Helsinki University of Technology
Laboratory of Computational Engineering
Technical report B49

Neurocognitive mechanisms of audiovisual speech perception

Ville Ojanen

An academic dissertation for the degree of Doctor of Philosophy to be presented for public examination and debate in Auditorium S2 at Helsinki University of Technology on the May 25th, 2005, at 12 o'clock noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering
Advanced Magnetic Imaging Centre (AMI)

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio
AMI-keskus

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
P.O. Box 9203
FIN-02015 HUT
FINLAND

Tel. +358-9-451 6157
Fax. +358-9-451 4830

http://www.lce.hut.fi
Online in PDF format: http://lib.hut.fi/Diss/2005/951227681X
E-mail: viloja@lce.hut.fi

ISBN 951-22-7680-1 (printed)
ISBN 951-22-7681-X (PDF)
ISSN 1455-0474

PicaSet Oy
Espoo 2005

Abstract

Face-to-face communication involves both hearing and seeing speech. Heard and seen speech inputs interact during audiovisual speech perception. Specifically, seeing the speaker's mouth and lip movements improves identification of acoustic speech stimuli, especially in noisy conditions. In addition, visual speech may even change the auditory percept. This occurs when mismatching auditory speech is dubbed onto visual articulation.

Research on the brain mechanisms of audiovisual perception aims at revealing where, when and how inputs from different modalities interact. In this thesis, functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG) and behavioral methods were used to study the neurocognitive mechanisms of audiovisual speech perception.

The results suggest that interactions during audiovisual and visual speech perception have an effect on auditory speech processing at early levels of processing hierarchy. The results also suggest that auditory and visual speech inputs interact in the motor cortical areas involved in speech production. Some of these regions are part of the “mirror neuron system” (MNS). The MNS performs a specialized primate cerebral function of coupling two fundamental processes - motor action execution and perception - together. It is suggested that this action-perception coupling mechanism might be involved in audiovisual integration of speech.

Keywords: auditory cortex, functional magnetic resonance imaging, magnetoencephalography, multisensory, audiovisual speech perception, lipreading, Broca, motor cortex, superior temporal sulcus

Author: Ville Ojanen
Laboratory of Computational Engineering
Helsinki University of Technology
Finland

Supervisor: Academy Professor Mikko Sams
Laboratory of Computational Engineering
Helsinki University of Technology
Finland

Preliminary examiners: Professor Christina Krause
Department of Psychology
University of Helsinki
Finland

Professor Matti Laine
Department of Psychology
Åbo Akademi
Finland

Official opponent: Professor Gemma Calvert
Department of Psychology
University of Bath
United Kingdom

Publications

The dissertation is based on following papers:

Study I: **Ojanen, V.**, Tuomainen, J., Sams, M. (2003) Effect of audiovisual primes on identification of auditory target syllables. In Schwartz, J.L., Berthommier, F., Cathiard M.A., Sodoyer, D. (eds.) Proceedings of AVSP 2003, 71-75, St. Jorioz, France.

Study II: Jääskelainen, I.P., **Ojanen, V.**, Ahveninen, J., Auranen, T., Levanen, S., Möttönen, R., Tarnanen, I., Sams, M (2004) Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. Neuroreport, 15(18), 2741-2744.

Study III: Pekkola, J., **Ojanen, V.**, Autti, T., Jääskelainen, I.P., Möttönen, R., Tarkiainen, A., Sams, M (2005) Primary auditory cortex activation by visual speech: an fMRI study at 3 T. Neuroreport, 16(2), 125-128.

Study IV: **Ojanen, V.**, Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. Sams, M. (2005) Processing of audiovisual speech in Broca's area, NeuroImage, 25(2), 333-338.

Study V: **Ojanen, V.**, Pekkola, J., Jääskeläinen, I.P. Möttönen, R., Autti, T., Jousmäki, V., Sams, M. (2005) Common brain areas activated by hearing and seeing speech. Laboratory of Computational Engineering Technical Report B49, ISBN 951-22-7679-8.

Contributions of the author

I was the principal author in Studies I, IV and V. I planned the experiments, prepared the stimuli, carried out the measurements, analyzed the data and wrote the papers. My co-authors provided contributions at all stages of the studies. I had an active role in planning the experiment, carrying out the measurements, analyzing the data and writing the paper in Studies II and III.

Abbreviations

A auditory
AV audiovisual
BA Brodmann's area
BOLD blood oxygenation level dependent
ECD equivalent current dipole
EEG electroencephalography
fMRI functional magnetic resonance imaging
FOV field of view
HG Heschl's gyrus
IPS Intraparietal sulcus
ISI interstimulus interval
MEG magnetoencephalography
M1 primary motor cortex
MNS mirror neuron system
MRI magnetic resonance imaging
MTG middle temporal gyrus
PAC primary auditory cortex
PFC prefrontal cortex
PMC premotor cortex
SNR signal-to-noise ratio
STG superior temporal gyrus
STS superior temporal sulcus
V visual
V1, V2, V5/MT visual cortical areas

Preface

This thesis work was carried out in the Laboratory Computational Engineering (LCE) an Advanced Magnetic Imaging (AMI) Centre at the Helsinki University of Technology. The work was financially supported by the Academy of Finland and Helsingin Sanomain 100-vuotissäätiö.

I wish to express deepest gratitude to academy Professor Mikko Sams for providing excellent supervision and guidance in my research.

I thank my co-authors Dr. Riikka Möttönen, Johanna Pekkola, Prof. Iiro Jääskeläinen and Dr. Taina Autti for comments and ideas to improve the manuscripts and for fruitful discussions.

Thanks also to my colleagues Tobias Andersen, Toni Auranen, Dr. Michael Frydrych, Laura Kauhanen, Jaakko Kauramäki, Cajus Pomren, Jari Kätsyri, Janne Lehtonen, Aapo Nummenmaa, Iina Tarnanen and Kaisa Mälkamäki in LCE.

I am grateful to the former and present personnel of the AMI-centre - Raimo Joensuu, Antti Tarkiainen, Veikko Jousmäki and Marita Kattelus who have made the complications of fMRI research tolerable for me.

I thank Riikka Möttönen, Iiro Jääskeläinen, Tobias Andersen and Kaisa Tiippana for useful comments on the manuscript, Professors Christina Krause and Matti Laine for review and Eeva Lampinen for help with the bureaucracy.

I wish to thank Professors Kimmo Kaski, Jouko Lampinen and Jukka Tulkki in LCE and Professor Riitta Hari, the scientific director of the AMI-centre, for their efforts in establishing the high standard research units I had the privilege to work in.

Thanks also to the staff and patients of the children's phoniatric ward in Helsinki University Central Hospital for giving me insights to the neurocognition of speech perception from the clinical perspective. I want to thank my former colleagues in the University of Turku, most of all Professor Antti Revonsuo, who prompted my interest in experimental scientific research.

My wife Johanna and my children Juho and Pietari have given me motivation, support and inspiration for the lifetime. This work is dedicated to my father Veikko Ojanen, who deceased January 19th 2005.

Ville Ojanen

Table of contents

Chapter 1: Literature review

The sections below review experimental studies and theoretical views on the neurocognitive mechanisms of audiovisual integration of speech. The first and second sections focus on behavioral studies and theoretical views on audiovisual integration of speech, the third and fourth sections focus on neurophysiological studies on the neural mechanisms of audiovisual multisensory integration in general and audiovisual integration of speech in particular.

AUDIOVISUAL SPEECH PERCEPTION

Think of a conversation with your friend in a noisy restaurant at lunch time. Does watching her face and mouth help to hear what she is saying? Most probably you would say that it does. From everyday situations like this we know that visual information from the talker's face facilitates speech perception and is often even crucial for intelligibility of speech. In other words, seeing speaker's articulatory gestures improves identification of acoustic speech, especially in noisy conditions. This phenomenon was experimentally characterized already in the 50's (Sumby and Pollack, 1954).

Sometimes visual speech input may even change the auditory percept, as occurs in the "McGurk effect". For example, simultaneously presented conflicting acoustic /ba/ and visual /ga/ are usually perceived as /da/ (McGurk and MacDonald, 1976). The McGurk effect is a captivating experience. Laboratory studies have shown that it is also a very robust phenomenon. It occurs even when the acoustic stimulus is loud and clear, but the strength of the McGurk effect may increase when the signal-to-noise ratio (SNR) of the acoustic speech decreases (Sekiyama and Tohkura, 1991). The effect occurs even though the observer is aware of how the stimuli are constructed. Dubbing a male voice onto female articulation does not influence the strength of the McGurk effect (Green et al., 1991), and it is not sensitive to the discrepancy in the spatial locations of auditory and visual speech (Jones and Munhall, 1997).

The acoustic and visual stimuli do not have to be in exact synchrony to be integrated. The asynchrony of acoustic and visual stimuli may be as much as 240 ms (Green, 1996; Munhall et al., 1996). Furthermore, instructing subjects to respond to audiovisual stimuli based on auditory or visual information only biases responses towards the instructed modality (Massaro, 1998). However, when visual attention is

directed towards a distractor stimulus presented together with the talking face, the McGurk effect is weaker indicating that visual attention modulates audiovisual speech perception (Tiippana et al., 2004).

Face and mouth are frequent visual stimuli from the birth. Infants learn to imitate mouth movements very early (Meltzoff, 1990) and they start to show interest in matching acoustic and visible speech at the age of 10-20 weeks (Dodd, 1979). At 3 months they have developed the ability to facilitate face recognition by voice information (Burnham, 1998). As young as 4-5 –months of age infants also show the McGurk effect (Burnham and Dodd, 1995). From there on, the effect of the visible speech on speech perception becomes stronger with age (McGurk and MacDonald, 1976).

MODELS OF AUDIOVISUAL SPEECH PERCEPTION

The central questions in audiovisual speech perception research are the following. When are the auditory and visual speech signals combined or integrated during speech processing? Where and how do the two signals interact? What is the common representation for the two very different sensory inputs?

Number of theoretical models have been proposed to explain the audiovisual integration of speech (for reviews, see Massaro, 1998; Schwartz et al., 1998; Summerfield, 1987). The models can be divided into early and late (Schwartz et al., 1998) and auditory and gestural (or “articulatory”) integration models (Green, 1998).

According to the auditory theories, visual information influences the processing of auditory speech. These influences may take place before phonetic categorization of unimodal inputs (early integration) or acoustic and visual speech inputs may be categorized separately before integration (late integration).

Early auditory integration models assume that the visual modality is recoded into or influences the auditory processing early in the processing hierarchy (Green, 1998). The visual signal is recoded into and combined with the auditory signal and processed as single input in the auditory processing stream until it is categorized (e.g., Schroeder and Foxe, 2004).

A well-known example of a late integration model is the Fuzzy Logical Model of Perception (FLMP) (for a review see, Massaro, 1998). According to this model the speech inputs are matched separately against unimodal phonetic prototypes. Then, the

separate classifications are fused through probabilistic computation. The integration is assumed to occur at a post-phonetic level.

According to the articulatory theories, the two inputs are integrated because both signals (e.g., seen /ba/ and heard /ba/) provide the observer with information about the motor act of speaking (e.g., uttering /ba/). These theories have sprung from the motor theory of speech, which advocates the idea that speech perception and production systems are intimately intertwined (Liberman and Mattingly, 1985; Liberman et al., 1967). The theory assumes speech to be perceived by recognition of the articulatory gestures of the speaker rather than the speech sounds. Specifically, speech inputs are mapped into the motor representations which control the vocal tract during the observer's own articulation. In other words, observers map speech input to the motor programs used in their own speech production. During audiovisual speech perception, observers are suggested to map information from seen lip movements and heard speech to the motor programs used in their own speech production. According to this theory, the motor act of articulation serves as the common metric for auditory and visual speech information in audiovisual speech perception (e.g., Skipper et al., 2005).

The articulatory theory is an example of early models as the motor mapping is thought to enable phonetic categorization in the first place. However, in recent neurophysiological literature, the role of articulatory processing in audiovisual speech perception has been incorporated into the late auditory interaction models. Specifically, articulatory processes have been suggested to have a secondary role by constraining and refining primary acoustic-phonetic processing (Callan et al., 2004; Calvert and Campbell, 2003).

NEURAL MECHANISMS OF AUDIOVISUAL MULTISENSORY PROCESSING

Combining and utilizing information from more than one modality is a general function of the nervous system. The following section gives a short summary of the neurophysiological mechanisms and the anatomical regions and connections known to underlie multisensory processing of audiovisual non-speech stimuli (see Calvert et al., 2004).

Multisensory convergence in superior colliculus

The most detailed account of the neuronal mechanisms of multisensory processing is based on single cell recordings directly from the mammalian superior colliculus

(SC) (for a review, see Stein and Meredith, 1993). The SC receives mostly input from the auditory, visual and somatosensory systems. The unimodal neurons in SC have sensory-specific receptive fields to which they respond.

Multisensory neurons in SC are defined by their multiple sensory inputs. These neurons display overlapping sensory receptive fields. The response of these neurons is substantially enhanced during spatially and temporally coinciding multisensory stimulation. The firing rate may be 12 fold of what would be expected by summing the responses of each unimodal input in isolation (Stein and Meredith, 1993). Because the combined response is no longer comparable to the response to either input, it is considered a new output signal. This process is referred to as multisensory integration (Calvert and Thesen, 2004). The enhancement of the neuronal response is often maximal when the responses to the individual inputs are weakest, a principle known as inverse effectiveness. In contrast, spatially or temporally disparate stimuli can induce response depression (Stein and Meredith, 1993).

The capability of noninvasive methods such as BOLD fMRI in demonstrating multisensory convergence (i.e., detecting sub- and supra-additive neuronal responses) might be limited (however, see Calvert et al., 2000). This is because the measured signal originates from a large neuronal population, sub- and supra-additive neurons make up a small portion of the total population of multisensory neurons, and they are not spatially segregated from other neurons (Beauchamp et al., 2004a).

Cortical multisensory areas

Multisensory cortical regions contain, by definition, neurons responsive to stimulation in more than one modality. Several such regions have been identified in the primate brain. These areas include the upper bank of the superior temporal sulcus (STS) (Bruce et al., 1981; Watanabe and Iwai, 1991), intraparietal sulcus (IPS) (Lewis et al., 2000), premotor (PMC) (Graziano and Gandhi, 2000) and prefrontal cortex (PFC) (Benevento et al., 1977; Bremmer et al., 2001; Romanski and Goldman-Rakic, 2002; Watanabe, 1992) (see Figure 1.1 depicting corresponding areas in the human brain).

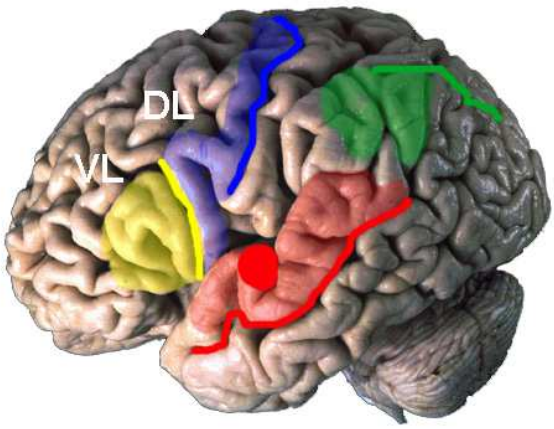

Figure 1.1. Putative cortical multisensory areas: posterior superior temporal sulcus (solid red line depicts entire STS), intraparietal sulcus (solid green line) and prefrontal cortex anterior to the precentral sulcus (solid yellow line; DL=dorsolateral PFC, VL=ventrolateral PFC). Mirror neuron areas: primary motor and premotor areas (blue) anterior to the central sulcus (solid blue line), Broca's area (yellow) and inferior parietal areas (green). Auditory processing areas: superior temporal gyrys (red) and the Heschl´s gyrys/primary auditory cortex (only the lateral end of which is visible, marked as solid red area). See sections above and below for details (The picture is modified from Williams et al., 1997).

Electrophysiological and functional neuroimaging studies have provided information on the topographic organization of these areas. Recent studies in rodents have shown that multisensory neurons are concentrated along boundaries between unisensory areas (Wallace et al., 2004) (Figure 1.2.). The pattern of alternating modality-specific and multisensory zones has been observed also in human. A recent imaging study suggests that auditory and visual inputs arrive in the human STS in separate patches, followed by integration in the intervening cortex (Beauchamp et al., 2004a). The intermixed composition of multisensory and sensory-specific neurons in the multisensory convergence zones suggest a functional architecture in which information from different modalities is brought into close proximity, followed by integration in the intervening cortex (Beauchamp et al., 2004a).

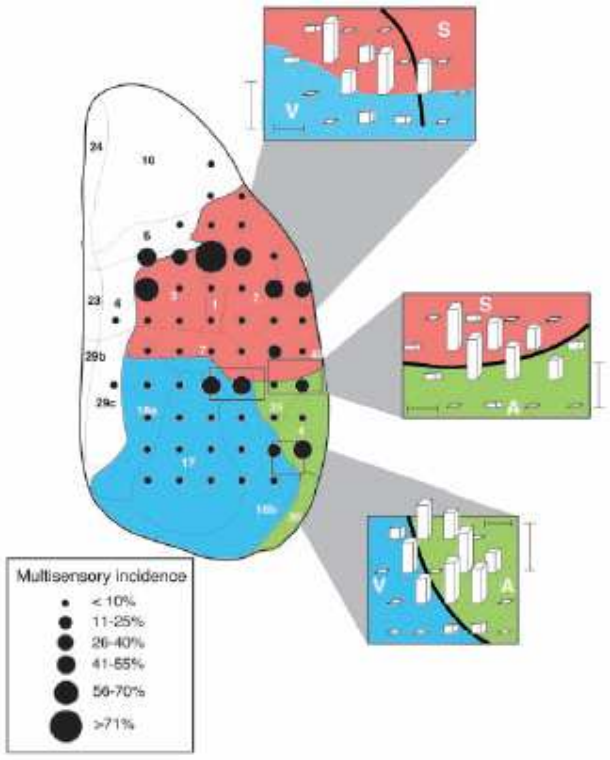

Figure. 1.2. The distribution of multisensory neurons in rat cortex. Major cortical subdivisions are shown in color shading: parietal in red, temporal in green and occipital in blue. Filled circles show locations of single cell recordings and circle size indicates the incidence of multisensory neurons displaying supra-additive responses. Insets show the results of higher-resolution sampling. Bar height indicates the incidence of multisensory neurons (vertical scale bar = 50% multisensory incidence, horizontal scale bar=250 μm). V=visual cortex, A=auditory cortex, S=somatosensory cortex. Adopted from (Wallace et al., 2004). Reprinted with permission of the National Academy of Sciences. Copyright (2004) National Academy of Sciences, U.S.A.

Connectivity of the cortical multisensory areas

In addition to having neurons responsive to stimulation from different modalities, the cortical multisensory areas have anatomical connectivity that relates them to more than one modality (for reviews see Kaas and Collins, 2004; Miller and Cohen, 2001). Most of the data on brain connectivity is based on anatomical tracing studies and electrophysiological experiments in monkeys. The cortico-cortical connections of the multisensory areas are briefly described below (see Figure 1.3.).

Superior temporal sulcus

STS receives inputs from a number of higher-order visual areas, like posterior parietal and inferotemporal areas (reviewed in Kaas and Collins, 2004). It encloses neurons with reciprocal connections to secondary auditory areas of the anterior and posterior STG (Seltzer and Pandya, 1991) and neurons with one-way connections to primary visual cortex (Falchier et al., 2002) (Figure 1.3.). The STS is also reciprocally

connected to PMC (Deacon, 1992), Broca's area in human (Catani et al., 2005), ventrolateral (VL) PFC (Petrides and Pandya, 2002) and IPS (Kaas and Collins, 2004).

Cortex of the intraparietal sulcus

The multisensory cortex inferior of intraparietal sulcus receives input from primary and secondary visual, and secondary auditory areas (Kaas and Collins, 2004). The IPS areas project reciprocally to PFC and ventral PMC. The most inferior part of IPS cortex has connections also with STS (see (Kaas and Collins, 2004). In human, inferior parietal cortex (BA 40) has reciprocal connections to posterior STG and Broca´s area (Catani et al., 2005).

Premotor cortex

Neurons in the monkey premotor area F5, which is considered to be the homologue of Broca´s area (BA 44) - respond to both sound and vision (Kohler et al., 2002). The IPS region is likely to be the main source of visual input to PMC (Graziano et al., 1999). For auditory input, area F5 is connected to STS and posterior STG regions (reviewed in Arbib and Bota, 2003). In human, Arcuate fasciculus connects Broca´s area to posterior temporal areas including STG, STS and MTG directly and indirectly through inferior parietal cortex (Catani et al., 2005). Premotor cortex is closely interconnected with M1 (Miller and Cohen, 2001).

Prefrontal cortex

The dorsolateral (DL) and ventrolateral (VL) PFC receive reciprocal projections from anterior and posterior STG as well as from secondary visual and parietal cortices. In addition PFC regions are extensively interconnected (for a review, see Miller and Cohen, 2001). The anterior STG projects mainly to ventral PCF and the posterior STG to the dorsal PCF (Romanski et al., 1999). This pattern is similar to that observed in the visual system – the ventral "what" stream projects to VLPFC and the dorsal "where" stream projects to DLPFC (Wilson et al., 1993). Both PFC areas receive inputs from the posterior superior temporal sulcus (Petrides and Pandya, 2002). Dorsal PFC is connected with premotor cortex (Lu et al., 1994).

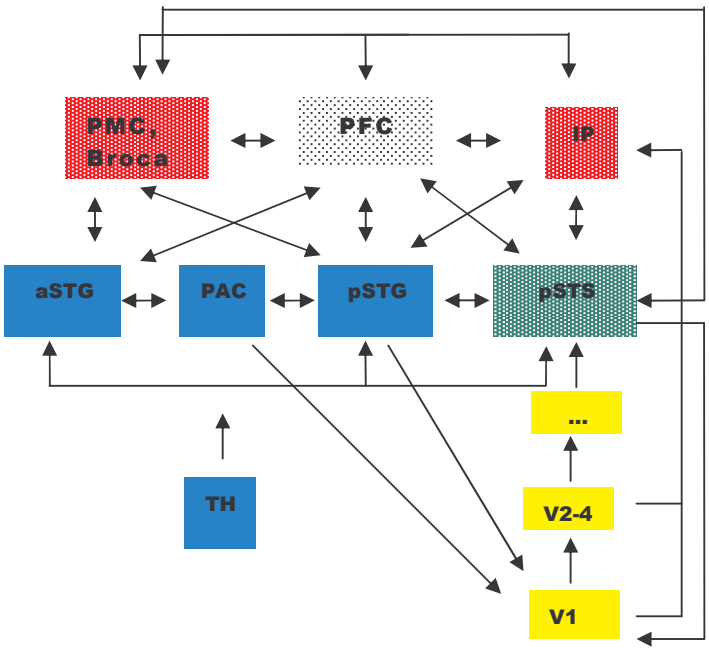

Figure 1.3. A schematic, simplified illustration of the anatomical structures, interconnectivity and multisensory properties (rasterized colour) of the key cortical areas involved in audiovisual speech perception: auditory areas (blue), multisensory posterior STS (green), visual areas (yellow) and mirror-neuron system (red). PMC=premotor cortex, PFC=prefrontal cortex, IP=inferior parietal cortex, PAC=primary auditory cortex, aSTG=anterior STG, pSTG=posterior STG, pSTS= posterior superior temporal sulcus, TH=thalamus, V1=primary visual area, V2-4=secondary visual areas, … = higher-order visual processing.

Multisensory convergence in sensory specific areas

Recent electrophysiological (Fort et al., 2002; Foxe et al., 2000; Giard and Peronnet, 1999; Molholm et al., 2004; Molholm et al., 2002; for a review, see Fort and Giard, 2004) and functional neuroimaging (Calvert et al., 1999; Laurienti et al., 2002; Macaluso et al., 2000) studies in humans as well as electrophysiological findings in monkeys (Fu et al., 2003; Schroeder et al., 2001) have suggested that cortical multisensory integration can occur at very early stages of the cortical processing hierarchy—stages previously thought to be purely unisensory.

In support, there is evidence that neurons at sensory specific cortical areas can be activated by stimulation from other modalities (Wallace et al., 2004) (Figure 1.2.). Specifically, in visual cortex there are neurons responsive to auditory inputs (Fishman and Michael, 1973; Morrell, 1972; Spinelli et al., 1968) and in auditory cortex neurons responsive to visual inputs (Schroeder and Foxe, 2002). Recent studies in cats have shown that primary and secondary auditory cortices project directly to primary

visual cortex (Falchier et al., 2002; Rockland and Ojima, 2003). These connections may not be reciprocal, since projections from V1 and V2 to auditory areas have not been demonstrated (Kaas and Collins, 2004). Single cell recordings in the monkey posterior auditory cortex have shown that the responses to visual stimuli in auditory cortex neurons are very early (~50ms from stimulus onset) and display feedback properties, suggesting that visual input originates from higher cortical region (Schroeder and Foxe, 2002; Schroeder et al., 2003).

AUDIOVISUAL INTEGRATION OF SPEECH IN THE HUMAN BRAIN

The previous chapter reviewed the general neural mechanisms involved in audiovisual multisensory processing. However, in audiovisual speech perception, both general and speech-specific multisensory mechanisms might be important (see Calvert et al., 2004; Klucharev et al., 2003). The following section gives a short description of the cortical processing of auditory speech, followed by a detailed review of the neuroimaging literature on audiovisual speech processing.

Cortical processing of auditory speech

Functional imaging studies of speech perception have typically shown bilateral responses to speech in STG/STS of the temporal lobes (Binder et al., 2000; Vouloumanos et al., 2001; Zatorre et al., 1992) (see Figure 1.1). Processing of the acoustic features of non-speech sounds has been attributed to the primary auditory cortex and dorsolateral portions of STG (Binder et al., 2000), whereas phonetic processing of speech signals involves ventral STG and the adjacent multisensory STS extending both anteriorly and posteriorly (Binder et al., 2000; Jäncke et al., 2002; Narain et al., 2003; Scott et al., 2000; Vouloumanos et al., 2001). The primary auditory cortex (PAC, BA 41), is in the medial portion of Heschl's gyrus (HG) (Rademacher et al., 1993). The secondary auditory cortex is in the surrounding regions of STG and STS encompassing Brodmann's Areas 42, 21 and 22 (Figure 1.1). Anatomical and functional boundaries between auditory processing areas are not precisely known, and may vary among individuals (Rademacher et al., 1993).

Recent theories of auditory speech processing have suggested that the ventral "what" and dorsal "where" streams of auditory processing (Romanski et al., 1999) would analyze the acoustic speech input into acoustic-phonetic and articulatory-based representations, respectively (Hickok and Poeppel, 2000; Hickok and Poeppel, 2004; Scott and Johnsrude, 2003; Scott and Wise, 2004). According to Scott et al. (2003,

2004) the ventral stream includes the anterior auditory cortical areas in STG/STS, connected with Broca's area. The dorsal stream includes posterior STG/STS which is connected to premotor cortex. Hickock and Poeppel (2004) propose slightly different brain areas. According to them the acoustic-phonetic representations are processed in STG. The posterior aspect of Sylvian fissure at the boundary between the parietal and temporal lobes serves as an interface between sound-based and articulatory-based representations (Hickok and Poeppel, 2004). Both theories suggest a link between speech perception and production.

In support, neuroimaging studies show that the "speech motor regions" (Broca's area and PMC, see Figure 1.1.) are involved in the processing of auditory speech information. Broca´s area is activated during speech production (Heim et al., 2003), passive listening to speech (Benson et al., 2001; Binder et al., 2000; Wilson et al., 2004) and in various acoustic speech processing tasks involving phonological processing (Zatorre et al., 1992), verbal working memory (Braver et al., 1997), and speech sound segmentation (Burton et al., 2000). Recently, it has been shown that *listening* to meaningless monosyllables activates a superior portion of ventral premotor cortex on the border of Brodmann areas 4a and 6, overlapping the motor activations during *production* of the same syllables and possibly extending to the most anterior part of M1 (Wilson et al., 2004). Furthermore, recent transcranial magnetic stimulation studies have found evidence of the activation of the speech motor areas during auditory speech perception (Fadiga et al., 2002; Watkins et al., 2003).

Activation of the speech motor regions during auditory speech perception has been suggested to be related to the functioning of "mirror neurons" (for reviews, see (Rizzolatti and Arbib, 1998; Scott and Johnsrude, 2003). Such neurons in the monkey brain are activated both during execution of goal-directed hand and mouth actions and during observation of similar actions performed by other individuals (di Pellegrino et al., 1992; Fadiga et al., 1995; Ferrari et al., 2003; Gallese et al., 1996). Cortical mirror neurons have been found in two main regions (mirror neuron system, MNS): the ventral premotor cortex (F5, the monkey homologue of Broca's area) and the rostral part of the inferior parietal lobule (for a review see Rizzolatti and Craighero, 2004). MNS has been demonstrated also in humans (Fadiga et al., 1995; Hari et al., 1998; Nishitani and Hari, 2000) (see Figure 1.1). In monkeys some of the mirror neurons are audiovisual, meaning that they are activated by both seen and heard actions (Keysers et al., 2003; Kohler et al., 2002). Because it seems to function as an interface between

motor actions and perception, the MNS has been suggested to play an important role in speech communication (for reviews, see Rizzolatti and Arbib, 1998; Rizzolatti and Craighero, 2004; Rizzolatti et al., 2001) in agreement with the motor theory of speech perception (Liberman and Mattingly, 1985; Liberman et al., 1967). Mirror neurons provide a plausible neural mechanism for the articulatory-based processing of speech information (Scott and Johnsrude, 2003).

Cortical processing of audiovisual speech

Neuroimaging and anatomical evidence suggest that the audiovisual integration of speech is achieved by processing in four key cortical areas, which are closely connected: posterior STS (Callan et al., 2004; Callan et al., 2003; Calvert et al., 2000; Macaluso et al., 2004; Möttönen et al., 2004; Sekiyama et al., 2003; Skipper et al., 2005; Wright et al., 2003) the sensory-specific auditory (Callan et al., 2004; Calvert et al., 1999; Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991) and visual cortices (Calvert et al., 1999) and the speech motor regions (Callan et al., 2004; Callan et al., 2003; Jones and Callan, 2003; MacSweeney et al., 2002b; Sekiyama et al., 2003; Skipper et al., 2005) (see Figures 1.1. and 1.3.).

Specific neurophysiological models of audiovisual integration of speech have been suggested emphasizing 1) multisensory convergence in STS (Calvert et al., 1997; Calvert et al., 2000), 2) interactions in the auditory cortical areas (Besle et al., 2004; Klucharev et al., 2003; Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991; Schroeder and Foxe, 2004; van Wassenhove et al., 2005), or 3) modulation of auditory or multisensory processing through back-projections from speech motor/mirror neuron areas (Callan et al., 2004; Calvert and Campbell, 2003; Skipper et al., 2005). The following section reviews the literature relevant to these models, although all studies have not explicitly suggested any model for audiovisual speech integration.

Multisensory convergence in the superior temporal region

The posterior part of STS has been identified as a brain area involved in audiovisual integration of speech in a number of imaging studies. The posterior STS responds to audiovisual speech stimulation (Callan et al., 2004; Callan et al., 2003; Calvert et al., 2000; Macaluso et al., 2004; Sekiyama et al., 2003; Skipper et al., 2005; Wright et al., 2003) as well as auditory (e.g., Binder et al., 2000) and visual speech

stimulation (Calvert and Campbell, 2003; Calvert et al., 1997; Campbell et al., 2001; MacSweeney et al., 2000; MacSweeney et al., 2001; MacSweeney et al., 2002b). In addition, posterior STS is responsive to human non-speech voices (Belin et al., 2000) and visual biological movements (for a review see Allison et al., 2000). The region is also important for integrating auditory and visual information about objects (Beauchamp et al., 2004b) and letters (Raij et al., 2000; van Atteveldt et al., 2004).

To study the brain areas of audiovisual integration of speech, Calvert et al. (2000) contrasted the brain activations to semantically matching and conflicting audiovisual speech (spoken extracts from a book) with the combined response to unimodal acoustic and visual speech. They used a paradigm in which matching and conflicting auditory and visual speech stimuli were presented simultaneously but so that the sequence resulted in periods of audiovisual, auditory and visual presentations of the stimuli. The matching condition was presumed to lead to multisensory integration. During conflicting AV stimulation the semantic as well as phonological and temporal coherence of the stimulation was disrupted. Analysis was targeted to find the voxels which display response properties analogous to those of superior colliculus multisensory integrative cells (Stein 1993). Only left STS exhibited significant supra-additive (AV > A+V) response enhancement to matching audiovisual speech and sub-additive (AV < A+V) response to conflicting audiovisual speech. As a plausible mechanism of audiovisual integration of speech, the authors suggested that the multisensory speech input would be initially integrated in STS and this area would subsequently modulate activity in the sensory-specific auditory cortices through feedback projections (Calvert et al., 2000).

Supporting the role of STS in audiovisual integration of speech, Wright et al. (2003) found both enhanced and suppressed activations in bilateral STS region during observation of matching meaningful AV words in comparison to the unimodal responses. In addition, in a recent fMRI study, left posterior STS was found to be more active during the observation of continuous audiovisual than auditory spoken stories (Skipper et al., 2005).

Furthermore, suppressed responses to audiovisual in comparison to the combined responses to unimodal speech stimuli (syllables) have been shown with MEG in the right STS region at 200-600 ms from stimulus onset (Möttönen et al., 2004). Klucharev et al. (2003) compared ERPs to audiovisual vowels which were phonetically either matching (e.g., acoustic /a/ and visual /a/) or conflicting (e.g.,

acoustic /a/ and visual /y/). They found differences in the ERPs to conflicting and matching audiovisual vowels peaking at three latencies, at 155 ms, 235 ms and 325 ms from stimulus onset. These relatively late effects were suggested to reflect AV interactions at phonetic level in the multisensory cortices.

The STS is implicated in audiovisual integration of speech by recent fMRI experiments which explored the brain mechanisms of enhanced perceptibility of degraded auditory speech by concordant visual speech (Callan et al., 2003; Sekiyama et al., 2003). Callan et al. (2003) presented audiovisual and auditory speech (meaningful words) with and without acoustic noise to the subjects. Increased activity in MTG and STG/STS was observed when audiovisual speech was presented with acoustic noise in comparison to audiovisual speech with no noise (Callan et al., 2003). Using similar approach, Sekiyama and coworkers (2003) investigated the McGurk effect with fMRI and positron emission tomography (PET) using conflicting audiovisual speech stimuli. Spoken syllables were presented auditorily, visually or audiovisually. The auditory component of the conflicting audiovisual stimuli was presented with and without added noise. Direct comparison between the audiovisual conditions showed increased activation with added noise in the posterior part of the left STS. Together these studies (Callan et al., 2003; Sekiyama et al., 2003) suggest that the responses of the multisensory neurons in STG/STS region display the principle of inverse effectiveness (see previous section) - the enhancement of STG/STS activity is greatest when the unimodal acoustic stimulus is the least effective.

Callan et al. studied audiovisual integration of speech using degraded visual speech stimuli (Callan et al., 2004). The purpose of the study was to control for multisensory responses resulting from cross-modal attentional modulation cued by aspects of visual speech information, which are not specific to place of articulation information. Visual components of audiovisual speech stimuli (sentences) were spatial wavelet bandpass filtered at low and middle frequencies (LF and MF, respectively) and stimuli were presented with background auditory noise. Unfiltered visual speech information (UF) signals the place of articulation and onsets and offsets of the acoustic speech signal. The place of articulation information is preserved in the MF stimuli but is not present in the LF stimuli. The multisensory responses selectively induced by place of articulation information were established by the presence of activity for both the UF

and MF conditions relative to the LF condition. These responses were detected in left MTG, and STG/STS, including the auditory cortex.

In a PET study, Macaluso et al. (2004) specifically manipulated the temporal and spatial synchrony of auditory and visual in audiovisual stimulation (words). Synchronous versus asynchronous audiovisual speech yielded increased activity in STS region. The spatial location of the sound source had no effect on STS activation. This indicates that temporal but not spatial synchrony of matching auditory and visual speech is critical to integrative effects in STS.

Some studies have failed to show audiovisual speech integration effects in STS region (Calvert et al., 1999; Jones and Callan, 2003; Olson et al., 2002). In an fMRI study by Calvert et al. (1999), audiovisual stimuli (spoken numbers) did not evoke more activation in STS than unimodal auditory stimuli. In an fMRI study of the McGurk effect with vowel-consonant-vowel (e.g., /aka/) stimuli (Jones and Callan, 2003), greater responses in STS were not observed for matching than conflicting audiovisual stimuli. Quite contrary, the conflicting stimuli activated larger areas of STS region. Also Olson et al. (2002) did not find enhanced activation in STS region when they compared the BOLD responses to synchronized over desynchronized conflicting audiovisual words producing the McGurk illusion.

Despite the growing body of evidence on the importance of STS region in audiovisual integration of speech, it is unclear which features of the audiovisual stimuli give rise to the enhanced activation in posterior STS region. The enhanced activation might be due at least to semantic, phonological or temporal coherence between auditory and visual speech during audiovisual speech stimulation. The differences between studies that support STS as an audiovisual speech integration area and those that do not, suggest that the nature of the stimuli (e.g., sentences, words or syllables), contrasts between unimodal and audiovisual combinations (A+V *vs.* AV, matching *vs.* conflicting or McGurk) and the way integration is manipulated (temporal synchrony, acoustic SNR, the strength of the McGurk effect) are important factors in determining whether or not activation is detected in STS.

STS activation seems to be particularly sensitive to auditory SNR (Callan et al., 2003; Sekiyama et al., 2003). However, it is difficult to accurately control and know the SNR of the acoustic stimuli in the MR scanner during conventional continuous imaging paradigms. Furthermore, the auditory BOLD-response itself is very sensitive to acoustic noise from the MR equipment (Shah et al., 1999), for reviews, see (Di

Salle et al., 2003; Moelker and Pattynama, 2003). This complicates the comparison of the results from separate studies and might explain some of the discrepant results.

Interactions in sensory specific areas

MEG studies have shown that visual speech modifies activity in the auditory cortices (BA 41/42, 22) during audiovisual speech observation ~50-200 ms after stimulus onset (Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991). Similarly, an EEG study by Klucharev et al. (2003) show that the amplitude of the N100 event-related potential (ERP) component peaking at 85 ms from stimulus onset is suppressed during audiovisual stimulation in comparison to the ERPs to the sum of the unimodal responses. The N100 suppression was suggested to reflect modified activity in the sensory-specific cortices (see also, Besle et al., 2004; van Wassenhove et al., 2005). These studies suggest that audiovisual integration of speech occurs early in the cortical auditory processing hierarchy.

Recent fMRI studies report response enhancement of the auditory cortex activity (identified with reference to an atlas by Rademacher et al. 2001) by visual speech in the presence of acoustic noise (Callan et al., 2003) and in comparison to varying levels of degraded visual speech (Callan et al., 2004) during audiovisual speech observation. Also the study by Calvert et al. (2000) showed supra-additive responses to matching audiovisual speech but not subadditive responses to conflicting audiovisual speech in sensory specific auditory (BA 41/42) and visual cortices (BA 19). Furthermore, during audiovisual speech perception, BOLD responses in the auditory cortex (BA 42/41) as well as in the visual motion cortex (V5/MT, BA 37/19) are enhanced in comparison to responses during auditory or visual speech stimulation (Calvert et al., 1999). These results demonstrate audiovisual interactions in the sensory specific cortices.

Neuroimaging studies have consistently shown that visual speech is processed in the auditory cortical areas of STG (Bernstein et al., 2002; Calvert and Campbell, 2003; Calvert et al., 1999; Calvert et al., 1997; Calvert et al., 2000; Campbell et al., 2001; MacSweeney et al., 2000; MacSweeney et al., 2002a; MacSweeney et al., 2001; Olson et al., 2002; Paulesu et al., 2003; Sekiyama et al., 2003; Wright et al., 2003). Whether this activation is limited to the hierarchically higher auditory areas in STG and STS or whether purely visual input can activate also PAC has, however, remained an open question (Bernstein et al., 2002; Calvert et al., 1997; MacSweeney et al.,

2000). Calvert et al. (1997) reported PAC and secondary auditory cortex activation by silent lip-reading. PAC activation during silent lip-reading was confirmed in a subsequent study using a silent event-related paradigm (MacSweeney et al., 2000). However, PAC activation was not found during silent lip-reading in a study, where probabilistic mapping was used as a tool to identify PAC (Bernstein et al., 2002).

It is not known which brain regions project the visual speech input to the auditory processing areas (for different possibilities, see Figure 1.3). It has been proposed that visual speech has access to sensory specific auditory cortex through feedback projections from multisensory neurons in posterior STS (Calvert et al., 2000). In support, there is evidence that responses to visual stimuli in auditory cortex neurons are projected from higher cortical regions (Schroeder and Foxe, 2002; Schroeder et al., 2003). On the other hand, multisensory interactions might occur earlier in the auditory cortices (150–200 ms) than in the right STS region (250–600 ms) (Möttönen et al., 2004).

Taken together, these findings have been interpreted to indicate that viewing speech influences the processing of acoustic speech in auditory sensory specific cortex at an early stage of speech processing possibly through feedback projections from the multisensory STS (Calvert et al., 1997; Calvert et al., 2000; Klucharev et al., 2003; Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991; Schroeder and Foxe, 2004).

Processing in the speech motor areas

Activity in brain regions involved with planning and execution of speech production (Broca's area, PMC and anterior insula (Dronkers and Ogar, 2004)) is also very consistently shown in studies of audiovisual speech perception (Callan et al., 2004; Callan et al., 2003; Calvert et al., 1999; Calvert et al., 2000; Jones and Callan, 2003; MacSweeney et al., 2002b; Olson et al., 2002; Sekiyama et al., 2003; Skipper et al., 2005). Additionally, most of the studies using unimodal visual speech stimulation have shown activation in these areas (Bernstein et al., 2002; Callan et al., 2004; Callan et al., 2003; Calvert and Campbell, 2003; Calvert et al., 1997; Campbell et al., 2001; MacSweeney et al., 2000; Nishitani and Hari, 2002; Olson et al., 2002; Paulesu et al., 2003; Sekiyama et al., 2003; Skipper et al., 2005). However, activation in the speech motor areas is not reported in all studies of lipreading (MacSweeney et al., 2002a; MacSweeney et al., 2001).

The speech motor regions are activated also during auditory speech perception. Broca's area and PMC are activated during passive listening (Benson et al., 2001; Binder et al., 2000; Wilson et al., 2004) and phonetic analysis of auditory speech (Burton et al., 2000; Paulesu et al., 1993; Zatorre et al., 1992; Zatorre et al., 1996).

Transcranial magnetic stimulation experiments demonstrate that observation of visual (Sundara et al., 2001; Watkins et al., 2003) and auditory (Fadiga et al., 2002; Watkins et al., 2003) speech increases the excitability of the orofacial motor system. Furthermore, left somatosensory cortex MEG responses to tactile lip stimulation are modulated during speech viewing but not during listening to speech (Möttönen et al., 2005).

Evidence of the roles of Broca's area and PMC in audiovisual integration of speech comes from studies contrasting audiovisual conditions with different levels of acoustic noise (Callan et al., 2003; Sekiyama et al., 2003) and degraded visual input (Callan et al., 2004). These studies (Callan et al., 2003; Sekiyama et al., 2003) show that as well as in STG/STS region, also in the speech motor areas the enhancement of activity during multisensory stimulation is greatest when the unimodal acoustic stimulus is the least effective, displaying inverse effectiveness.

Together these findings suggest that the motor regions of speech production might participate in unimodal auditory and visual as well as in audiovisual speech perception. This is further supported by the multisensory properties of the precentral gyrus (Graziano and Gandhi, 2000), PFC (Romanski and Goldman-Rakic, 2002) and ventral premotor cortex (monkey homologue of Broca's area) (Kohler et al., 2002).

Mirror neurons and audiovisual speech perception

The speech motor regions have been suggested to be engaged in visual and audiovisual speech perception through the functioning of mirror neurons in Broca's area, PMC and inferior parietal lobule (Callan et al., 2004; Callan et al., 2003; Calvert and Campbell, 2003; Campbell et al., 2001; MacSweeney et al., 2000; Nishitani and Hari, 2002; Paulesu et al., 2003; Skipper et al., 2005). MacSweeney et al. (2000) observed activation in response to silent lipreading bilaterally in PMC and Broca's area. They suggest that the activations might be related to speech comprehension possibly via the mirror function of these areas (see also Paulesu et al., 2003). Campbell et al. (2001) observed activation of Broca's area during silent lip-reading when contrasted to the activation during observation of meaningless facial movements

(gurning). The authors attribute the activation to either mirror neurons through imitation of observed biological actions, greater articulatory and/or lexical demands during visual speech than gurning observation, or prefrontal language rehearsal system (Paulesu et al., 1993) irrespective of the 'mirror function' (Campbell et al., 2001). Interestingly, this result was not replicated in a second experiment where the visual speech stimuli included the upper half of the body of the speaker, whereas in the first experiment only the mouth was visible. This suggests that activation in Broca's area related to action observation may be sensitive to the visible area of face in the images displayed (Campbell et al., 2001).

Nishitani and Hari (2002) measured cortical evoked magnetic responses to observation and imitation of static lip forms as well as execution of similar facial gestures. They found that the MNS is activated during observation and imitation in a distinct temporal sequence - from STS to inferior parietal areas, ending in activation in Broca's area and M1 at 220-340 ms after stimulus onset. They propose that the results agree with the motor theory of speech perception (Liberman and Mattingly, 1985; Liberman et al., 1967).

Calvert and Campbell (2003) observed activation in Broca's area and left PMC/M1 during moving visual speech and stilled pictures of articulatory movements. They suggest that following processing in the visual cortical areas, visual speech information is routed to the MNS. Phonetic representations in left posterior STS regions are then accessed through back-projections from the MNS (Calvert and Campbell, 2003). This is supported by the known neural projections in human brain from the prefrontal areas (including PMC and Broca's area) to STG/STS regions (Catani et al., 2005) (see Figure 1.3.).

Similarly, Callan et al. (2003, 2004) suggested that during audiovisual stimulation, both multisensory integration in STS, and internal articulatory simulation of the intended speech act of the observed speaker through MNS activation, facilitate auditory speech perception. Furthermore, Callan et al. (2004) suggest that the internal articulatory simulation mechanism might be dependent on task demands so that the harder the task the more internal articulatory simulation is involved in perceptual processing.

Recently, Skipper et al. (2005) have elaborated the role of speech motor regions in AV integration. Their model relies on the neurophysiological evidence for the parallel functional properties of the posterior STS and Broca's area in multisensory speech

perception and their anatomical connectivity. According to the model these two functionally and anatomically closely connected regions interact to produce multisensory representations of speech input by matching acoustic-phonetic sensory patterns to internal articulatory motor commands that the observer uses for own articulation. At the motor end of the sensory-motor continuum the mapping is carried out by interaction between Broca's area, PMC and M1. This sensory to motor mapping generates predictions of the sensory consequences of the motor match. Predictions are used to constrain the phonetic interpretation of the sensory input. This process involves interaction between STS and/or inferior parietal areas as well as feedback from premotor or motor cortices (Skipper et al., 2005).

Silent articulation as a possible confound

A plausible criticism of the results and models reviewed above would be that the speech motor regions are activated during speech perception because subjects might be engaged in some degree in overt articulation (involving motor activity of the vocal tract) or silent articulation (involving "inner speak" but no motor activity of the vocal tract) during speech perception (Paulesu et al., 1993; Sekiyama et al., 2003). Broca's area is activated during overt and silent articulation (Huang et al., 2002). Therefore, activation of Broca's area due to silent articulation is a natural concern in studies where the subjects might be engaged in silent articulation during the active condition but the control condition does not involve silent articulation (Bernstein et al., 2002; Jones and Callan, 2003; MacSweeney et al., 2002b; Sekiyama et al., 2003). However, it may be difficult to separate activations due to redundant silent articulation from actual perceptual mechanisms involving speech motor processing.

Several studies of lipreading and audiovisual speech perception have tried to overcome these confounds by controlling for articulatory as well as other motor demands between the conditions. In the lipreading studies by Calvert et al. (1997), MacSweeney et al. (2000), Campbell et al. (2001) and Calvert and Campbell (2003) all conditions, including control conditions, required participants to generate silent speech at the same rate, with no other motor tasks. However, enhanced premotor or Broca's area activation in these studies might still reflect small differences in the articulation demands (Calvert et al., 1997; MacSweeney et al., 2000).

Olson et al. (2002) had no motor task, but an articulatory task (rate matched to stimulation) in the control but not in the active condition. Calvert et al. (1999) had no

motor task and used direct contrasts between active conditions (AV-A and AV-A) with equal subvocal articulation in both conditions. Skipper et al. (2005) presented highly engaging unimodal and audiovisual stories to the subjects while they were simply attending to them. No overt motor response was required. Activation in the speech motor regions was detected in AV-A and AV-V contrast. Paulesu et al. (2003) were able to demonstrate a negative correlation between identification scores of visual speech stimuli and rCBF (regional cerebral blood flow) in speech motor regions. The authors argue that if inner articulation is related to visual speech perception, then increasingly successful identification of the stimuli should lead to a positive correlation with rCBF within the speech motor regions. The negative correlation suggests that the rCBF is increased the fewer accurate responses the subject made, indicating that the activation was related to more demanding phonological analysis of the stimuli. The MEG studies by Nishitani and Hari (2003) demonstrating motor and Möttönen et al. (2005) demonstrating somatosensory activations during visual speech perception controlled for overt articulation by measuring electromyography (EMG) signal from the orbicularis oris muscle during stimulation. To my knowledge, there are no auditory speech perception studies which would have used silent articulation as a control condition.

SUMMARY

Neuroimaging studies suggest that the audiovisual integration of speech is achieved by processing in posterior STS, the sensory-specific auditory and visual cortices and the speech motor regions/MNS. Some of these regions have been suggested to subserve the analysis of auditory speech input into acoustic-phonetic and articulatory representations. However, the role of sensory-specific as well as multisensory convergence regions in audiovisual speech perception is unclear and the contribution of the MNS and speech motor regions to audiovisual speech integration needs to be further clarified.

Chapter 2: Brain imaging methods used in the studies

The primary brain imaging method used in the experiments of this study was functional magnetic resonance imaging (fMRI). Magnetoencephalography (MEG), as well as behavioral reaction time measurements were used in two of the five experiments. The fMRI section below is largely based on the book by Jezzard et al. (Jezzard et al., 2001). The MEG section is based on the review article by Hari (Hari, 1999).

FUNCTIONAL MAGNETIC RESONANCE IMAGING

Functional magnetic resonance imaging is a method which allows precise localization of brain activity by measuring a signal which is dependent on changes in blood circulation and oxygenation due to neuronal activity.

The method is based on the dissimilarity of different types of tissue in terms of their magnetic susceptibility, i.e., difference in response when placed in a strong magnetic field and excited with a radio-frequency pulse (see Jezzard et al., 2001). In practice, due to sensory stimulation, a local increase in neuronal activity causes local increase in blood flow and in the amount of oxygenated hemoglobin flowing to the brain area. The increase in oxygenated haemoglobin is beyond the metabolic need and thus the proportion of deoxyhaemoglobin is reduced. The two different forms of hemoglobin have different magnetic properties. Deoxyhaemoglobin disturbs the local magnetic field whereas oxygenated haemoglobin does not. As the net result, the measured local signal intensity is reduced and regions of the brain that have enhanced activity appear brighter in the image. Although the Blood Oxygenation Level Dependent (BOLD) signal has been shown to correlate with neural activity measured using microelectrodes (Logothetis et al., 2001), the exact nature of the coupling of neuronal and vascular responses is unknown. The physiological time course of cerebral blood flow following neuronal activation (3-4 seconds to maximum) is the major limitation to the temporal resolution of fMRI.

FMRI experiments

In an fMRI experiment, usually hundreds of consecutive MR-images are measured. The resulting data are a time series of signal intensities in all of the individual volume elements (voxels) which make up the image.

Several preprocessing steps are performed before analyzing the data (see Jezzard et al., 2001). Head motion correction has to be done for each subject, because due to the subjects head movement during the experiment the same voxel does not represent the same location in the brain throughout time. Spatial smoothing increases the SNR in the fMRI signal and enables the use of Gaussian Randon Field Theory to correct for multiple comparisons in the statistical analysis by making the data more normally distributed. Temporal filtering removes high frequency fluctuations and long term drifts from the time series.

One of the advantages of fMRI is its ability to detect robust signal changes in individual subjects. For intra-subject anatomical alignment of the activation maps, the low-resolution fMRI scan is co-registered to a high-resolution anatomical scan from the same individual. For group analysis, the brains of individual subjects are transformed to match a standard brain. Spatial transformations results in some reduction of spatial accuracy. The group average fMRI analysis is based on the assumption that cognitive and perceptual functions are mapped onto roughly same anatomical locations across individuals. However, in some cases this approximation might be too rough. Therefore, per-participant analysis might be advantageous in estimating the exact location, extent and intensity of brain activations.

The most commonly used type of analysis, the General Linear Model (GLM), sets up a model derived from the stimulation that was applied during the scanning, and fits it to the data. The fitting is done by performing voxel-wise tests of the hypothesis that the observed time course is not significantly related to the model. An activation map is then constructed by applying a significance threshold to the resulting statistics. The analysis involves hundreds of thousands of statistical tests and the statistics has to be corrected for multiple comparisons to avoid false positive results. One way to correct for false positives is to divide the probability threshold by the number of tests (Bonferroni correction). This is, however, too stringent correction, since all individual voxels are not independent measurements. Due to smoothing, interpolation during preprocessing and the initial MRI reconstruction, signals in nearby voxels correlate.

The number of independent measurements in spatially smoothed fMRI data can be accurately estimated using the Gaussian Random Field Theory (Friston et al., 1994; Worsley et al., 1992). In this approach, the probability threshold is divided by the number of independent resells (resolution elements) making correction more valid and less stringent.

Acoustic noise in auditory fMRI

Functional MRI is a challenging method for auditory neuroscience due to the acoustic noise. There are two sources of acoustic noise in the imaging environment: the coolant pump and the noise generated by the rapidly shifting gradient coils. The noise poses difficulties for studies using sound stimuli by masking the stimuli, and inducing brain activity that is not related to the stimuli. The masking effects of the gradient noise can be avoided with a “clustered acquisition” paradigm (Hall et al., 1999). It involves imaging a volume of slices in a "cluster" and leaving a quiet interval for sound stimuli presentation between the clusters. The technique improves the sensitivity of auditory cortex fMRI measurements in comparison to continuous imaging (Edmister et al., 1999).

MAGNETOENCEPHALOGRAPHY

When a neuron is active, ions flow in and out of the cell through the membrane creating small currents and magnetic fields in the intra- and extracellular space. The magnetic fields generated by synchronous activation of several tens of thousands of neurons sum up, producing a measurable field outside the head (Hari, 1999). The physiological generator of the MEG signal is the electrophysiological activity of neurons, specifically dendritic postsynaptic potentials. MEG measures mostly currents generated in the apical dendrites of the cortical pyramidal neurons that lie parallel to each other and are tangential with respect to the surface of the head (e.g., in sulci) (Figure 2.3.). However, these fields are extremely weak and their detection requires especially sensitive instruments. The MEG signal is measured in an electrically shielded room using SQUID (Superconducting Quantum Interference Device) sensors immersed in liquid helium. In contrast to fMRI, MEG is a direct measure of neuronal activity and it has sub-millisecond temporal resolution. MEG has a spatial resolution of few millimeters.

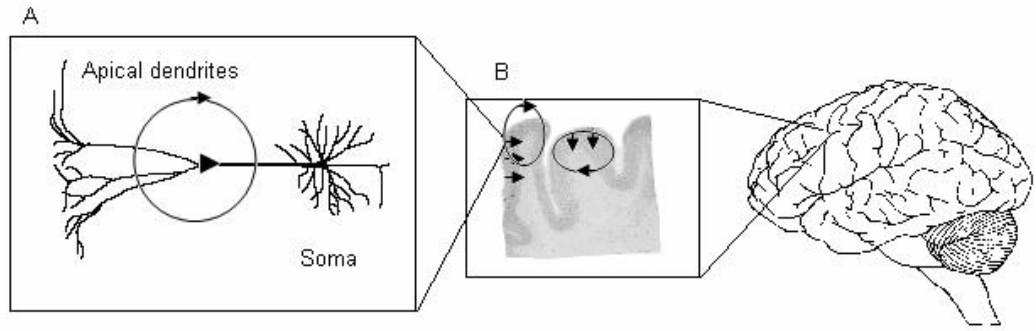

Figure 2.3: A: A typical pyramidal cell of the cerebral cortex. Postsynaptic potential propagating in the apical dendrite is coupled with a magnetic field. B: The MEG is most sensitive to the tangential (sulcus dipoles).

Contrary to the electric potentials measured with EEG, magnetic fields generated by an active neuronal population travel undisturbed through the skull, scalp and brain tissue. Therefore, MEG allows relatively accurate localization of the sources that produced the measured magnetic field. However, the estimation of the current sources in the brain on the basis of the measured signals is complicated by the so-called electromagnetic inverse problem that has no unique solution. Different current distributions inside the brain may produce the same magnetic field patterns outside the skull. Solving the inverse problem requires using assumptions which limit the number of possible solutions. The equivalent current dipole (ECD) is the most commonly used model of the current sources. An ECD (or multiple ECDs) is calculated by minimizing the difference between a calculated and the measured magnetic fields using least-squares search. The ECD has location, orientation and strength.

Chapter 3: Aims of the studies

The aim of this thesis was to investigate the neurocognitive mechanisms of audiovisual speech perception using fMRI and MEG as well as behavioral methods. The specific aims of Studies I-V were the following.

Study I: To study the representations underlying audiovisual integration of speech with a priming paradigm and reaction time measurements (see chapter 4 for details and motivation for the experiments).

Study II: To explore the adaptation of auditory cortical MEG responses to auditory speech by preceding visual speech information.

Study III: To study the primary auditory cortex activation during visual speech observation with fMRI.

Study IV: To reveal the brain areas involved in the processing of phonetic features of audiovisual speech with fMRI.

Study V: To map the network of common processing areas for auditory and visual speech information with fMRI.

Chapter 4: Experiments

SUMMARY OF METHODS

Subjects

In all studies subjects were healthy and had normal hearing and vision (self reported) (see Table 4.1 for details). All subjects were native speakers of Finnish. Prior to participation the subjects gave an informed consent to the protocol that had been approved by the local ethics committee (fMRI studies) in accordance with the Helsinki declaration.

Stimuli

Speech stimulus material was videotaped in a sound attenuated chamber. The speaker was native Finnish male in Study I and female in studies II-V. Sound files (digitized at 44 100 Hz) and visual stimuli (a sequence of bitmap files, frame rate 25 Hz) were extracted from the digital video. The acoustic and visual speech stimuli were presented with Presentation software.

In Study I the experiment was conducted in an acoustically shielded booth with background noise of about 30 db. The auditory stimuli were presented through two loudspeakers located symmetrically in front of the subject on both sides of a monitor which was located 50 cm away from the subjects shoulder. The visual speech stimuli covered an area from the bottom of the talker's nose to the middle of his chin, with no background visible. The visual stimuli subtended a visual angle of 11.3°.

In study II the auditory stimuli were presented at 60 dB over the individually determined hearing threshold to the right ear of the subjects. The second-formant midpoint stimulus (/æ/-/ø/) between /æ/ and /ø/ was created with Praat software.

In studies IV and V the acoustic stimuli were presented binaurally through MR-compatible electrostatic headphones. In study IV, the onset of the acoustic vowel was 95 ms later than the onset of the articulatory lip movement. Such an asynchrony is natural in uttering single vowels.

In studies II-V, the view in the visual speech stimuli was frontal, including the head and the upper part of the shoulders. In the fMRI studies III-V the visual stimuli were back projected to a mirror attached to the birdcage head coil.

Data acquisition

In study II, MEG (VectorView 306-channel system, Neuromag Ltd., Finland) was recorded (digitization rate 600 Hz, passband 0.01–197 Hz) during presentation of the stimuli.

In studies III-V subjects were scanned using 3.0 T GE Signa system retrofitted with Advanced NMR operating console and a quadrature birdcage head coil. For anatomical co-alignment, a T1-weighted volume with a slice prescription similar to the functional volume and a high-resolution whole-head sagittal 3D spoiled-gradient echo-pulse sequence (voxel size 1x1x1.4mm) were acquired during the same imaging session.

In study III, 30 second periods of active and baseline condition were intermittently varied and gradient echo planar magnetic resonance images depicting BOLD-contrast acquired (Time to echo, TE=32 ms; time to repetition, TR=2500 ms; flip angle=90º, 28 contiguous 3.4 mm-thick axial oblique slices; field of view, FOV=22x22 cm, matrix=64x64). The experiment consisted of two 6 min runs (vowels and circles, see Table 4.1 for details), the order of which was randomized across subjects.

In studies IV and V subjects were scanned during a 10 min session, which was repeated three times in study IV and twice in study V. To ensure silence during acoustic stimulus presentation, we used 'clustered volume acquisition' with 2.5 s periods of imaging separated by 3.5 s periods of stimulation, during which no scanner noise was present (Edmister et al., 1999). The scanner's coolant pump was also switched off during the functional imaging sessions. During the 3.5 seconds of stimulation three stimuli of one stimulus type were presented. A stimulus block consisted of two to five consecutive 3.5 s periods of one stimulus type (see Table 4.1 for details).

Imaging parameters in study IV were: slices= 26, slice thickness= 4mm, TE=40 ms, TR=2.5 s, flip angle = 90 degrees, matrix size = 96 × 96, interslice gap = 1 mm, FOV=22 mm. Imaging parameters in study V were: slices= 26, slice thickness= 3.4mm, TE=30 ms, TR=2.5 s, flip angle = 90 degrees, matrix size = 96 × 96, FOV=22 mm.

Data analysis

In study I, wrong responses and outlying RT's (longer than mean ± 3.5 standard deviations) were excluded from each subject's data prior to statistical analyses. The data from the McGurk block (see Table 4.1 for details) was analysed with the Fisher's exact test. The data from the two priming blocks were analyzed separately for the /ba/ and /va/ targets with repeated measures ANOVA. Significant interactions were further analyzed with a Fisher LSD Post-hoc test.

In study II at least 60 artifact-free (i.e., peak to peak electro-oculogram (EOG) and planar gradiometer sensor amplitudes <150 μV and <3000 femtoTesla/cm, respectively) MEG responses were collected and averaged per stimulus category (see

Table 4.1 for details). Response amplitudes were then quantified from the averaged responses to the test stimuli using ECDs fitted in a least-squares sense at the individually determined peak latency of the N1 response using a subset of 34 planar gradiometers over the left hemisphere temporal areas. Between-condition differences in the latencies and amplitudes of the ECD responses were statistically tested using repeated-measures ANOVA.

In studies III-V, fMRI data were analyzed using FMRI Expert Analysis Tool (FEAT) sofware, version 3.1, part of FMRIB's Software library (FSL, www.fmrib.ox.ac.uk/fsl). Pre-processing steps included discarding the two first volumes, non-brain tissue extraction, motion correction, spatial smoothing using a Gaussian kernel with 5 mm full-widht-half-maximum, mean-based intensity normalization of all volumes by the same factor, and high-pass temporal filtering. Time-series analyses were performed using general linear model with local autocorrelation correction (Woolrich et al., 2001). The model used independent predictor for each stimulus condition. In studies IV and V the model was not convolved to a hemodynamic response function, due to the sparseness of the data sampling. Resulting statistical maps were thresholded for subsequent clustering and then a cluster-wise significance threshold was set, corrected for multiple comparisons across the whole acquisition volume (Forman et al., 1995; Friston et al., 1994; Worsley et al., 1992).

In study III each individual's functional images were co-aligned to their own anatomical scans and the data was analyzed only in individual subject level. In studies IV and V subjects' functional images were co-registered with their anatomical scans and a standard brain (Jenkinson et al., 2002; Jenkinson and Smith, 2001). Then a mixed-effects (often referred to as 'random-effects') group analysis was carried out. MNI-coordinates were transformed into Talairach space using Matthew Brett's method (http://www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml). Anatomical regions were automatically determined for within cluster peak activation (study IV) or activations' Center of gravity (Study V) coordinates using the Talairach Daemon v. 1.1 software (University of Texas Health Science Center at San Antonio, TX).

Table 1.

Study	Subjects	Stimuli	Tasks	Method
I	Baseline: N=18, 15 females 16-19 years right handed Audiovisual: N=26 19 females 16-19 years right handed	Baseline: Auditory /va/ and /ba/ ISI=1,5 s Audiovisual: 1) Auditory /ba/ + visual /va/ and vice versa 2) Auditory /ba/ + visual /va/ and vice versa, followed by auditory /ba/ or /va/ 3) Audiovisual /ba/ or /va/ followed by auditory /ba/ or /va/ Durations: /va/=315 ms, /ba/=328 ms (acoustic), 1 s (visual), ISI (prime-target)=580 ms Auditory intensity=50 and 60 dB	Two-choice auditory identification task	Reaction time
II	N=8 right handed	1) auditory /æ/, /ø/ or /æ-ø/ (midpoint between /æ/ and /ø/) followed by auditory /æ/ or /ø/ 2) visual/ æ/ or /ø/ followed by auditory /æ/ or /ø/ 3) Auditory /æ/ or /ø/ Durations: 450 ms (visual), ISI (prime-target)=500 ms, SOA=3,5 s	Two-choice auditory identification task	MEG, 306-channels
III	N=10 3 females, 21-30 years right handed	Run 1 active condition: Visual articulation (a, e, o and y) Run 2 active condition: Expanding and constricting ovals overlaid on a still face Duration=560 ms, ISI=640 ms Baseline condition: still face	Press a key when two identical stimuli were presented in row	fMRI, 3T
IV	N=10 5 females 22-31 years right handed	Matching AV: /a/, /o/, /i/, /y/ Conflicting AV: Acoustic /a/ + Visual /y/, A /y/ + V /a/, A /o/ + V /i/, A /i/ + V /o/ Duration=440 ms (acoustic), 780 ms (visual), ISI=100-400 ms Baseline: still face	Press a key when the type of stimulation changes	fMRI 3T
V	N=13 6 females 20-31 years	Active condition: Auditory vowels (/a/, /o/, /i/ and /y/, the same visual vowels Baseline condition: still face Duration=440 ms (acoustic), 780 ms (visual), ISI=100-400 ms	Press the key when /i/, or target stimulus during baseline is presented	fMRI, 3T

STUDY I: AUDIOVISUAL INTEGRATION OF A PRIME STIMULUS AFFECTS THE PROCESSING SPEED OF A TARGET STIMULUS

Introduction and methods

The purpose of this study was to investigate the representations underlying audiovisual interactions in speech perception using a priming paradigm. We measured the effects of a prior exposure to matching and conflicting (McGurk-type) audiovisual prime stimuli to the identification speed of auditory targets (see table 4.1. for details). The acoustic component of the audiovisual prime was presented at two different intensities.

Behavioral studies of selective adaptation and phonetic context effects have shown that preceding auditory stimulation has a striking effect on the identification of following speech stimuli (Mann, 1980; for a review, see Diehl et al., 2004). The effects are generally thought to be specific to auditory speech (Roberts and Summerfield, 1981; Saldaña and Rosenblum, 1994) but it is not certain whether these effects might be affected by the visual component of an audiovisual stimulus (Bertelson et al., 2003).

The experiment consisted of three consecutive stimulus blocks. The McGurk stimuli were presented alone in the first block to have a measure of the strength of audiovisual integration at both auditory intensity levels for each subject. The following blocks consisted of the prime-target pairs. The subjects were instructed to pay attention to the first audiovisual stimuli similarly as in the McGurk block but respond only to the second auditory stimulus. The subjects' task was to discriminate between the target stimuli.

Results and discussion

As can be seen from Figure 4.1 the conflicting audiovisual stimuli were strongly integrated. The mean proportion of visual responses was 83±5% in the 60-db condition and 78±6% in the 50-db condition (range 45%-100%). The most important finding was that the identification speed of target /ba/ varied as the intensity of the prime's acoustic component varied, but only after exposure to the conflicting prime (Figure 4.2). RT was 81 ms faster in the 50-db than in the 60-db condition (Table 4.2). Furthermore, in the 60-db condition the effect of the conflicting prime was similar to

those of the other prime stimuli. In contrast, the RT in the 50-db condition was about 90 ms faster after the conflicting prime than after the audiovisual /ba/ prime. The interaction of the factors Prime type x Auditory intensity was significant, $F(2,32)=3.3$, $p<0.05$. The mean error rates at the different stimulus conditions showed similar pattern of results.

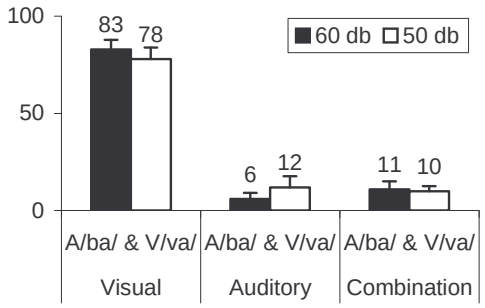

Figure 4.1. Mean proportions of visual, auditory and combination responses to the conflicting audiovisual stimuli. Error bars represent standard error of mean.

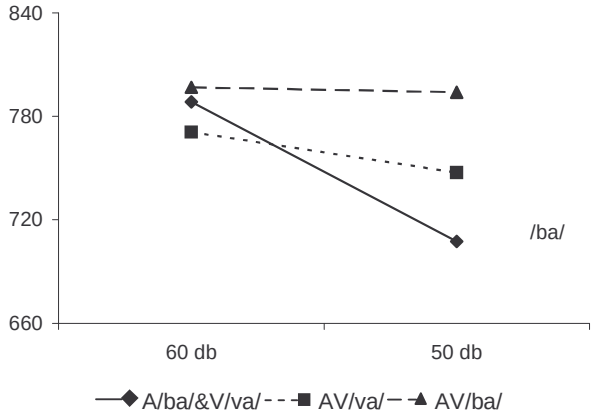

Figure 4.2. The mean RT's to the target /ba/ at the two auditory intensity levels of the conflicting and the two matching prime stimuli. Statistically significant effects were observed only for the target /ba/, not for /va/.

We suggest that the observed effects were due to audiovisual interactions in the processing of the auditory and the visual components of the conflicting prime. Specifically, in accordance with the inverse effectiveness principle (Stein and Meredith, 1993), the visual component of the conflicting prime might have influenced auditory processing more when the intensity of the auditory stimulus was 50 db than 60 db. Increased visual influence might have directly changed the acoustic-phonetic presentation of the prime stimulus, changing therefore the preceding acoustic-

phonetic context and the processing of the auditory target stimulus. Such influence is plausible on the basis of neurophysiological studies demonstrating that visual speech has access to auditory cortex (Calvert et al., 1997; Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991).

STUDY 2: ADAPTATION OF THE NEUROMAGNETIC AUDITORY N1 RESPONSE AMPLITUDE BY VISUAL SPEECH INFORMATION

Introduction and methods

The purpose of this study was to investigate whether seeing a visual articulation causes adaptation of auditory cortex MEG responses to a subsequently presented auditory target speech stimulus (see table 4.1. for details). We hypothesized that the adaptation is larger when the target stimulus is preceded by an auditory than a visual stimulus and that the adaptation effects would be phonetic category specific (e.g., stronger adaptation when /æ/ preceded /æ/ than when /ø/ preceded /æ/).

The target stimulus was preceded (500 ms) by another auditory or visual stimulus or without any preceding stimuli in separate stimulus blocks. The subjects' task was to discriminate between the target stimuli.

Results and discussion

The amplitude of the left-hemisphere N1 (peaking at ~110 ms after stimulus onset) response to test stimuli was significantly suppressed, when the test stimuli were preceded by auditory ($P<0.001$) and visual ($P<0.05$) stimuli, as compared with the responses to the test stimuli when they were presented alone. The effect was significantly greater when auditory stimuli preceded the test stimuli than when the preceding stimuli were visual ($P<0.01$) (Figure 4.3 and 4.4). Category-specific adaptation of the responses was weak and statistically non-significant.

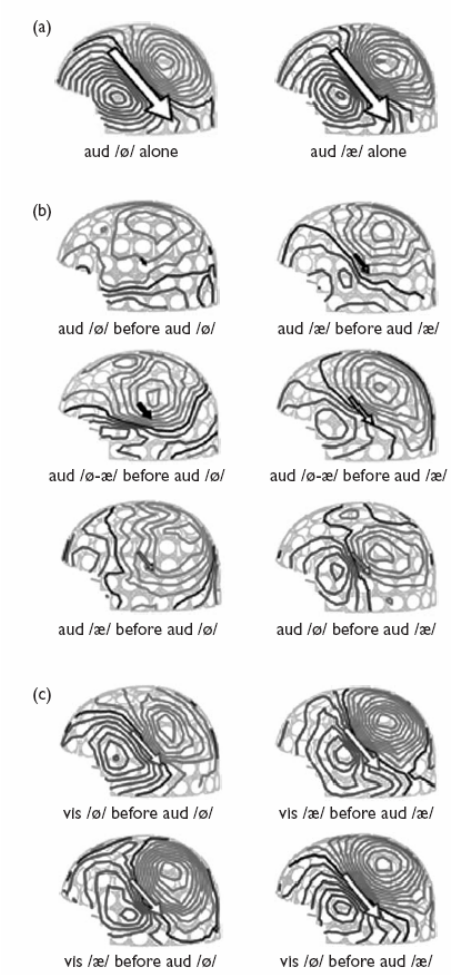

Figure 4.3. Single-subject ECD fits at N1m peak latency with the arrow depicting estimated source strengths and orientations. (a) Responses to the auditory phonemes when presented without preceding stimuli. (b) Responses to the auditory phonemes when preceded by auditory stimuli. (c) Responses to the auditory phonemes when preceded by the visual articulations.

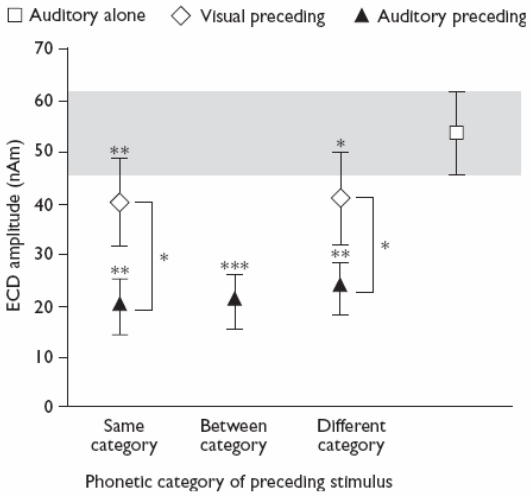

Figure 4.4. Mean (±standard error of mean) amplitude of the MEG responses to the phonetic stimuli when preceded by auditory and visual stimuli, and when presented alone.

These results support the hypothesis according to which seeing the articulatory lip movements of a speaker causes adaptation of feature-specific neurons within the human auditory cortex. The adaptation effect caused by the preceding auditory stimuli was significantly larger than that caused by the preceding visual stimuli, most probably as the result of adaptation of auditory cortex neurons to both phonetic and simple acoustic (e.g., stimulus intensity) stimulus features. The results corroborate previous findings showing that seeing speech influences auditory cortex processing of heard speech (Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991) possibly already at the level of primary auditory cortex (Calvert et al., 1997).

STUDY 3: SILENT LIP-READING ACTIVATES HUMAN PRIMARY AUDITORY CORTEX

Introduction and methods

Silent lip-reading is known to activate auditory cortices and STS (Calvert and Campbell, 2003; Calvert et al., 1997; Paulesu et al., 2003). Whether this effect is limited to the hierarchically higher auditory areas in STG and STS or whether purely visual input can modulate also PAC function has, however, remained an open question (Bernstein et al., 2002).

PAC is anatomically located in the medial half of the transverse gyrus of Heschl (HG) in the temporal lobe. We defined HG in each subject's high-resolution MR images and evaluated signal changes during silent lip-reading inside this area in a per-participant basis. During the MR-imaging the subjects were instructed to fixate their gaze on a marker constantly visible in the mouth region of the face, lip-read the vowels during the active condition and indicate by a button press whenever any vowel occurred twice in succession. During the circles run, they were to detect occurrence of two circles successively moving in the same direction.

Results and discussion

Nine subjects showed significant BOLD signal changes within the left HG during visual speech perception. The activation extended to the medial half of the HG in seven (Fig. 4.5.). Five of the nine subjects had activation also in the right HG. The signal changes extended to its medial half in three.

Six subjects showed significant activation within the left HG (bilateral in one subject) also during watching the moving circles (three in the medial half). All of

them had HG activation also during lip-reading. One subject showed no significant activation during either condition.

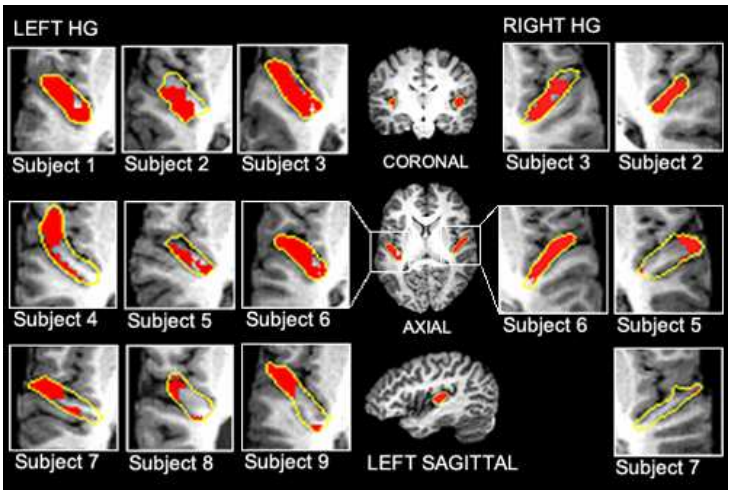

Figure 4.5. Watching speech activates primary auditory cortex. Significant (Z>2.3, p>0.01, corrected), activations during visual speech perception within Heschl's gyri are shown, overlaid on axial MR images. The yellow line outlines Heschl's gyri, the medial parts of which accommodate primary auditory cortex. The middle column displays coronal, axial, and left sagittal high-resolution MR images of subject 6 with overlaid activations.

Our findings demonstrate that PAC in normal-hearing individuals can be activated by visual speech perception, an issue that had remained controversial according to earlier studies (Bernstein et al., 2002; Calvert et al., 1997; MacSweeney et al., 2000). Compared to earlier studies, directly determining HG in each participant may have enhanced the anatomical accuracy of signal-change localization in the present study, since error-sources inherent to standard-space registration could be avoided, and the need to use any atlas was circumvented. We also used a higher field-strength MR-scanner (3 Tesla) and a smaller voxel size (3.4 x 3.4 x 3.4mm) in image acquisition than in previous studies, which may have provided increased sensitivity in BOLD-signal change detection when examining the small anatomical area of PAC.

Several subjects exhibited activation within HG also while watching the moving circles. However, the activation was significantly stronger during visual speech perception than when viewing circles. This, combined with left-hemisphere dominance for visual speech, suggests that the left HG is specifically tuned to phonetic features of visually perceived articulations.

The results support the views about visual speech influencing auditory speech processing at early stages of information processing. As a tentative mechanism, PAC could be receiving visual input through feedback connections from the multisensory

STS region (Calvert et al., 2000; Schroeder and Foxe, 2002), or possibly through direct connections from sub-cortical structures. Alternatively, seeing speech may tune PAC to the acoustic features of speech, via multisensory attentional mechanisms, and this anticipation then modulate processing of simultaneous acoustic input (e.g., the scanner noise) in PAC, even in the absence of heard speech.

STUDY 4: INCREASED ACTIVITY IN BROCA'S AREA DURING PROCESSING OF CONFLICTING VISUAL AND ACOUSTIC PHONETIC INPUTS

Introduction and methods

The purpose of this study was to investigate the processing of the phonetic features of audiovisual speech. We used two types of audiovisual stimuli (matching and conflicting vowels) which differed only with respect to phonetic congruency (see table 4.1. for details). The matching vowels produced a unified audiovisual percept, but the conflicting ones were perceptually clearly incongruous.

Results and discussion

Matching and conflicting audiovisual speech activated the auditory and the visual cortical areas and the inferior frontal, the premotor and the visual-parietal areas bilaterally (Fig 4.6.). Conflicting stimulation evoked significantly greater activity than matching in three left hemisphere areas: Broca´s area (BA44/45), superior parietal lobule (BA7) and prefrontal cortex (BA10). No statistically significant voxels were detected when the BOLD signal during matching stimulation was contrasted to the signal during conflicting stimulation.

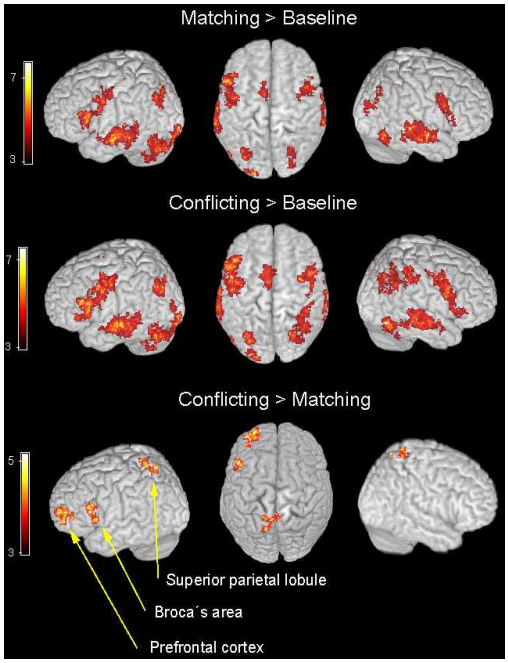

Figure 4.6. Across-subjects z-statistic maps overlaid on an anatomical template (Z>3.0 and cluster-wise p<0.05, corrected for multiple comparisons). Matching audiovisual speech activated the acoustic and the visual cortical areas, as well as the inferior frontal, the premotor and the visual-parietal areas bilaterally (upper panel). Conflicting audiovisual speech caused a similar but more extensive pattern of brain activity (middle panel). The difference in the contrast conflicting > matching AV-stimulation reached significance in three left-hemisphere areas: Broca´s area (BA44/45), superior parietal lobule (BA7) and prefrontal cortex (BA10) (lower panel).

Broca´s area is activated during speech production (Dronkers and Ogar, 2004) and it participates also in various acoustic speech processing tasks (Braver et al., 1997; Burton et al., 2000; Zatorre et al., 1992) and visual speech processing (Callan et al., 2003; Campbell et al., 2001; Nishitani and Hari, 2002; Paulesu et al., 2003; Sekiyama et al., 2003).

We suggest that stronger activation to conflicting than to matching stimuli in Broca's area was due to processing of two instead of a single phonetic input. Neuronal activity related to the processing of unimodal components of the matching audiovisual stimuli probably converged on the same neurons in Broca's area. On the other hand, conflicting components probably activated partly same but also separate neurons.

The activation of Broca´s area and motor speech regions by auditory and visual speech has been argued to be related to the functioning of the mirror neurons. We suggest that Broca's area would contain amodal motor representations of articulatory gestures into which both acoustic and visual phonetic inputs are mapped during audiovisual speech observation, possibly through the activation of the mirror neurons.

We were surprised not to see stronger activation in the left STS for the matching than conflicting audiovisual speech stimulation in the present study. The main difference between our study and that of Calvert et al. (2000) is the nature of the stimuli. Calvert et al. (2000) used meaningful acoustic and visual speech inputs (paragraphs from a book), which were either in temporal synchrony or not. It might be that the left STS is not involved in the processing of phonetic, but rather temporal or semantic, factors of audiovisual speech.

STUDY 5: SPEECH PRODUCTION REGIONS ARE ACTIVATED DURING AUDITORY AND VISUAL SPEECH PERCEPTION

Introduction and methods

In the present study, we specifically aimed at mapping common brain areas processing both auditory and visual speech using the same set of vowels in both modalities as stimuli (see table 4.1. for details). We also tested the hypothesis that auditory and visual speech both activate Broca's area, but its partially different neural pools (Study IV) by comparing the activations' center of gravities (COG) within the common areas. The subjects' task was to press a button every time he/she saw or heard /i/ and when a small white square appeared below the lower lip of the talker during baseline.

Results and discussion

Areas activated by both hearing and seeing speech included: left motor and premotor motor cortex, Broca's area, left inferior parietal area, STG/STS bilaterally, left dorsolateral prefrontal cortex, and left anterior cingulate cortex (Figure 4.7).

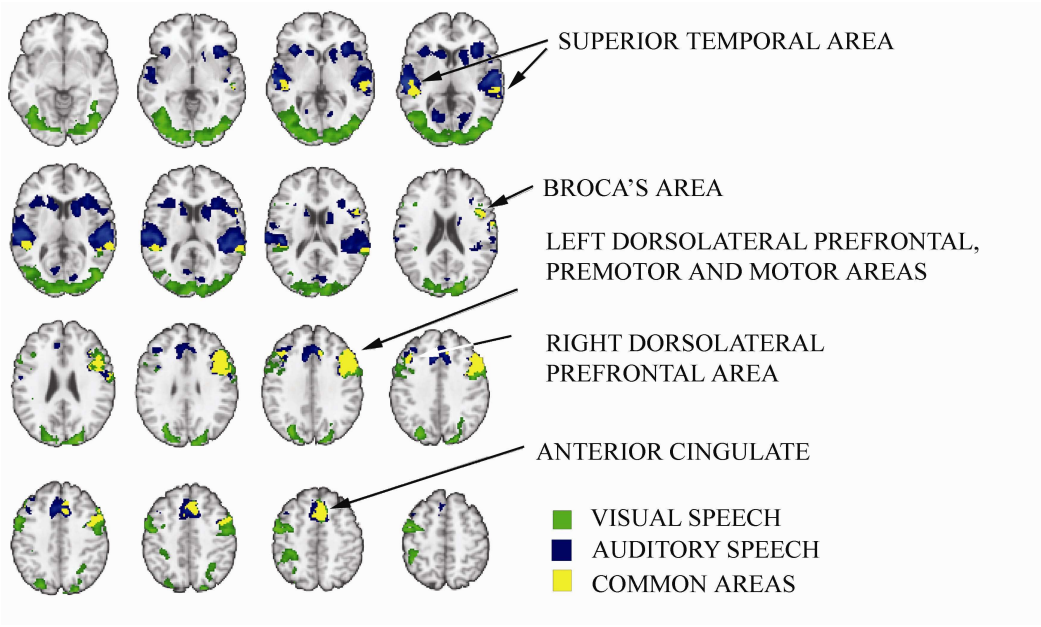

Figure 4.7. Across-subjects z-statistic maps overlaid on an anatomical template (thresholded at Z>2.0, p<0.05, corrected). Brain areas that were activated during lipeading are indicated in green and those activated by listening to speech in blue (compared to the baseline condition). Areas activated by both types of speech perception are shown in yellow.

The COGs during visual speech perception were located more superiorly than the COGs during auditory speech perception within the left premotor/primary motor cortex, Broca's area, the inferior parietal area and the bilateral superior temporal area (Fig. 4.8A). In the left premotor/primary motor cortex, the COGs of all individual subjects were located in the ventrolateral part of the primary motor cortex area 4a (Geyer et al., 1996), which however can not compellingly be differentiated from the premotor area 6 with MRI (Fig. 4.8B). The activations were below the approximate representation area of finger movements, in the representation areas of lip and tongue movements as well as in the more inferior part of area 4a/6 (Fig. 4.8C). During lipreading, most COGs of individual subjects (mean z-coordinate value 43±1.6 mm) were located in the lip and tongue areas. During auditory speech perception, individual COGs (mean z-coordinate value 37±2.6 mm) were more evenly distributed within 4a/6 (Fig. 4.8C). The difference in the mean z-coordinate values was statistically significant, $t(8) = 2.38$, $p < .05$

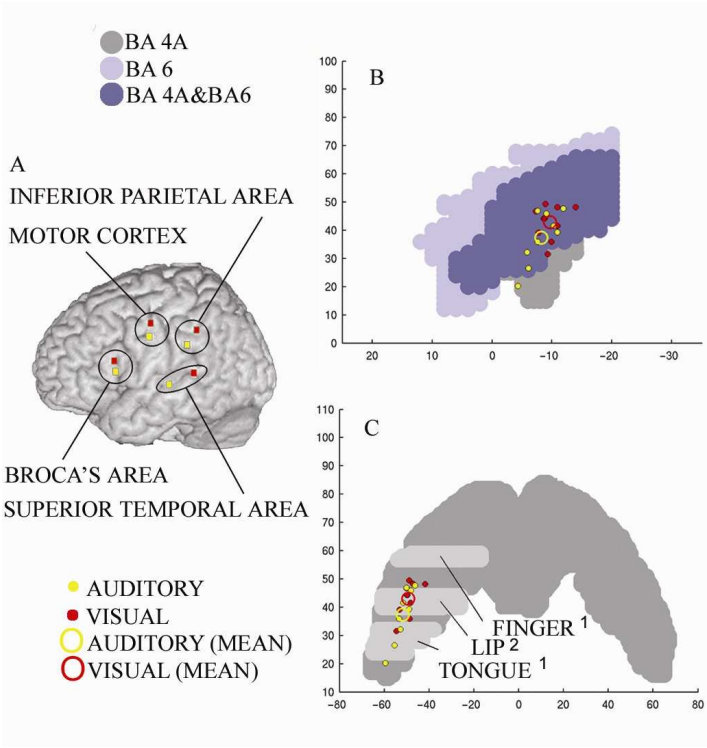

Figure 4.8. A: The activations' center of gravities during lipreading (red) were located superior to those during listening to speech (yellow) in the left motor cortex and in Broca´s area and superior and posterior in the left posterior superior temporal area, and in the inferior parietal lobule. B: Sagittal view of the ventrolateral part of the cytoarchitectoniccytoarchitectural maps of Brodmann motor cortical areas 4a (grey) and 6 (light blue) and their intersection (dark blue) and individual subjects' (small circles) and mean (large circles) COG locations during auditory (yellow) and visual (red) speech perception. C: Coronal view of the cytoarchitectonic map of Brodmann motor cortical area 4a divided in three sections corresponding to the approximate motor representation areas for finger, lip and tongue movement, [1] (Alkadhi et al., 2002), [2] (Hesselmann et al., 2004).

The results corroborate and extend findings from several neuroimaging studies of auditory, visual and audiovisual speech processing which suggest that STS and speech motor areas might be involved in processing both heard and seen speech (e.g., Binder et al., 2000; Callan et al., 2003; Calvert and Campbell, 2003; Wilson et al., 2004). In addition, our data suggest that visual speech activates the primary motor cortex, even though we can not rule out that at least some of these activations are in the premotor cortex.

The COG analysis indicates that visual speech activates systematically more superior parts of the MNS areas than auditory speech. We suggest that this might be related to the difference in information provided by auditory and visual speech. Visual vowels contain information from orofacial gestures, such as movements of the lips and surrounding areas of the face, not of the whole vocal tract. In contrast, acoustic speech provides information of the activity of the whole vocal tract including the non-visible parts. In support, motor-cortex activation during auditory speech perception

was distributed along the lip, tongue and more inferior areas of the area 4a/6, whereas activation during visual speech perception was more concentrated to the lip and tongue areas (Fig. 4.8)

The observed overlap of MNS activation during auditory and visual speech perception suggests that there is an audiovisual neuronal subpopulation, which processes speech input independently of whether it is produced, heard or seen. On the other hand, different COGs within the common areas suggest that there are auditory and visual modality-specific subpopulations as suggested by Study IV and previous research (Keysers et al., 2003; Kohler et al., 2002). Furthermore, the observed somatotopic activation tentatively suggests that auditory and visual speech inputs are mapped to specific motor representations corresponding to the perceived motor origin of the input.

In conclusion, we suggest that the network of brain areas identified here supports the mapping of the auditory and visual speech inputs into motor based phonetic representations, via the mirror neuron system (see also Skipper et al., 2005).

CHAPTER 5: GENERAL DISCUSSION

The current thesis summarizes the results from five experiments which investigated the neurocognitive mechanisms of audiovisual speech perception by using fMRI, MEG and behavioral methods. The results of the Studies I-III corroborate earlier findings on audiovisual integration of speech taking place early on in the auditory processing hierarchy. Study I provided behavioral evidence for this, although the interpretation of the complex behavioral result has to be taken with proper caution. The interpretation is, however, supported by studies II-III, where visual speech was found to modulate the reactivity of the left auditory cortical areas (Study II) and importantly, the PAC (Study III). Studies IV and V provided new findings suggesting that acoustic and visual speech signals might interact in the speech motor regions (Study IV). Furthermore, these regions were tentatively suggested to support the mapping of the auditory and visual speech inputs into common motor based phonetic representations, possibly via the somatotopic activation of the mirror neuron system (Study V). The following sections discuss these main findings.

Audiovisual interaction in auditory cortex

The results of studies I-III corroborate previous findings showing that visual speech has access to the early levels of auditory processing hierarchy (Klucharev et al., 2003; Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991; Besle et al., 2004; Calvert et al., 1997; van Wassenhove et al., 2005) and support the auditory integration models (see Chapter 1). Importantly, Study III disclosed that PAC is activated during visual speech perception. However, these studies cannot disclose where the input to auditory processing stream comes from. Electrophysiological studies in monkeys suggest, that auditory cortex responses to visual stimuli are due to projections from higher cortical regions (Schroeder and Foxe, 2002; Schroeder et al., 2003). However, additional research is needed to find out whether visual speech has access to PAC directly from subcortical structures or visual processing areas or through feedback from STS or other multisensory regions.

The phonetic categorization of speech starts 100-150 ms after stimulus onset (Rinne et al., 1999). The modulation of auditory cortex reactivity by preceding visual speech stimulation in Study II may have occurred at any time between the prime and the target presentation. However, EEG studies show that the auditory N100 amplitude is suppressed during audiovisual speech stimulation in comparison to the sum of

unimodal responses (Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005). These studies indicate that in terms of processing time there are early (within 100 ms from the stimulus onset) audiovisual interactions in the auditory cortical areas. In support, electrophysiological studies in monkeys show that responses to visual stimuli in auditory cortex neurons are very early (~50ms from stimulus onset) (Schroeder and Foxe, 2002; Schroeder et al., 2003). These results suggest that there are audiovisual interactions in auditory cortical areas before the phonetic categorization of the speech input. Interactions occur also during or after the approximate time-window of phonetic categorization (> 150 ms) possibly through feedback to PAC/STG from STS or other multisensory areas (Möttönen et al., 2002; Möttönen et al., 2004; Sams et al., 1991).

Study II failed to reveal phonetic category specific effects of visual speech on auditory cortex reactivity and thus leaved open the possibility that the effect is not specific to speech at all but might be due to any dynamic visual movement. In support, Study III showed that a dynamically moving circle placed above the still face baseline image activated PAC but the activation was not left-lateralized as it was during the visual speech condition. This suggests that dynamic visual non-speech stimuli modulate PAC reactivity to some extent but the left PAC is especially tuned to visual speech information. Therefore it is likely that the left auditory N1 suppression after visual speech stimulus presentation was at least in part due to adaptation of neurons coding phonetic features of visual speech.

The role of multisensory convergence in audiovisual speech integration

Integration of auditory and visual non-speech information is primarily based on temporal and spatial coincidence of the stimuli (Stein and Meredith, 1993). These mechanisms are important in audiovisual integration of speech as well (Macaluso et al., 2004). However, seeing and hearing speech provide also phonetic information. Therefore, both general and speech-specific multisensory mechanisms might be important in audiovisual perception of speech (for recent discussions, see Calvert et al., 2004; Klucharev et al., 2003).

The lack of stronger activation in STS during matching than conflicting AV stimulation in Study IV suggests that STS region might not be involved in the processing of phonetic features of audiovisual speech, although one should be cautious in interpreting negative results. It was suggested in the Study IV that STS

might be involved in processing temporal or semantic factors of audiovisual speech instead.

According to the hypothesis by Calvert et al. (2000), unimodal speech signals are integrated in STS and fed back onto primary auditory areas. This mechanism predicts activation of auditory cortices during visual (Study IV; Calvert et al., 1997) and enhanced activation during audiovisual speech processing (Callan et al., 2004; Callan et al., 2003; Sekiyama et al., 2003). However, the AV interactions in the left auditory cortex might precede those in the right STS (Möttönen et al., 2004) and the audiovisual responses measured with EEG and MEG are often smaller during audiovisual stimulation than the sum of responses during unimodal stimulation (Besle et al., 2004; Klucharev et al., 2003; Möttönen et al., 2004; van Wassenhove et al., 2005).

It has been suggested that rather than in establishing multisensory perceptual representations, the function of the multisensory areas could be the weighting of one sensory stream against the other (van Wassenhove et al., 2005). In support, studies using non-speech audiovisual stimuli have shown suppression of activity in sensory-specific cortices together with enhanced activation of multisensory sites (Bushara et al., 2001; Laurienti et al., 2002). This suggests that there would be reciprocal interaction between multimodal and unimodal areas during audiovisual integration (Bushara et al., 2001).

The vast extent of cerebral areas demonstrating multisensory convergence suggests that it is likely to be a general cerebral processing principle. However, the functional role of multisensory integration in posterior STS or in other multisensory cortical areas is unclear.

Audiovisual integration in speech motor areas

Studies IV and V provide corroborating and new evidence for visual and auditory speech inputs being converted to motor representations during auditory, visual and audiovisual speech perception. One of the underlying neurocognitive mechanisms of audiovisual speech integration could therefore be the convergence of information from seen lip movements and heard speech into a common, specific motor representation matching the observers own speech production. According to this hypothesis the phonetic features of the auditory and visual components of audiovisual speech would be integrated in the multisensory neurons of the speech motor regions

(possibly after multisensory integration of non-speech features in other cortical areas) followed by a motor-based categorization of the speech input and refinement and constraining of sound-based phonetic representations in STG/STS through back-projections. According to this model AV integration in the speech motor areas would precede phonetic categorization as well as constrain and facilitate sound-based phonetic processing. A similar model has been recently proposed by Skipper et al. (2005).

In support, Study IV demonstrated that the speech motor regions are involved in the phonetic analysis of audiovisual speech input. Phonetically conflicting audiovisual stimulation enhanced activation in Broca's area in comparison to matching stimulation. It is possible that there are supra-additive responses during phonetically matching audiovisual speech stimulation in these areas, although these responses were not detected. The increased activation during conflicting stimulation indicates, that phonetic analysis is performed by both multisensory and sensory-spesific neuronal populations within Broca's area. In support, in Study V auditory and visual speech activated overlapping areas of the speech motor regions indicating that there might be multisensory neurons. The interpretation is, however, tentative since one can not differentiate between fMRI signal originating from densely mixed populations of unisensory cells from actual multisensory neurons by overlapping activations (Calvert and Thesen, 2004). However, the existence of multisensory neurons in the speech motor areas is plausible given the results from single-cell recordings in animals (see Chapter 1). Different COGs within the common areas suggests that there are auditory and visual modality-specific subpopulations as suggested by Study IV. Furthermore, the observed somatotopic activation tentatively suggests that auditory and visual speech inputs might be mapped to somatotopically specified motor representations corresponding to the perceived motor origin of the speech input.

According to Calvert et al. (2003) and Callan et al. (2004) both integration in STS and the internal articulatory simulation of the intended speech act of the (visually) observed speaker facilitate auditory speech perception through back-projections to auditory cortical areas. Articulatory simulation of the visual speech input would have a secondary role in audiovisual speech perception and is used to facilitate primary acoustic-phonetic processing especially in sub-optimal conditions (Callan et al., 2004; Calvert and Campbell, 2003). Despite the similarities, this account is different from the one presented above. The important difference is that in these models the visual

speech input is assumed to be processed independently of the auditory input in the speech motor regions. Visual influence on auditory processing is achieved without convergence of A and V inputs into motor representations.

Insights for future research

The main challenge for future research is to further characterize the roles of the sensory specific, multisensory and speech motor systems in cortical audiovisual speech processing. In addition, it is important to identify and differentiate the neurocognitive mechanisms of general and speech-specific multisensory processing.

Combining fMRI and MEG would be beneficial in studying the source and time-scale of feedback projections to the PAC and STG/STS during audiovisual speech perception. Using multiple phased-array surface coil techniques in fMRI for high spatial resolution and better SNR might enable detecting multisensory organization and possibly also supra- and subadditive responses to audiovisual speech stimulation in other brain areas than STS.

Only a limited set of vowels were used as stimuli in the studies. Vowels were used in the studies to focus onto the prelexical, phonetic level of speech perception. Vowels lack several key properties of other types of speech information. At the pre-lexical level consonant-vowel syllables involve short-duration spectral changes, due to the format transitions of consonants. Vowels and consonant-vowel syllables activate the superior temporal areas bilaterally (Scott and Wise, 2004) but consonant-vowel syllables activate the left planum temporale and right STG/STS more than speech sounds incorporating spectral changes of a longer duration (such as vowel sounds) (Jäncke et al., 2002). The activation of the auditory cortical areas and the speech motor regions during speech perception are dependent on the level of phonetic/linguistic processing involved in the stimulation and the experimental task (Binder et al., 2000; Burton et al., 2000; LoCasto et al., 2004; Zatorre et al., 1992) for a review, see (Hickok and Poeppel, 2004). Therefore, parametric experimental designs which manipulate spatial and temporal as well as acoustic/phonetic and linguistic features of the stimuli may allow segregating the hierarchy of brain areas involved in audiovisual integration of speech.

Precise control of the auditory SNR, for example by determining the hearing threshold of the subjects, would improve the comparability of the results of separate

studies. Investigation of speech specific integration mechanisms would require comparing responses to audiovisual speech and non-speech stimulus combinations.

One plausible way to investigate the role of speech motor processing in audiovisual speech perception would be by comparing the timing and sequence of activated brain areas during silent articulation, visual, auditory and audiovisual speech perception. This requires combining fMRI and MEG for accurate spatial and temporal resolution. Another way to directly investigate this would be by measuring the effect of transcranial magnetic stimulation of motor cortical areas to unimodal and audiovisual speech perception. Furthermore, the suggested transformation of heard and seen vocal tract functions into somatotopic activations of the motor homunculus predicts that differences in the place of articulation (e.g., /apa/ vs. /ala/) would be mapped onto different locations of M1 during auditory speech perception and that conflicting visual input might have an effect on this location.

In addition, it might be fruitful to study audiovisual speech processing in specific patient populations. Especially interesting are the children with specific language impairment. SLI-children have problems in either speech production or production and comprehension. They might be worse than their healthy peers in auditory and visual speech perception and possibly have different responses to conflicting audiovisual stimuli (Hayes et al., 2003).

Conclusions

The results of this thesis support the view on the neurocognitive mechanisms of audiovisual speech perception which emphasizes the involvement of multiple, hierarchically organized and mutually interacting brain mechanisms (Calvert and Thesen, 2004). The results add to and extend the evidence suggesting that auditory and visual speech interact in the auditory cortical regions early on in the processing hierarchy. Furthermore, the results indicate that auditory and visual speech inputs might interact in the motor cortical areas involved in speech production and that the cortical mechanisms of coupling motor action execution and perception might be involved in audiovisual integration of speech.

REFERENCES

Alkadhi H, Crelier GR, Boendermaker SH, Golay X, Hepp-Reymond MC, Kollias SS. Reproducibility of primary motor cortex somatotopy under controlled conditions. AJNR Am J Neuroradiol 2002; 23: 1524-32.

Allison T, Puce A, McCarthy G. Social perception from visual cues: role of the STS region. Trends Cogn Sci 2000; 4: 267-278.

Arbib M, Bota M. Language evolution: neural homologies and neuroinformatics. Neural Netw 2003; 16: 1237-60.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 2004a; 7: 1190-2.

Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 2004b; 41: 809-23.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature 2000; 403: 309-12.

Benevento LA, Fallon J, Davis BJ, Rezak M. Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. Exp Neurol 1977; 57: 849-872.

Benson R, Whalen D, Richardson M, Swainson B, Clark V, Lai S, et al. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain Lang 2001; 78: 364-396.

Bernstein LE, Auer ET, Jr., Moore JK, Ponton CW, Don M, Singh M. Visual speech perception without primary auditory cortex activation. Neuroreport 2002; 13: 311-5.

Bertelson P, Vroomen J, De Gelder B. Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol Sci 2003; 14: 592-7.

Besle J, Fort A, Delpuech C, Giard MH. Bimodal speech: early suppressive visual effects in human auditory cortex. Eur J Neurosci 2004; 20: 2225-34.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, et al. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 2000; 10: 512-28.

Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC. A parametric study of prefrontal cortex involvement in human working memory. Neuroimage 1997; 5: 49-62.

Bremmer F, Schlack A, Shah NJ, Zafiris O, Kubischik M, Hoffmann K, et al. Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. Neuron 2001; 29: 287-96.

Bruce C, Desimone R, Gross CG. Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. Neurophysiol 1981; 46: 369-384.

Burnham D. Language specificity in the development of auditory-visual speech perception. In: Campbell R, Dodd B and Burnham D, editors. Hearing by Eye 2: Advances in the Psychology Speechreading and Auditory-visual Speech. Hove, East Sussex, UK: Psychology Press Ltd, 1998.

Burnham D, Dodd B. Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In: Stork D and Hennecke M,

editors. Speechreading by Humans and Machines. Berlin: Springer, 1995: 103-114.
Burton MW, Small SL, Blumstein SE. The role of segmentation in phonological processing: an fMRI investigation. J Cogn Neurosci. 2000; 12: 679-690.
Bushara KO, Grafman J, Hallett M. Neural correlates of auditory-visual stimulus onset asynchrony detection. J Neurosci 2001; 21: 300-4.
Callan DE, Jones JA, Munhall K, Kroos C, Callan AM, Vatikiotis-Bateson E. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. J Cogn Neurosci 2004; 16: 805-16.
Callan DE, Jones JA, Munhall KG, Callan AM, Kroos C, Vatikiotis-Bateson E. Neural processes underlying perceptual enhancement by visual speech gestures. Neuroreport 2003; 14: 2213-2217.
Calvert G, Campbell R. Reading speech from still and moving faces: The neural substrates of visible speech. Journal of Cognitive Neuroscience 2003; 15: 57-70.
Calvert G, Spence C, Stein BE. The Handbook of Multisensory Processes. Cambidge, Massachusetts: The MIT Press, 2004.
Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS. Response amplification in sensory-specific cortices during crossmodal binding. Neuroreport 1999; 10: 2619-2623.
Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, et al. Activation of auditory cortex during silent lipreading. Science 1997; 276: 593-6.
Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Opinion in Biology 2000; 10: 649-57.
Calvert GA, Thesen T. Multisensory integration: methodological approaches and emerging principles in the human brain. J Physiol Paris 2004; 98: 191-205.
Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, et al. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Brain Res Cogn Brain Res 2001; 12: 233-243.
Catani M, Jones DK, ffytche DH. Perisylvian language networks of the human brain. Ann Neurol 2005; 57: 8-16.
Deacon TW. Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. Brain Res 1992; 573: 8-26.
di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G. Understanding motor events: a neurophysiological study. Exp Brain Res 1992; 91: 176-80.
Di Salle F, Esposito F, Scarabino T, Formisano E, Marciano E, Saulino C, et al. fMRI of the auditory system: understanding the neural basis of auditory gestalt. Magn Reson Imaging 2003; 21: 1213-24.
Diehl RL, Lotto AJ, Holt LL. Speech perception. Annu Rev Psychol 2004; 55: 149-79.
Dodd B. Lip reading in infants: attention to speech presented in- and out-of-synchrony. Cognit Psychol 1979; 11: 478-84.
Dronkers N, Ogar J. Brain areas involved in speech production. Brain 2004; 127: 1461-2.
Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM. Improved auditory cortex imaging using clustered volume acquisitions. Hum Brain Mapp 1999; 7: 89-97.

Fadiga L, Craighero L, Buccino G, Rizzolatti G. Speech listening speficically modulates the excitability of tongue muscles: a TMS study. European Journal of Neuroscience 2002; 15: 399-402.
Fadiga L, Fogassi L, Pavesi G, Rizzolatti G. Motor facilitation during action observation: a magnetic stimulation study. Journal of Neurophysiology 1995; 73: 2608-2611.
Falchier A, Clavagnier S, Barone P, Kennedy H. Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 2002; 22: 5749-59.
Ferrari PF, Gallese V, Rizzolatti G, Fogassi L. Mirror neurons responding to the observation of ingestive and communicative mouth actions in monkey ventral premotor cortex. European Journal of Neuroscience 2003; 17: 1703-1714.
Fishman MC, Michael P. Integration of auditory information in the cat's visual cortex. Vision Res 1973; 13: 1415-9.
Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. Magnetic Resonance in Medicine 1995; 33: 636-647.
Fort A, Delpuech C, Pernier J, Giard MH. Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. Cereb Cortex 2002; 12: 1031-9.
Fort A, Giard MH. Multiple Electrophysiological Mechanisms of Audiovisual Integration in Human Perception. In: Calvert G, Spence C and Stein BE, editors. The Handbook of Multisensory Processes. Cambidge, Massachusetts: The MIT Press, 2004: 503-514.
Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE. Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 2000; 10: 77-83.
Friston KJ, Worsley KJ, Frakowiak RSJ, Mazziotta JC, Evans AC. Assessing the significance of focal activations using their spatial extent. Human Brain Mapping 1994; 1: 214-220.
Fu KM, Johnston TA, Shah AS, Arnold L, Smiley J, Hackett TA, et al. Auditory cortical neurons respond to somatosensory stimulation. J Neurosci 2003; 23: 7510-5.
Gallese V, Fadiga L, Fogassi L, Rizzolatti G. Action recognition in the premotor cortex. Brain 1996; 119: 593-609.
Geyer S, Ledberg A, Schleicher A, Kinomura S, Schormann T, Burgel U, et al. Two different areas within the primary motor cortex of man. Nature 1996; 382: 805-7.
Giard MH, Peronnet F. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J Cogn Neurosci 1999; 11: 473-90.
Graziano MS, Gandhi S. Location of the polysensory zone in the precentral gyrus of anesthetized monkeys. Exp Brain Res 2000; 135: 259-66.
Graziano MS, Reiss LA, Gross CG. A neuronal representation of the location of nearby sounds. Nature 1999; 397: 428-30.
Green K, Kuhl P, Meltzoff A, Stevens E. Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. Perception & Psychophysics 1991; 50: 524-36.

Green KP. The use of auditory and visual information in phonetic perception. In: Stork DG and Hennecke ME, editors. Speechreading by Humans and Machines : Models, Systems, and Applications. Vol 150. Berlin: Springer, 1996: 55-77.
Green KP. The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In: Campbell R, Dodd B and Burnham D, editors. Hearing by Eye 2: Advantages in the Psychology of Speechreading and Auditory-Visual Speech. Hove, UK: Psychology Press Ltd, 1998.
Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliot MR, et al. "Sparse" Temporal Sampling in Auditory fMRI. Human Brain Mapping 1999; 7: 213-223.
Hari R. Magnetoencephalography as a tool of clinical neurophysiology. In: Niemermeyer E and Da Silva FL, editors. Electroencephalography: Basic principles, Clinical applications, and Related Fields. Maryland, USA: Lippincott Williams & Wilkins, 1999: 1107-1134.
Hari R, Forss N, Avikainen S, Kirveskari E, Salenius S, Rizzolatti G. Activation of human primary motor cortex during action observation: a neuromagnetic study. Proceedings of the National Academy of Sciences of the United States of America 1998; 95: 15061-15065.
Hayes EA, Tiippana K, Nicol TG, Sams M, Kraus N. Integration of heard and seen speech: a factor in learning disabilities in children. Neurosci Lett 2003; 351: 46-50.
Heim S, Opitz B, Muller K, Friederici AD. Phonological processing during language production: fMRI evidence for a shared production-comprehension network. Brain Res Cogn Brain Res 2003; 16: 285-96.
Hesselmann V, Sorger B, Lasek K, Guntinas-Lichius O, Krug B, Sturm V, et al. Discriminating the cortical representation sites of tongue and up movement by functional MRI. Brain Topogr 2004; 16: 159-167.
Hickok G, Poeppel D. Towards a functional neuroanatomy of speech perception. Trends Cogn Sci 2000; 4: 131-138.
Hickok G, Poeppel D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 2004; 92: 67-99.
Huang J, Carr TH, Cao Y. Comparing cortical activations for silent and overt speech using event-related fMRI. Hum Brain Mapp 2002; 15: 39-53.
Jäncke L, Wustenberg T, Scheich H, Heinze H-J. Phonetic Perception and the Temporal Cortex. NeuroImage 2002; 15: 733-746.
Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002; 17: 825-841.
Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001; 5: 143-56.
Jezzard P, Matthews PM, Smith SM. Functional MRI: An Introduction to Methods: Oxford University Press, 2001.
Jones JA, Callan DE. Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. Neuroreport 2003; 14: 1129-1133.
Jones JA, Munhall KG. The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics 1997; 25: 13-19.
Jäncke L, Wüstenberg T, Scheich H, Heinze H-J. Phoentic perception and the temporal cortex. NeuroImage 2002; 15: 733-746.

Kaas JH, Collins CE. The resurrection of multisensory cortex in primates: Connection patterns that integrate modalities. In: Calvert G, Spence C and Stein BE, editors. The Handbook of Multisensory Processes. Cambidge, Massachusetts: The MIT Press, 2004: 285-293.
Keysers C, Kohler E, Umilta MA, Nanetti L, Fogassi L, Gallese V. Audiovisual mirror neurons and action recognition. Exp Brain Res 2003; 153: 628-36.
Klucharev V, Möttönen R, Sams M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Research. Cognitive Brain Research 2003; 18: 65-75.
Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G. Hearing sounds, understanding actions: action representation in mirror neurons. Science 2002; 297: 846-848.
Laurienti PJ, Burdette JH, Wallace MT, Yen Y-F, Field AS, Stein BE. Deactivation of sensory-specific cortex by cross-modal stimuli. Journal of Cognitive Neuroscience 2002; 14: 420-429.
Lewis JW, Beauchamp MS, DeYoe EA. A comparison of visual and auditory motion processing in human cerebral cortex. Cereb Cortex 2000; 10: 873-88.
Liberman A, Mattingly IG. The motor theory of speech perception revised. Cognition 1985; 21: 1-36.
Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review 1967; 74: 431-461.
LoCasto PC, Krebs-Noble D, Gullapalli RP, Burton MW. An fMRI investigation of speech and tone segmentation. J Cogn Neurosci 2004; 16: 1612-24.
Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. Nature 2001; 412: 150-7.
Lu MT, Preston JB, Strick PL. Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. J Comp Neurol 1994; 341: 375-92.
Macaluso E, Frith CD, Driver J. Modulation of human visual cortex by crossmodal spatial attention. Science 2000; 289: 1206-1208.
Macaluso E, George N, Dolan R, Spence C, Driver J. Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage 2004; 21: 725-32.
MacSweeney M, Amaro E, Calvert GA, Campbell R, David AS, McGuire P, et al. Silent speechreading in the absence of scanner noise: an event-related fMRI study. Neuroreport 2000; 11: 1729-33.
MacSweeney M, Calvert GA, Campbell R, McGuire PK, David AS, Williams SC, et al. Speechreading circuits in people born deaf. Neuropsychologia 2002a; 40: 801-7.
MacSweeney M, Campbell R, Calvert GA, McGuire PK, David AS, Suckling J, et al. Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. Proc R Soc Lond B Biol Sci 2001; 268: 451-7.
MacSweeney M, Woll B, Campbell R, McGuire PK, David AS, Williams SC, et al. Neural systems underlying British Sign Language and audio-visual English processing in native users. Brain 2002b; 125: 1583-93.
Mann VA. Influence of preceding liquid on stop-consonant perception. Percept Psychophys 1980; 28: 407-12.
Massaro DW. Perceiving talking faces. Cambridge, Massachusetts: MIT Press, 1998.
McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976; 264: 746-748.

Meltzoff A. Towards a developmental cognitive science. The implications of cross-modal matching and imitation for the development of representation and memory in infancy. Ann N Y Acad Sci 1990; 608: 1-31; discussion 31-7.
Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu Rev Neurosci 2001; 24: 167-202.
Moelker A, Pattynama PM. Acoustic noise concerns in functional magnetic resonance imaging. Hum Brain Mapp 2003; 20: 123-41.
Molholm S, Ritter W, Javitt DC, Foxe JJ. Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. Cereb Cortex 2004; 14: 452-65.
Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 2002; 14: 115-28.
Morrell F. Visual system's view of acoustic space. Nature 1972; 238: 44-46.
Möttönen R, Järveläinen J, Sams M, Hari R. Viewing speech modulates activity in the left SI mouth cortex. Neuroimage 2005; 24: 731-7.
Möttönen R, Krause CM, Tiippana K, Sams M. Processing of changes in visual speech in the human auditory cortex. Brain Res Cogn Brain Res 2002; 13: 417-25.
Möttönen R, Schurmann M, Sams M. Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neurosci Lett 2004; 363: 112-5.
Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. Percept Psychophys 1996; 58: 351-62.
Narain C, Scott SK, Wise RJ, Rosen S, Leff A, Iversen SD, et al. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb Cortex 2003; 13: 1362-8.
Nishitani N, Hari R. Temporal dynamics of cortical representation for action. Proceedings of the National Academy of Sciences of the United States of America 2000; 97: 913-8.
Nishitani N, Hari R. Viewing lip forms: cortical dynamics. Neuron 2002; 36: 1211-1220.
Olson IR, Gatenby JC, Gore JC. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Research. Cognitive Brain Research 2002; 14: 129-138.
Paulesu E, Frith CD, Frackowiak RS. The neural correlates of the verbal component of working memory. Nature 1993; 362: 342-5.
Paulesu E, Perani D, Blasi V, Silani G, Borghese NA, De Giovanni U, et al. A functional-anatomical model for lipreading. J Neurophysiol 2003; 90: 2005-13.
Petrides M, Pandya DN. Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. Eur J Neurosci 2002; 16: 291-310.
Rademacher J, Caviness VS, Jr., Steinmetz H, Galaburda AM. Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology. Cereb Cortex 1993; 3: 313-29.
Raij T, Uutela K, Hari R. Audiovisual integration of letters in the human brain. Neuron 2000; 28: 617-25.

Rinne T, Alho K, Alku P, Holi M, Sinkkonen J, Virtanen J, et al. Analysis of speech sounds is left-hemisphere predominant at 100-150ms after sound onset. Neuroreport 1999; 10: 1113-7.
Rizzolatti G, Arbib MA. Language within our grasp. Trends in Neurosciences 1998; 21: 188-194.
Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci 2004; 27: 169-92.
Rizzolatti G, Fogassi L, Gallese V. Neurophysiological mechanisms underlying the understanding and imitation of action. Nat. Rev. Neurosci. 2001; 2: 661-670.
Roberts M, Summerfield Q. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Perception & Psychophysics 1981; 30: 309-314.
Rockland KS, Ojima H. Multisensory convergence in calcarine visual areas in macaque monkey. Int J Psychophysiol 2003; 50: 19-26.
Romanski LM, Goldman-Rakic PS. An auditory domain in primate prefrontal cortex. Nat Neurosci 2002; 5: 15-6.
Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 1999; 2: 1131-6.
Saldaña HM, Rosenblum LD. Selective adaptation in speech perception using a compelling audiovisual adaptor. Journal of the Acoustical Society of America 1994; 95: 3658-3661.
Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu S-T, et al. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters 1991; 127: 141-145.
Schroeder CE, Foxe JJ. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Brain Res Cogn Brain Res 2002; 14: 187-98.
Schroeder CE, Foxe JJ. Multisensory Convergence in Early Cortical Processing. In: Calvert G, Spence C and Stein BE, editors. The Handbook of Multisensory Processes. Cambidge, Massachusetts: The MIT Press, 2004: 295-310.
Schroeder CE, Lindsley RW, Specht C, Marcovici A, Smiley JF, Javitt DC. Somatosensory input to auditory association cortex in the macaque monkey. J Neurophysiol 2001; 85: 1322-7.
Schroeder CE, Smiley J, Fu KG, McGinnis T, O'Connell MN, Hackett TA. Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. Int J Psychophysiol 2003; 50: 5-17.
Schwartz J-L, Robert-Ribes J, Escudier P. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In: Campbell R, Dodd B and Burnham D, editors. Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech. Hove, U.K.: Psychology Press, 1998: 85-108.
Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 2000; 123: 2400-2406.
Scott SK, Johnsrude IS. The neuroanatomical and functional organization of speech perception. Trends Neurosci 2003; 26: 100-107.
Scott SK, Wise RJ. The functional neuroanatomy of prelexical processing in speech perception. Cognition 2004; 92: 13-45.
Sekiyama K, Kanno I, Miura S, Sugita Y. Audio-visual speech perception examined by fMRI and PET. Neuroscience Research 2003; 47: 277-287.

Sekiyama K, Tohkura Y. Japanese subjects hearing Japanese syllables of high auditory intelligibility. J Acoust Soc Am 1991; 90: 1797-1805.
Seltzer B, Pandya DN. Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. J Comp Neurol 1991; 312: 625-40.
Shah NJ, Jäncke L, Grosse-Ruyken ML, Muller-Gartner HW. Influence of acoustic masking noise in fMRI of the auditory cortex during phonetic discrimination. J Magn Reson Imaging 1999; 9: 19-25.
Skipper JI, Nusbaum HC, Small SL. Listening to talking faces: motor cortical activation during speech perception. NeuroImage 2005; 25: 76-89.
Spinelli DN, Starr A, Barrett TW. Auditory specificity in unit recordings from cat's visual cortex. Exp Neurol 1968; 22: 75-84.
Stein BE, Meredith MA. Merging of the senses. Cambridge, Massachusetts: The MIT Press, 1993.
Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America 1954; 26: 212-215.
Summerfield Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B and Campbell R, editors. Hearing by Eye: The Psychology of Lip-reading. London: Lawrence Erlbaum Associates, 1987: 3-51.
Sundara M, Namasivayam AK, Chen R. Observation-execution matching system for speech: a magnetic stimulation study. Neuroreport 2001; 12: 1341-1344.
Tiippana K, Andersen TS, Sams M. Visual attention modulates audiovisual speech perception. European Journal of Cognitive Psychology 2004; 16: 457-472.
Wallace MT, Ramachandran R, Stein BE. A revised view of sensory cortical parcellation. Proc Natl Acad Sci U S A 2004; 101: 2167-72.
van Atteveldt N, Formisano E, Goebel R, Blomert L. Integration of letters and speech sounds in the human brain. Neuron 2004; 43: 271-82.
van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 2005; 102: 1181-6.
Watanabe J, Iwai E. Neuronal activity in visual, auditory and polysensory areas in the monkey temporal cortex during visual fixation task. Brain Res Bull 1991; 26: 583-92.
Watanabe M. Frontal units of the monkey coding the associative significance of visual and auditory stimuli. Exp Brain Res 1992; 89: 233-47.
Watkins KE, Strafella AP, Paus T. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 2003; 41: 989-994.
Williams TH, Gluhbegovic N, Jew JJ. The Human Brain: Dissections of the Real Brain. Vol 2005, 1997.
Wilson FA, Scalaidhe SP, Goldman-Rakic PS. Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 1993; 260: 1955-8.
Wilson SM, Saygin AP, Sereno MI, Iacoboni M. Listening to speech activates motor areas involved in speech production. Nat Neurosci 2004; 7: 701-2.
Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 2001; 14: 1370-86.
Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 1992; 12: 900-18.
Vouloumanos A, Kiehl KA, Werker JF, Liddle PF. Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to

speech and nonspeech. Journal Of Cognitive Neuroscience 2001; 13: 994-1005.
Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cerebral Cortex 2003; 13: 1034-1043.
Zatorre RJ, Evans AC, Meyer E, Gjedde A. Lateralization of phonetic and pitch discrimination in speech processing. Science 1992; 256: 846-849.
Zatorre RJ, Meyer E, Gjedde A, Evans AC. PET studies of phonetic processing of speech: review, replication, and reanalysis. Cereb Cortex 1996; 6: 21-30.