analysis:

The responses to the two unisensory conditions will be “summed” for each condition respectively and all comparisons will be made between the multisensory and summed unisensory responses. First, we will see whether the inversion effect on amplitude or latency of the N170 component is modulated by multisensory stimulation. If this is the case, we will see whether this modulation is also present in children with autism. Second, we will determine whether the expected multisensory modulation of the face-familiarity effect, seen at the P250 and the N400 components, will be present in autism.

Figure 1.

Percentage of correctly identified words in the A- condition (blue) and the AV-condition (green). The red curve shows audiovisual gain (AV-A). In the left panel (a) the visual signal contains regular articulation whereas in the right panel (b) the visual signal contains only timing information.

Figure 1. Performance in the A, AV and modified AV conditions and gain graphs

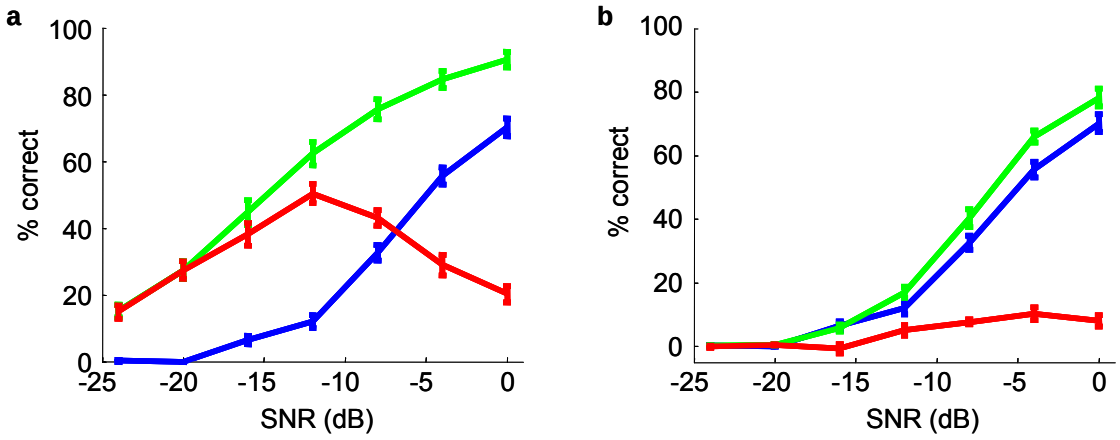


Figure 2.

Modeled correct recognition of an auditory (A: blue) and audiovisual (AV: green) source stimulus and its gain curve (red: AV-A) depending on the dimensionality of the source stimulus (n), the amount of visual stimulation (r_V) and stimulus reliability (r_A). a), AV- cue combination with a one dimensional source-stimulus; b), Predicted performance for a source stimulus with 20 dimensions (n = 20) with little visual information r_V = 1; c), 90 dimensions (n = 90) and little visual information (r_V = 1); d), 20 dimensions and more visual information (r_V = 2); e), 90 dimensions (n = 90) and more visual information (r_V = 2).

Figure 2: AV-cue combination with a source stimulus of different dimensionality

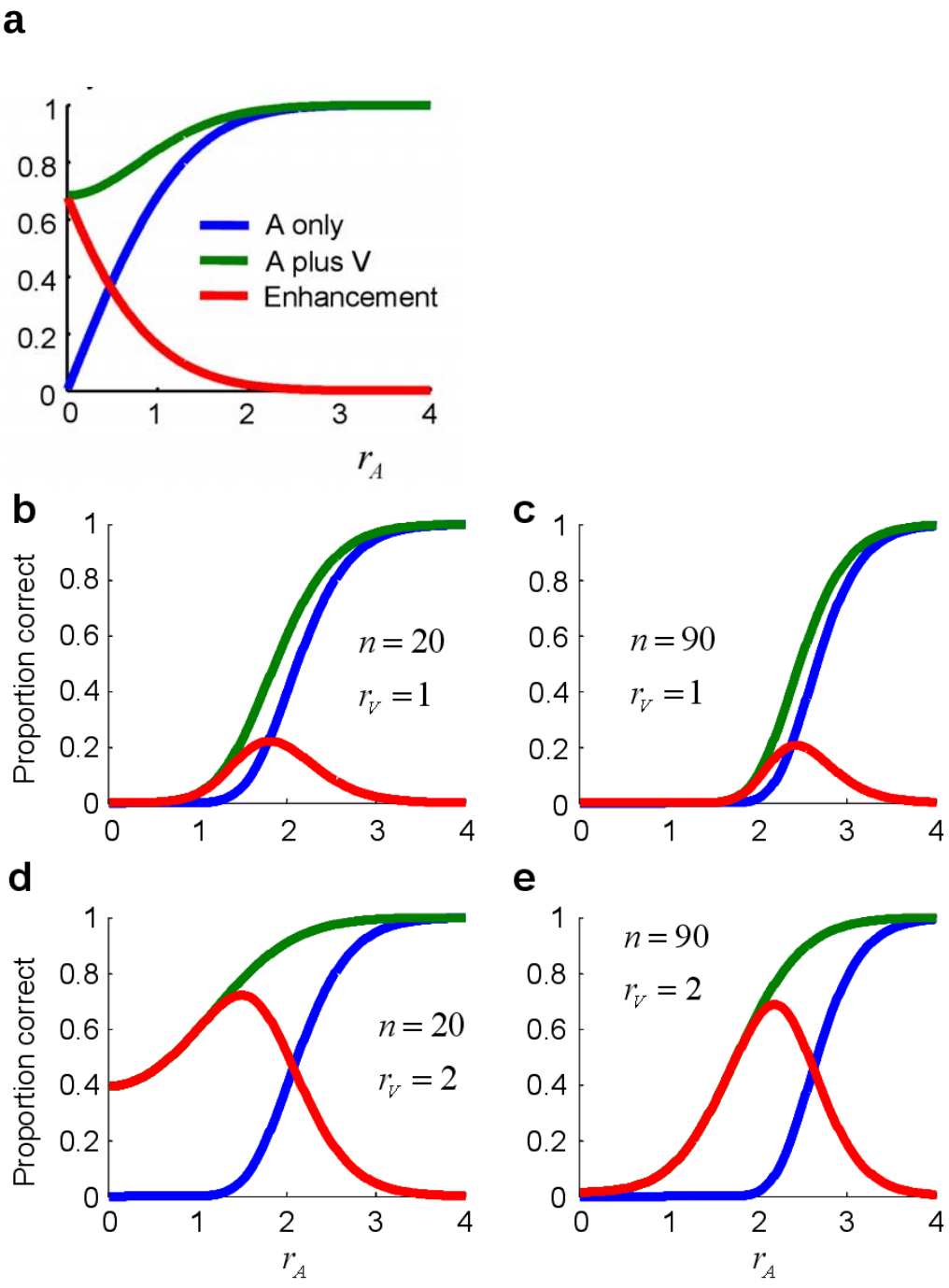
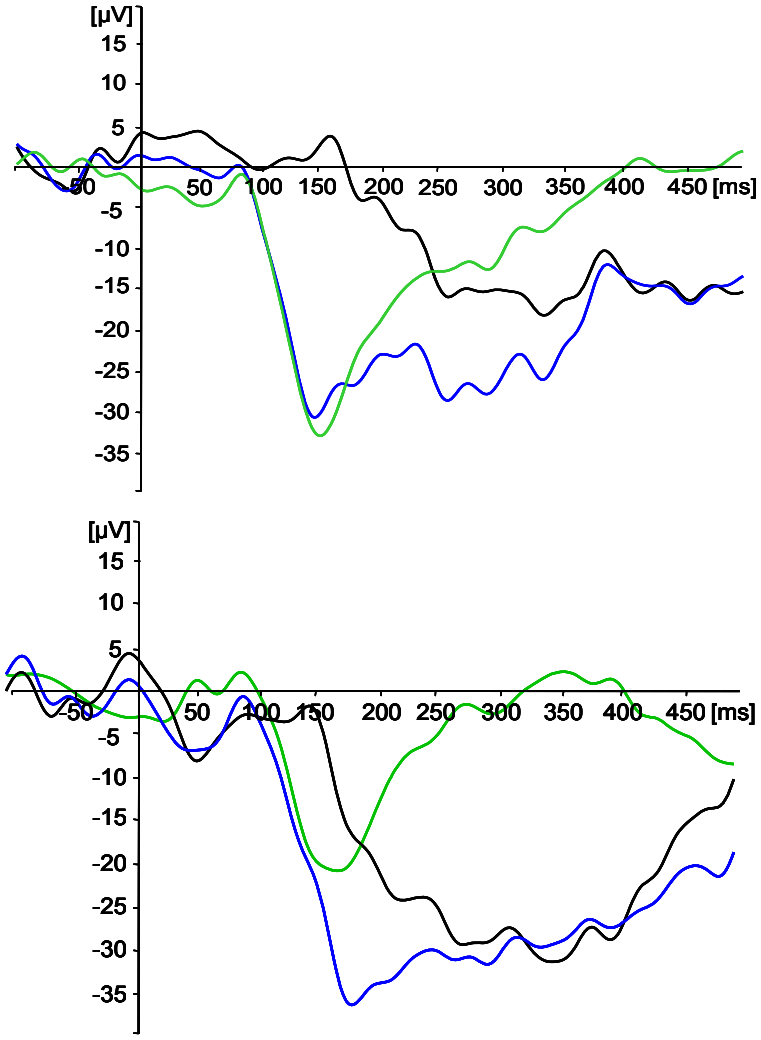


Figure 3.

McGurk (top) and acoustic (bottom) MMN. Green: standard; blue: deviant; black: difference waveform.

Figure 3: Acoustic and McGurk MMN from an intracranial recording in one subject.



References Chapter I and V

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol, 14*(3), 257-62.

Alcantara, J. I., Weisblatt, E. J., Moore B. C., & Bolton P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's Syndrome. *J Child Psychol Psychiatry, 45,* 1107-1114.

Allison, T., Puce, A., McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends Cogn Sci, 4*(7), 267-278.

Amedi, A., Von Kriegstein, K., Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Exp Brain Res, 166,* 559-571.

Anderson, P. A. V., & Schwab, W. E. (1982). Recent advances and model systems in coelenterate neurobiology. *Prog Neurobiol, 19,* 213-236.

Behrmann, M., Thomas, C., & Humphreys, K. (2006). Seeing it differently: Visual processing in autism. *Trends Cogn Sci, 10*(6), 258-264.

Beneveneto, L. A., Fallon, J., Davis, B. J., & Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbitofrontal cortex of the macaque monkey. *Experimental Neurology, 57*(3), 849-872.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8,* 551-565.

Bernstein, L. E., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., & Kaufmann, J. N. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512-528.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex, 17*(9), 2172-2189.

Blake, R., Turner, L. M., Smoski, M. J., Pozdol, S. L., & Stone, W. L. (2003). Visual recognition of biological motion is impaired in children with autism. *Psychol Sci, 14*(2), 151-157.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annu Rev Psychol, 58,* 47-73.

Boddaert , N., Chabane, N., Gervais, H., Good, C. D., Bourgeois, M., Plumet, M. H., Barthelemy, C., Mouren, M. C., Artiges, E., Samson, Y., Brunelle, F., Frackowiak, R. S. J., & Zilbovicius, M. (2004). Superior temporal sulcus anatomical abnormalities in childhood autism: A voxel-based morphometry MRI study. *Neuro Image, 23,* 364-369.

Bodner M., Kroger J., & Fuster, J. M. (1996, August 12). Auditory memory cells in dorsolateral prefrontal cortex. *Neuro Report, 7*(12), 1905-1908.

Bolognini, N., F. Rasi, F., Coccia, M. & Lavadas, E. (2005). Visual localization of sounds. *Neuropsychologia, 43,* 1655-1661.

Breeuwer, M., & Plomp, R. (1985). Speechreading supplemented with formant-frequency information from voiced speech. *Journal of the Acoustical Society of America, 77,* 314-317.

Brodmann, K. (1908). Beitraege zur histologischen Lokalisation der Grosshirnrinde: VI Mitteilung. Die Cortexgliederung des Menschen. *Journal of Psychological Neurology, 10,* 231-246.

Buus, S. (1985, December). Release from masking caused by envelope fluctuations. *J Acoust Soc Am, 78*(6), 1958-1965.

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

Calvert, G. A., Bullmore, E. T., Campbell, R., Williams, S. C., & McGuire, P. K. (1997). Activation of auditory cortex airing silent lipreading. *Science, 276,* 593-596.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999, Aug 20). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*(12), 2619-2623.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *J Cogn Neurosci, 1*(151), 57-70.

Calvert, G. A., & Lewis, J. W. (2004). Hemodynamic studies of Audiovisual Interactions. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes.* Cambridge, MA: MIT Press.

Carver, L. J., Dawson, G., Panagiotides, H., Meltzoff, A. N., McPartland, J., Gray, J., & Munson, J. (2003). Age-related differences in neural correlates of face recognition during the toddler and preschool years. *Dev Psychobiol, 42*(2), 148-159.

Cauller, L. J., & Connors, B. W. (1994). Synaptic physiology of horizontal afferents to layer I in slices of rat SI neocortex. *Journal of Neuroscience, 14,* 751-762.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J. P., Berri, I., & Nespoulos, J. L. (1999). Differential fMRI responses in the left superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *Neuroimage, 9,* 135-144.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115,* 1989–2000.

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

Corneil, B. D., M. Van Wanrooij, Munoz, D.P. & Van Opstal, A. J. (2002). Auditory–visual interactions subserving goal-directed saccades in a complex scene. *J Neurophysiol, 88,* 438-454.

Dawson, G., Munson, J., Estes, A., Osterling, J., McPartland, J., Toth, K., Carver, L., & Abbott, R. (2002). Neurocognitive function and joint attention ability in young children with autism spectrum disorder. *Child Development, 73,* 345-358.

Dawson, G., & Zanolli, K. (2003). Early intervention and brain plasticity in autism. Novartis Found Symp, 251:266-74; discussion 274-80, 281-97.

Dawson, G., Webb, S. J., & McPartland, J. (2005). Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Dev Neuropsychol, 27*(3), 403-424.

Dawson, G., Webb, S. J., Wijsman, E., Schellenberg, G., Estes, A., Munson, J., & Faja, S. (2005). Neurocognitive and electrophysiological evidence of altered face processing in parents of children with autism: Implications for a model of abnormal development of social brain circuitry in autism. *Dev Psychopathol, 17*(3), 679-697.

De Haan, E. H. (1999). A familial factor in the development of face recognition deficits. *J Clin Exp Neuropsychol, 21*(3), 312-315.

De Ceccactty, M. P. (1974). The origin of the integrative systems: A change in view derived from research on coelenterates and sponges. *Perspect Biol Med, 17,* 379-390.

Diederich, A., & Colonius, H. (2004). Modeling the time course of multisensory interaction in manual and saccadic responses. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Eimer, M. (1998). Does the face specific component reflect the activity of a specialized eye processor? *Neuro Report, 9,* 2945-2948.

Eimer, M. (2000a). Effects of face inversion on the structural encoding and recognition of faces: Evidence from event-related brain potentials. *Cognitive Brain Research, 10,* 145-158.

Eimer, M. (2000b). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology, 111,* 694-705.

Eimer, M. (2004). Electrophysiological studies of multisensory attention. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisenory processes*. Cambridge, MA: MIT Press.

Elberling, C., Bak, C., Kofoed, B., Lebech, J., & Saermark, K. (1982). Auditory magnetic from the human cerebral cortex: Location and strength of an equivalent current dipole. *Acta Neurologica Scandinavica, 65,* 553-569.

Falchier, A., Clavagnier, S., Barone, P. & Kennedy, H. (2002). Anatomical evidence for multimodal integration in the primate striate cortex. *J Neuroscience, 22,* 5749-5759.

Fechner, T. (1871). *Vorschule der Aesthetik.* Leipzig: Breitkopf und Hartel.

Felleman, D. J. , Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1-47.

Fowler, C. A. (2004). Speech as a supramodal phenomenon. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisenory processes*. Cambridge, MA: MIT Press.

Foxe, J. J., Morocz, I. A., Higgins, B. A., Murray, M. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing. *Brain Res Cogn Brain Res, 10,* 77-83.

Foxe, J. J., & Simpson, G. V. (2002, January). Flow of activation from V1 to frontal cortex in humans: A framework for defining "early" visual processing. *Exp Brain Res, 142*(1), 139-150.

Foxe, J. J., & Schroeder, C. E. (2005). Multisensory contributions to low-level, “unisensory” processing. *Curr Opin Neurobiol, 15*(4), 454-458.

George, N., Evans, J., Fiori, N., Davidoff, J., & Renault, B. (1996, September). Brain events related to normal and moderately scrambled faces. *Brain Res Cogn Brain Res, 4*(2), 65-76.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., & Zilbovicius, M. (2004). Abnormal cortical voice processing in autism. *Nat Neurosci, 7*(8), 801-802.

Giard, M., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *J Cogn Neurosci, 11,* 437-490.

Gigerenzer, G. (1989). A general algorithm for pattern recognition? *Behavioral and Brain Sciences, 12*(4), 764-765.

Gillmeister, H., & Eimer, M. (2007). Tactile enhancement of auditory detection and perceived loudness. *Brain Res, 1160,* 58-68.

Grant, K. W. (2001, May). The effect of speechreading on masked detection thresholds for filtered speech. *J Acoust Soc Am, 109*(5 Pt 1), 2272-2275.

Graziano, M. S. A., Reiss, L. A., & Gross, C. G. (1999). A neuronal representation of the location of nearby sounds. *Nature, 397*, 428-430.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye: Part II. The psychology of speechreading and audiovisual* speech (pp. 85-108). East Sussex, England: Psychology Press.

Grelotti, D. J., Gauthier, I., & Schultz, R. T. (2002). Social interest and the development of cortical face specialization: What autism teaches us about face processing. *Dev Psychobiol, 3,* 213-225.

Grill-Spector, K., Koutzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41,* 1409-1422.

Gutfreund, Y., & Knudsen, E. I. (2004). Visual instruction of the auditory space map in the midbrain. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Hackett, T. A., Stepnievska, I., & Kaas, J. H. (1998). Subdivisions of the auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology, 394*, 475-495.

Haxby, J. V., Grady, C. L., Horwitz, B., Salerno, J., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Shapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in the human extrastriate cortex. *Proc Natl Acad Sci USA, 88,* 1621-1625.

Haxby, J. V., Ungerleider, L. G., Horwitz, B., Maisog, J. M., Rapoport, S. I., & Grady, C. L. (1996). Face encoding and recognition in the human brain. *Proc Natl Acad Sci USA, 93,* 922-927.

Henson, R. N., Goshen-Gottstein, Y., Ganel, T., Otten, L. J., Quayle, A., & Rugg, M. D. (2003). Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cereb Cortex, 13*(7), 793-805.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed neural system for face perception. *Nature Neurosci, 3,* 80-84.

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. London: Wiley.

Howard, M. A., Volkov, I. O., Mirsky, R., Garell, P. C., Noh, M. D., & Granner, M. (2000). Auditory cortex on the human posterior superior temporal gyrus. *Journal of Comparative Neurology, 416,* 79-92.

Huckins, S. C., Turner, C. W., Doherty, K. A., Fonte, M. M., & Szeverenyi, N. M. (1998). Functional Magnetic Resonance imaging measures of blood flow patterns in the human auditory cortex in response to sound. *Journal of Speech, Language and Hearing Research, 41,* 538-548.

James, W. (1890). *The principles of psychology.* London: Macmillan and Co.

Jemel, B., Mottron, .L, & Dawson, M. (2006). Impaired face processing in autism: Fact or artifact? *J Autism Dev Disord, 36*(1), 91-106.

Jiang, W., Wallace, M. T., Jiang, H., Vaughan, J. W., & Stein, B. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology, 85,* 506-522.

Jiang, W., Jiang, H., & Stein, B. E. (2002). Two corticotectal areas facilitate multisensory orientation behavior. *Journal of Cognitive Neuroscience, 14,* 1240-1255.

Jokisch, D., Daum, I, Suchan, B., & Troje, N. F. (2005). Structural encoding and recognition of biological motion: Evidence from event-related potentials and source analysis. *Behav Brain Res, 157*(2), 195-204.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc Nat Acad Sci USA, 97,* 11793-11799.

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *J Neuroscience, 27*(8), 1824-1835.

Kelemen, O., Erdelyi, R., Pataki, I., Benedek, G., Janka, Z., & Keri, S. (2005). Theory of mind and motion perception in schizophrenia. *Neuropsychology, 19*(4), 494-500.

Kim, J., Doop, M. L., Blake, R., & Park, S. (2005). Impaired visual recognition of biological motion in schizophrenia. *Schizophr Res, 77*(2-3), 299-307.

King, A. J., & Calvert, G. A. (2001). Multisensory integration: Perceptual grouping by eye and ear. *Current Biology, 11*, 322-325.

King, A. J., Doubell, T. P., & Skaliora, I. (2004). Epigenetic factors that align visual and auditory maps in the ferret midbrain. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes.* Cambridge, MA: MIT Press.

Lehn-Schioler, T. (2005). *Making faces: State-space models applied to multi-modal signal processing, Ph.D. Thesis.*

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005, October). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp Brain Res, 166*(3-4), 289-297.

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron,18,*53(2),279-292.

Lauter, J. L., Herscovitch, P., Formby, C., & Raichle, M. E. (1985). Tonotopic organization in human auditory cortex revealed by positron emission topography. *Hearing Research, 20,* 199-205.

Lehmann, S., Murray, M. M. (2005, July). The role of multisensory memories in unisensory object discrimination. *Brain Res Cogn Brain Res, 24*(2), 326-334.

Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex, 10,* 873-888.

Liegois-Chauvel, C., Musolino, A., Badier, J. M., Marquis, P., & Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man: Evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology, 92,* 204-214.

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

Luethke, L. E., Krubitzer, L. A., & Kaas, J. H. (1989). Connections of primary auditory cortex in the new world monkey. *Journal of Comparative Neurology, 285,* 487-513.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science, 289,* 1206-1208.

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

MacSweeney, M., Calvert, G. A., Campbell, R., & McGuire, P. K. (2002). Speechreading circuits in people born deaf. *Neuropsychologia, 40,* 801-807.

MacSweeney, M., Campbell, R., Calvert, G. A., McGuire, P. K., David, A. S., Suckling, J., Andrew, C., Woll, B., & Brammer, M. J. (2001, March 7). Dispersed activation in the left temporal cortex for speechreading in congenitally deaf people. *Proc Biol Sci, 268*(1466), 451-457.

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

Mahling, F. (1926). Das Problem der `audition colorée: Eine historisch-kritische Untersuchung. *Archiv Für die Gesamte Psychologie, 57,* 165-301.

Marks, L. E. (1978). *The unity of the senses: Interrelations among the modalities.* New York: Academic.

Marzi, T., & Viggiano, M. P. (2007, April 13). Interplay between familiarity and orientation in face processing: An ERP study. *Int J Psychophysiol, 65*(3), 182-192.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes.* Cambridge, MA: MIT Press.

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

McPartland, J., Dawson, G., Webb, S. J., Panagiotides, H., & Carver, L. J. (2004). Event-related brain potentials reveal anomalies in temporal processing of faces in autism spectrum disorder. *J Child Psychol Psychiatry, 45*(7), 1235-1245.

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science, 221,* 289-291.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory and somatosensory convergence on cells in the superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56,* 640-662.

Mesulam, M. M. (1999). From sensation to cognition. *Brain, 121* (Pt. 6), 1013-1052.

Mnatsakanian, E. V., & Tarkka, I. M. (2004). Familiar-face recognition and comparison: Source analysis of scalp-recorded event-related potentials. *Clin Neurophysiol, 115*(4), 880-886.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Brain Res Cogn Brain Res, 14,* 115-128.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cereb Cortex, 14,* 452-465.

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Research Cognitive Brain Research, 13,* 417–425.

Morel, A., & Kaas, J. H. (1992). Subdivisions and connections of auditory cortex in owl monkeys. *Journal of Comparative Neurology, 318,* 27-63.

MRC Psycholinguistic Database (www.pst.uwa.edu.au/mrcdatabase/uwa_mrc.htm) Provided by School of Psychology of the University of Western Autralia.

Murray, M. M., Foxe, J. J., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2001). Visuo-spatial response interactions during early sensory processing in humans: A high-density electrical mapping study. *Neuropsychologia, 39,* 828-844.

Murray, M. M., Wylie, G. R., Higgins, B. A., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). The spatiotemporal dynamics of illusory contour processing: Combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging. J *Neurosci, 22*(12), 5055-5073.

Murray, M. M., Foxe, J. J., & Wylie, G. R. (2005, Aug 15). The brain uses single-trial multisensory memories to discriminate without awareness. *Neuroimage, 27*(2), 473-478.

Näätanen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology, 38,* 1–21.

Newell, F. N. (2004). Cross-modal object recognition. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Nunn, J. A., Gregory, L. J., Brammer, M., Williams, S. C. R., Parslow, D.M., & Morgan, M. J. (2002). Functional magnetic resonance imaging of synesthesia: Activation of V4/V8 by spoken words. *Nature Neuroscience, 5,* 371-375.

O' Doherty, J., Rolls, E. T., & Kringelbach, M. (2004). In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Pandya, D. N., Hallett, M., & Mukherjee, S. K. (1969). Intra- and inter-hemispheric connections of neocorticalauditory system in the rhesus monkey. *Brain Res, 14,* 49-65.

Paulesu, E., Harrison, J., Baron-Cohen, S., Watson, J. D. G., Goldstein, L., Heather, J. (1995). The physiology of coloured hearing: A PET activation study of color-word synesthesia. *Brain, 118,* 661-676.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic imaging study. *Journal of Neuroscience, 16,* 5205-5215.

Puce, A., Allison, T., Bentin, S., Asgari, M., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 16,* 5205-5215.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Phil Trans R Soc, 358,* 435-445.

Puce, A., Syngeniotis, A., Thompson, J. C., Abbott, D. F., Wheaton, K. J., & Castiello, U. (2003). The human temporal lobe integrates facial form and motion: Evidence from fMRI and ERP studies. *NeuroImage, 19,* 861–869.

Rach, S., & Diederich, A. (2006). Visual-tactile integration: Does stimulus duration influence the relative amount of response enhancement? *Exp Brain Res, 173*(3), 1246-1266.

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science, 268,* 111-114.

Rauschecker, J. P., Tian, B., Pons, T., & Mishkin, M. (1997). Serial and parallel processing in the rhesus monkey auditory cortex. *Journal of Comparative Neurology, 382,* 89-103.

Recanzone, G. H. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences USA, 95,* 869-875.

Rezak, M., & Benevento, L. A. (1979). A comparison of the organization of the projections of the dorsal lateral geniculatenucleus, the inferior pulvinar, and adjacent lateral pulvinar to primary visual cortex (area17) in the macaque monkey. *Brain Research, 167,* 19-40.

Ritter, W., Deacon, D., Gomes, H., Javitt, D. C., & Vaughan, H. G. (1995, February). The mismatch negativity of event-related potentials as a probe of transient auditory memory: A review. *Jr Ear Hear, 16*(1), 52-67.

Rockland, K. S., Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkeys. *Int Journal of Psycholophysiology, 50,* 19-26.

Ross, L. A., Saint-Amour, D., Leavitt, V., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I'm saying? Optimal visual enhancement of speech recognition in noisy environments. *Cerebral Cortex, 17,* 1147-1153.

Rowland, B. T., Stanford, T. & Stein, B. (2007). A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Exp Brain Res, 180,* 153-161.

Saint-Amour, D., De Sanctis, P., Molholm, M., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45,* 587-597.

Sams, M., H¨am¨al¨ainen, M., Antervo, A., Kaukoranta, E., Reinikainen, K., & Hari, R. (1985). Cerebral neuromagnetic responses evoked by short auditory stimuli. *Electroencephalographic Clinical Neurophysiology, 61,* 254–266.

Sasson, N. J. (2006). The development of face processing in autism. *J Autism Dev Disord, 36*(3), 381-394.

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

Schroeder, C. E., Lindsley, R. W., Specht, C., Marcovici, A., Smiley, J. F., & Javitt, D. C. (2001). Somatosensory input to auditory association cortex in the macaque monkey. *J Neurophysiol, 85*(3), 1328-1327.

Schroeder, C. E., & Foxe, J. J. (2002, June). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Res Cogn Brain Res, 14*(1), 187-198.

Schroeder, C., & Foxe, J. J. (2005). The case for feedforward multisensory convergence during early cortical processing. *NeuroReport, 16*(5), 419-423.

Schwartz, J. L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye: Part II. The psychology of speech-reading and audiovisual speech* (pp. 85-108). East Sussex, England: Psychology Press.

Schwartz, J. L., Berthommier, F. & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition 93*(2), B69-78.

Schwartz, J. L. (2006, October). The 0/0 problem in the fuzzy-logical model of perception. *J Acoust Soc Am, 120*(4), 1795-1798.

Schweinberger, S. R., Burton, A. M. (2003). Covert recognition and the neural system for face processing, *Cortex, 39*(1), 9-30.

Seltzer, B., & Pandya, D. N. (1980). Converging visual and somatic sensory cortical input to the intraprietal sulcus of the rhesus monkey. *Brain Research, 192*(2), 339-351.

Serino, A., Farne, A., Rinaldesi, M. L., Haggard, P. & Lavadas, E. (2007). Can vision of the body ameliorate impaired somatosensory function? *Neuropsychologia, 45,* 1101-1107.

Shenton, M. E., Frumin, M., & McCarley, R. W. (2001) A review of MRI findings in schizophrenia. *Schizophr Res, 49,* 1-52.

Stein, B. E., Huneycutt, W. S., & Meredith, M. A. (1988). Neurons and behavior: The same rules of multisensory integration apply. *Brain Research, 448,* 355-358.

Stein, B. E., Meredith, M. A., Huneycutt, W. S., & McDade, L. (1989). Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience, 1,* 12-24.

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

Stein, B. E., Jiang, W., & Stanford, T. R. (2004). Multisensory integration in single neurons in the midbrain. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard III, M. A. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from the human auditory cortex. *Journal of Neurophysiology, 82,* 2346-2357.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The J Acoust Soc Am, 26,* 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-12). London: Erlbaum.

Tautz, J. (1987). Interneurons in the tritocerebrum of the crayfish. *Brain Res, 407,* 230-239.

Taylor, M. J., Edmonds, G. E., McCarthy, G., & Allison, T. (2001). Eyes first! Eye processing develops before face processing in children. *NeuroReport, 12,* 1671-1676.

Taylor, M. J., Batty, M., & Itier, R. J. (2004). The faces of development: A review of early face processing over childhood. *J Cogn Neurosci,16*(8), 1426-1442.
Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *J Exp Psychol Hum Percept Perform, 30*(5), 873-888.

Todd, J. W. (1912). Reaction to multiple stimuli. In R. S. Woodworth (Ed.), *Archives of psychology*, 25 (Columbia contributions to Philosophy and Psychology, Vol XXI, No. 8). New York: Science Press.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology, 17,* 257-279.

Urbantschitsch, V. (1888). Ueber den Einfluss einer Sinneserregung auf die uebrigen Sinnesempfindungen. *Pfluegers Archiv European Journal of Physiology, 42,* 1.

Valdizan, J. R. (2005). Cognitive evoked potentials in face recognition in autism. *Rev Neurol,5* (40),163-5.

Von Hornbostel, E.M. (1927). "Die Einheit der Sinne", In: Melos, Zeitschrift für Musik 4, 290-297 (1927); H. Werner, "Unity of the Senses" in S.S. Barten and M.B. Franklin, eds., Developmental Processes: Heinz Werner's Selected Writings, Vol.I (New York: International Universities Press, 1978, originally published in 1934)

Wallace, M. T., & Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *J Neurosci, 17,* 2429-2444.

Wallace, M. T. (2004a). The development of multisensory integration. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Wallace, M. T., Perrault Jr., T. J., Hairston, W. D., & Stein, B. E. (2004b). Visual experience is necessary for the development of multisensory integration. *J Neurosci, 24,* 9580-9584.

Wallace, M. T., & Stein, B. E. (2006). Early experience determines how the senses will interact. *J Neurophysiol, 16,* 921-926.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America, 54*(5), 1248-1266.

Wang, M. D., Reed, C. M., & Bilger, R. C. (1978). A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions. *Journal of Speech and Hearing Research, 21,* 5-36.

Wang, X., Mezernich, M. M., Beital, R., & Schreiner, CE (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spectral characteristics. *Journal of Neurophysiology, 74,* 2685-2706.

Wessinger, C. M., Buonocore, M. H., Kussmaul, C. L., & Mangun, R. (1997). Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. *Human Brain Mapping, 5,* 18-25.

Wiersma, C. A. G., & Mill, P. G. (1965). "Descending" neuronal units in the commissure of the crayfish central nervous system and their integration of visual, tactile and proprioceptive stimuli. *J Com Neurol, 125,* 67-94.

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., & Warburton, E. A. (2001). Separate neural subsystems within "Wernicke's area". *Brain, 124,* 83-95.

Zietz, K. (1931). Gegenseitige Beeinflussung von Farb- und Tonerlebnissen. *Zeitschrift für Psychologie, 121,* 257-356.

Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., & Boddaert, N. (2006). Autism: The superior temporal sulcus and social perception. *Trends Neurosci, 29*(7), 359-366.