HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering

Juuso Tujunen

The Effect of Visual Speech on FM-Sweep Evoked MEG Responses

Supervisor: Professor Iiro Jääskeläinen

Instructor: Researcher Jaakko Kauramäki

TEKNILLINEN KORKEAKOULU **DIPLOMITYÖN TIIVISTELMÄ**

Tekijä: Juuso Tujunen
Otsikko: Visuaalisen puheen vaikutus FM-pyyhkäisyiden aiheuttamiin MEG vasteisiin

Päivämäärä: 28.11.2006 Sivumäärä: 73

Osasto: Sähkö- ja tietoliikennetekniikan osasto
Professuuri: S-114, Kognitiivinen teknologia

Työn valvoja: Professori Iiro Jääskeläinen
Työn ohjaaja: DI Jaakko Kauramäki

Tiivistelmäteksti:

Multimodaalisuus eli useiden eri aisteista tulevan tiedon yhdistäminen yhdeksi yhtenäiseksi havainnoksi on keskushermoston yleinen ominaisuus. Tämä koskee myös audiovisuaalista toimintaa eli kuullun ja nähdyn yhdistämistä. Eräs tunnetuimmista ja vaikuttavimmista esimerkeistä tästä on niin sanottu McGurk-illuusio, jossa tiettyä tavua vastaava ääni ja eri tavua vastaava videokuva puhuvasta henkilöstä aiheuttavat kuuloaistimuksen, joka eroaa näistä kahdesta ärsykkeestä.

Tässä kokeessa tutkittiin visuaalisen puheen vaikutusta formantin kaltaisten sinipyyhkäisyiden aiheuttamiin aivovasteisiin käyttämällä magnetoenkefalografiaa (MEG) tutkimusmenetelmänä. MEG mittaa aivoissa olevien sähkövirtojen pään ympärille muodostamaa magneettikenttää; tästä kentästä päätellään taas aivoissa tapahtuvat aktivaatiot. Visuaalisina puheärsykkeinä toimi joko henkilö toistamassa tavua /ba/, tavua /ga/ tai still-kuva samasta henkilöstä. Auditorisina ärsykkeinä toimi kuusi sinipyyhkäisyä, joiden alku ja lopputaajuudet olivat seuraavat: 200–700 (F1), 400–1800 (F2a), 1000–1800 (F2b), 1600–1800 (F2c), 2200–1800 (F2d) ja 2800–1800 Hz (F2e). Tutkimusoletuksena oli, että kun visuaalinen ja auditorinen ärsyke vastaisivat toisiaan, aktivaatio aivoissa olisi voimakkaampaa tai heikompaa, kun jos ne eivät vastaisi toisiaan. Myös vasteiden latenssit saattaisivat erota toisistaan. Kokeessa tuli aina sarja joko /ba/-, /ga/- tai still-tilannetta videolta, joiden aikana kuului sinipyyhkäisyjä satunnaisessa järjestyksessä. Visuaaliset tilanteet vaihtuivat myös satunnaisesti. Koehenkilöiden tuli aina visuaalisen tilanteen vaihtuessa toiseksi vastata nostamalla sormeaan.

Kokeen lopputulokset olivat ristiriitaiset: kun dataa tarkasteltiin yhtenä kokonaisuutena, mitään yhteisvaikutusta nähdyn ja kuullun ärsykkeen välillä ei havaittu. Kun taas nähdyn ärsykkeen vaikutusta aivovasteisiin tutkittiin eri tilanteissa, havaittavissa saattaa olla tietyissä yksittäistapauksissa esiintyvää modulaatiota vasteiden amplitudeissa, mutta tämä on epävarmaa. Mahdollisia visuaalisia efektejä testattiin useammalla tilastollisella testillä. Eroja aktivaatioista löytyi vasemmalta puolelta aivoja. Mahdollinen interaktioefekti visuaalisen ja auditorisen ärsykkeen välillä oli olemassa, mutta tämän efektin tarkka luonne on epäselvä. Koe kuitenkin paljasti muita ärsykkeisiin liittyviä efektejä. Kuultu ääni vaikutti sekä vasemmassa että oikeassa aivopuoliskossa sekä mitattaessa amplitudia että latenssia siihen, millainen aktivaatio syntyi.

Avainsanat: magnetoenkefalografia, herätepotentetiaali, audiovisuaalinen integraatio

HELSINKI UNIVERSITY OF TECHNOLOGY | **ABSTRACT OF THE MASTER THESIS**

Author: Juuso Tujunen

Title: The Effect of Visual Speech on FM-Sweep Evoked MEG Responses

Date: 28.11.2006 Number of pages: 73

Department: Department of Electrical and Communications Engineering

Professorship: S-114, Cognitive Technology

Supervisor: Iiro Jääskeläinen, Professor

Instructor: Jaakko Kauramäki, MSc(Tech)

Abstract:

Multimodality (combination of information coming from several senses as a unified perception) is a common property of central nervous system. One example of this is combination of auditory and visual information; that is, what is seen and what is heard. One of the better known and most impressive examples of this is the McGurk illusion, where a sound of a syllable and a video picture of a person pronouncing another syllable produce a completely new audio sensation, which is different from the audio and visual stimuli alone.

This experiment examined the effect of visual speech on brain responses evoked by formant like sine wave sweeps using magnetoencephalography (MEG) as a research method. MEG measures the magnetic field outside the head, which is caused by electrical currents on our brains; from this magnetic field the electric current distribution inside the head is then deducted. Visual speech stimuli were either a video of a person pronouncing /ba/, pronouncing /ga/ or a still picture of the same person. Auditory stimuli were six different sine sweeps, with the following initial and final frequencies: 200-700 (F1), 400-1800 (F2a), 1000-1800 (F2b), 1600-1800 (F2c), 2200-1800 (F2d) and 2800-1800 Hz (F2e). The hypothesis was that when auditory and visual stimulus match each other, the activation in the brains would be stronger, than when they do not match each other. In the experiment, a series of /ba/-, /ga/- or still situation came from the video, during which the subject heard sound stimuli coming in a random order. The order of visual series was random. Whenever the visual series changed to another, the subject was supposed to answer by lifting a finger.

The results of the experiment were contradictory: when the data was observed as a hole, no interaction effect between the visual and audio stimulus was observed. When the effect of visual stimuli in different situations was being observed, there might have been some kind of interaction effect present in isolated cases, but this is uncertain. Possible visual effects were tested with several statistical tests. Differences in activations were present in the left hemisphere. A potential interaction effect between auditory and visual stimuli was detected, but the exact nature of this effect remains unclear. The experiment did, however, reveal other effects. The heard sound affected both in the left and in the right hemisphere, with both amplitude and latency to that, which kind of activation occurred.

Keywords: magnetoencephalography, event-related potential, audiovisual integration

Foreword

This thesis was done in the Laboratory of Computational Engineering (LCE) in the Helsinki University of Technology (HUT). The supervisor for this work was Professor Iiro Jääskeläinen and the instructor M.Sc. Jaakko Kauramäki

Thanks for this work go to my supervisor and instructor for their help and expertise on this thesis. I would also like to thank all the people working at the Magnet House; you've been very helpful, whenever I have needed help. Special thanks go to all my subjects, who didn't get any financial compensation for their time, and still were willing to participate in a time consuming experiment. I would also like to thank my parents and other relatives, also, for their support. Finally, I would like to thank all my friends, wherever and whoever you are.

In Espoo, 21st November 2006

Juuso Tujunen

1. INTRODUCTION 1

2. BACKGROUND 3

2.1. Range of hearing 3

2.2. Human ear 4
2.2.1 The outer ear 5
2.2.2 The middle ear 5
2.2.3 The inner ear 6

2.3. Human brains and central nervous system 9
2.3.1 Electric signals in the brain 14
2.3.2 Auditory pathway 15
2.3.3 Audiovisual and speech areas in the brain 18

2.4 Speech 25
2.4.1 Development of speech and audiovisual integration 25
2.4.2 Audiovisual integration of speech 26
2.4.3 Production, characteristics and recognition of speech 28

2.5 Magnetoencephalography 30

2.6 Purpose of the study and specific hypotheses 36

3. METHODS 37

3.1 Subjects 37

3.2 Used stimuli 37

3.3 Proceeding of the experiment 38

3.4 Equipment 39

3.5. MEG analysis 40

4. RESULTS 43

5. DISCUSSION 55

5.1. Summary of the results and some general thoughts 56

5.2 Suggestions for further studies 58

6. BIBLIOGRAPHY 60

APPENDIX A 65

PRESENTATION SCRIPTS USED FOR STUDY 65

A.1 Code for visual stimulus 65
A.1.1 visual.sce 65
A.1.2 visual_2.sce 66
A.1.3 visual_program.pcl 70

A.2 Code for audio stimulus 71
A.2.1 phonmod.sce 71
A.2.2 phonmod.pcl 72

Abbreviations and notations

$\vec{B}$	magnetic field density
$\vec{E}$	electric field
$\vec{J}$	current density
$\vec{J}^{\,p}$	primary current
$\vec{J}^{\,v}$	return or volume current
ε_0	permittivity of free space
μ_0	permeability of free space
ρ	free electric charge density
$\sigma(\vec{r})$	macroscopic conductivity
ANOVA	analysis of variance
AEF	Auditory Evoked Fields
AL	anterior lateral auditory belt
CL	caudal lateral auditory parabelt
CM	caudomedial area
CM	caudomedial auditory belt
CNS	central nervous system
CP	caudal auditory parabelt
dB	decibel
EEG	electroencephalography
EOG	electrooculography
ERP	event related potential
FLMP	Fuzzy Logical Model of Perception
FM	frequency modulation
fMRI	functional magnetic resonance imaging
fT	femtoTesla
Hz	Hertz, unit of frequency (1/s)

IC	inferior colliculus
MEG	magnetoencephalography
MGC	medial geniculate complex
MGN	medial geniculate nucleus
ML	middle lateral auditory belt
N100	negative peak in the ERP approximately 100 ms from the stimulus onset
P200	positive peak in the ERP approximately 200 ms from the stimulus onset
P50	positive peak in the ERP approximately 50 ms from the stimulus onset
R	rostral area
RM	rostromedial region
RP	rostral auditory parabelt
RT	rostrotemporal
RTL	lateral rostrotemporal auditory belt
RTM	rostromedial auditory belt
SEF	somatosensory evoked fields
SEM	standard error of mean
SPL	sound pressure level
SQUID	superconducting quantum interference device
STG	superior temporal gyrus
STS	superior temporal sulcus
t	time
V1	striate cortex
VEF	visually evoked fields

1. Introduction

For a while it has been known that our perception of speech is not only affected by what kind of a sound reaches our ear, but also what we see. What we hear while perceiving speech is a combination of two of our senses: audition and vision. The reason why brains operate this way is quite simple: combination of these two modalities makes speech perception easier when speech is heard in a noisy environment (Sumby and Pollack, 1954) and when the semantic content of speech is difficult (Reisberg et al., 1987). Sometimes this combination of different modalities causes a person to hear something completely different from what is presented either auditorily or visually. This illusionary combination of these two modalities was first discovered by McGurk and McDonald in their famous article “Hearing lips and seeing voices” (McGurk and McDonald, 1976). Roughly, brains form a compromise between what is the visual stimulus and the audio stimulus, and thus a person hears a third, intermediate phoneme from what the visual phoneme and the auditory phoneme are. For example, the syllables ba, da and ga form an acoustic continuum; if, then, a person sees /ga/ and hears /ba/, what he perceives is /da/. This is not the case another way around, that is, when a person sees /ba/ and “hears” /ga/, what (s)he hears is usually a combination of the modalities (instead of fusion, as in the case of seen /ga/ and “heard” /ba/); the heard combinations are gabga, bagba, baga and gaba.

The affect of visual input to speech perception is not restricted to visual speech only. Also written language affects perception of heard speech (Massaro, 1999).

The question is that where, when and how exactly in the brains does this integration of different modalities take place. In this study, specifically signals from the temporal lobes appearing 100 ms after the stimulus onset were being observed; these signals are called the N100 response. Here, the video was a person pronouncing either a syllable /ba/, /ga/, or the same persons still face was shown. The audio signal was such, that in a sense, it was highly simplified from actual syllables (more about this later), sounding like beeps. So the situation was somewhat similar to McGurk effect, although no actual perceptual effect was present here, and the auditory and visual stimuli were not synchronized, which is essential in the McGurk illusion.

Before this topic is further discussed, it is appropriate to go through roughly the structure of cortex (the outer layer of brains), the structure of the ear and also discuss briefly about visual and auditory pathways.

2. Background

2.1. Range of hearing

The human auditory system is able to hear only certain frequencies, and is more sensitive to some frequencies than others. The range of frequencies we can hear is called the *range of hearing*, and in humans this range is about 20-20 000 Hz. Humans are most sensitive to frequencies between 2000 and 4000 Hz; these are the frequencies most important for understanding speech. Figure 2.1, called *audibility curve*, shows the ears sensitivity to different frequencies.

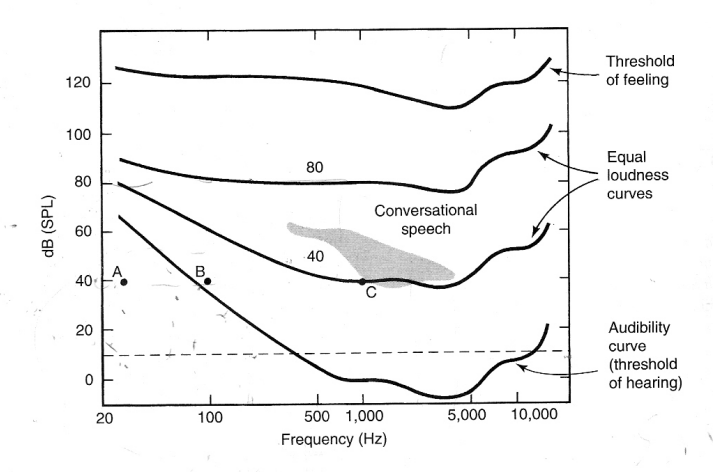

Figure 2.1. Equal loudness curves for human. Curves show as a function of frequency and sound pressure level (SPL) which frequency-sound pressure level combinations are perceived as equally loud. Sounds below the lowest curve can't be heard; they are below the threshold of hearing. Sounds above the highest curve (threshold of feeling) result in feeling of pain, and can cause damage to the cochlea. The area above the audibility curve is called the *auditory response area* because tones falling in this area can be heard (adapted from Goldstein, 2002).

As can be seen from Figure 2.1, the sound pressure level (SPL) alone doesn't determine, how loud a sound is; also the frequency of the sound determines the experience of loudness. *Loudness* is "the magnitude of auditory sensation" (Goldstein, 2002). Also, it has to be noted, that above low SPLs (about 20 dB or so), the loudness of the sound increases linearly with the SPL (the loudness approximately doubles, as the intensity increases 10 dB), but below these SPL:s, the loudness increases faster as a function of intensity (Figure 2.2).

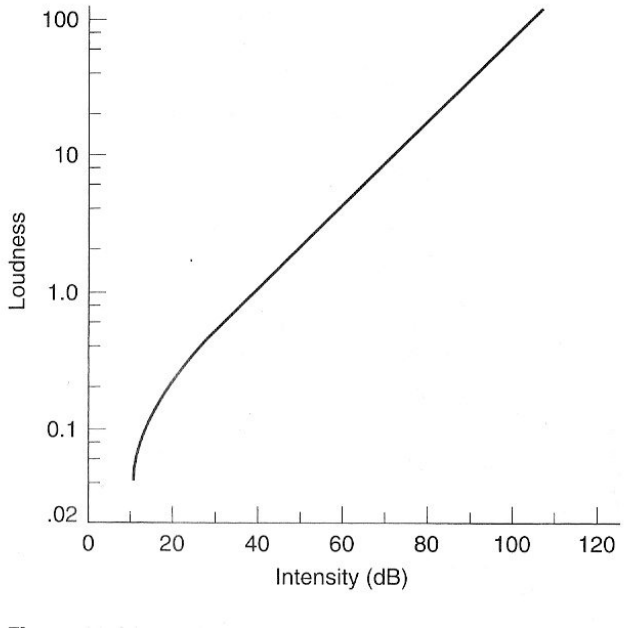

Figure 2.2 Loudness for a 100-Hz tone as a function of intensity (adapted from Goldstein, 2002).

2.2. Human ear

The human ear consists of three regions: the outer ear, the middle ear and the inner ear. For structure of the ear, see Figure 2.3. For a more detailed description of the structure and the functions of the ear see Karjalainen (1999) and Goldstein (2002).

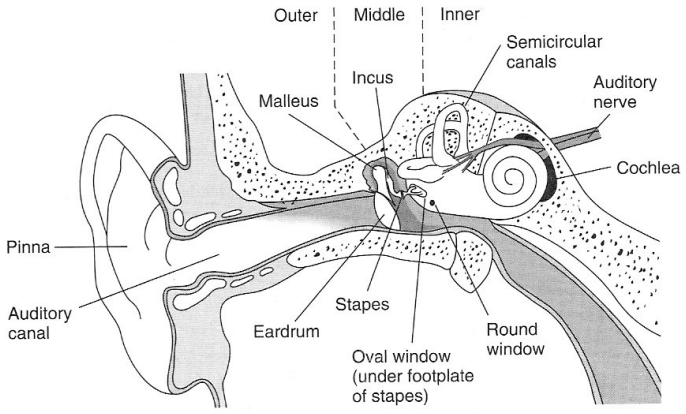

Figure 2.3 Structure of the ear (from Goldstein, 2002)

2.2.1 The outer ear

The outer ear consists of the pinna and the auditory canal; tympanic membrane forms an interface between the outer ear and the middle ear.

The pinna is the most distinguished part of the ear, and it helps us determine from which direction the sound is coming. It is worth mentioning that also heads diffraction affects the acoustical functioning of the ear. It also helps to determine the direction from which the sound is coming

The ear canal works as an acoustic tube, which carries sound wave from the pinna to the eardrum. The ear canal has a resonance frequency at 4 kHz, and thus it enhances the level of that frequency by 10 dB. The eardrum (tympanic membrane) transforms a sound pressure variation into a mechanical pressure variation in the ossicles (Figure 2.3).

2.2.2 The middle ear

The middle ear extends from the ear drum to the oval window at the beginning of the inner ear. Middle ear functions as an impedance adapter between outer and inner ear. This is because inner ear is filled with liquid, which has characteristic impedance of around 4000 times bigger than that of air. Without impedance matching, almost all energy would be reflected back from the inner ear.

There are three small bones (actually the smallest bones in human body) in the middle ear: the malleus, the incus and the stapes. Together they are the ossicles (Figure 2.3). These ossicles reach from the eardrum to the oval window, and mediate pressure signal between these two.

The impedance matching is based on two factors: 1) the eardrum has a larger area than the oval window and 2) the ossicles form a lever. Together these two systems multiply the pressure at eardrum by the factor of 18 at the oval window (compared to

the pressure at eardrum). The increased pressure is a tradeoff between lower particle velocity in the liquid than in the air.

2.2.3 The inner ear

There are two organs in the inner ear: semicircular canals and the cochlea. Semicircular canals do not contribute to hearing; instead, they work as a vestibular organ. The cochlea, however, does an important task in hearing by transforming pressure variations in to neural impulses.

Cochlea is a liquid filled bony structure, which is curled snail-like around itself about 2.7 times (Figure 2.4 a)). The cochlea is divided into two halves from the inside: the scala vestibule (upper half, if the cochlea would be uncoiled) and the scala tympani (lower half). These two parts are separated from each other by the cochlear partition, which almost extends from end-to-end of the cochlea. Base of the cochlear partition is located near the stapes and the apex is located at the far end (Figure 2.4b).

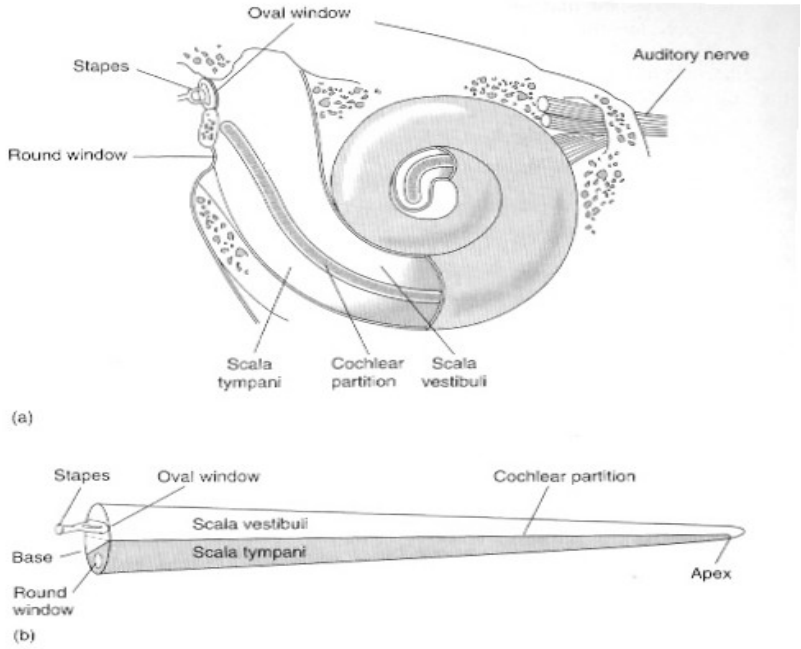

Figure 2.4 Part a) shows a partially uncoiled cochlea. b) A fully uncoiled cochlea (Adapted from Goldstein, 2002).

Figure 2.5 shows a more detailed view of the organ of Corti in the cochlear partition. Movement of the cilia on inner hair cells produces neural signals, which are then send forward by auditory nerves. Vibrations in the stapes at the oval window makes the liquid and basilar membrane inside cochlea vibrate, creating a travelling wave in the basilar membrane. Near the windows the basilar membrane is narrow and light, and at the far end (helicotrema) it is wider and also more flexible. So basilar membrane works as mechanical impedance, which qualities change as a function of location. Due to this, each part of the membrane reacts differently to sounds of different frequencies. The maximum amplitude of the wave is located at the beginning of the membrane with high frequency sounds, and at the end with low frequency sounds (and between them according to the sounds frequency). The wave in question is a *travelling wave* which maximum amplitude value changes based on its current location, this value peaking at the location, which matches sounds frequency on the membrane.

The frequency selectivity of the auditory nerve is better than would be expected based on the functioning of the basilar membrane. There are several explanations for this, but one thing is relatively certain; hair cells affect the vibrations of the basilar membrane via some sort of feedback system, thus increasing the selectivity of the auditory system for different frequencies. More precisely, the outer hair cells react to sound by moving, and this movement (slight tilting and change of length in outer hair cells, called motile response) affects the vibration of the basilar membrane. This movement is tuned to frequency: high frequency sounds cause motile response in the outer hair cells near the base of the basilar membrane, and low frequency sounds in turn cause motile response in the outer hair cells near the apex of the basilar membrane. The outer hair cells push on the basilar membrane, and this pushing amplifies the motion of the membrane and sharpens its response to specific frequencies.

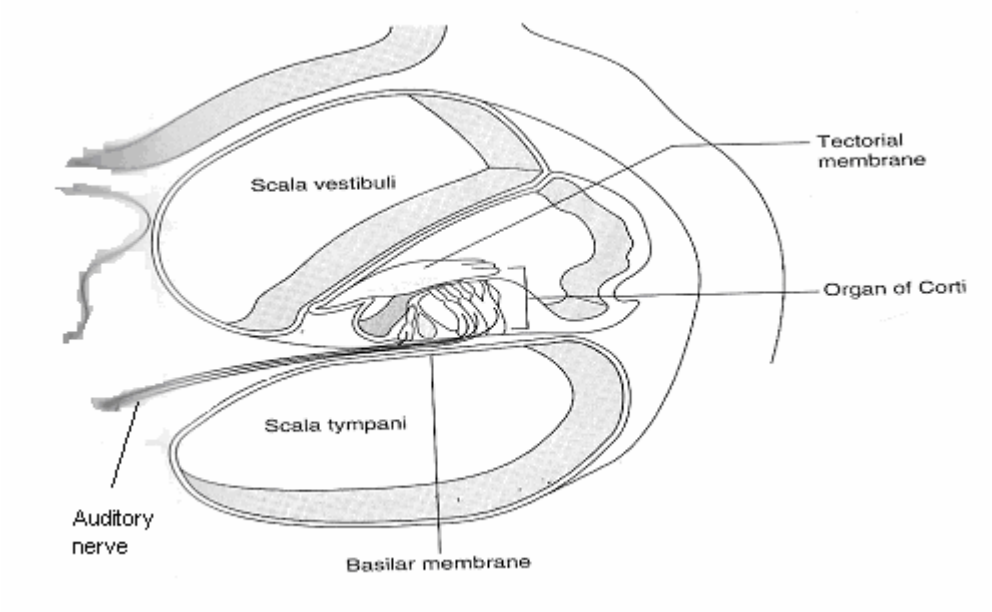

Figure 2.5. a cross section of the cochlea (adapted from Goldstein, 2002)

Hair cells are responsible for picking up the vibrations in the basilar membrane. The hair cells code the intensity of the signal so, that the higher the intensity is, the higher is the hair cells firing rate; however, this process is not linear. Firstly, the hair cells have a spontaneous activity, and secondly, after increasing its firing frequency *approximately* linearly with increasing sound pressure, the cell saturates, and an increase in the sound pressure level doesn't cause increase in the firing frequency (Figure 2.6, Karjalainen).

How exactly do hair cells code the information coming from the basilar membrane so that the spectrum of the sound is preserved? Two different mechanisms work together in the cochlea to accomplish this: place coding and phase locking.

Place Coding refers to a system in the cochlea, which codes frequency based on *which* nerve fibers connected to the hair cells are firing. Place coding was first discovered by Békésy, who proposed place theory of hearing.

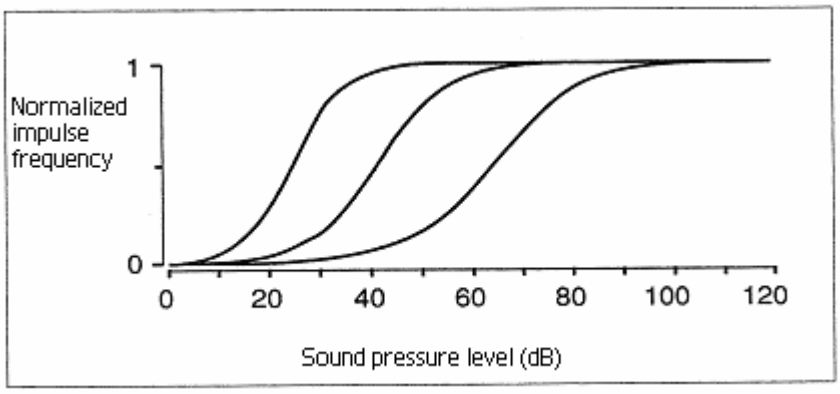

Figure 2.6. Impulse frequency of hair cells as a function of sound pressure level (adapted from Karjalainen, 1999)

Frequency of the sound stimulus can be represented by the timing of neural firing. Nerve fibers fire, in addition to the random firing, when the sound stimulus is at or near the peak of its value. This property is called *phase locking*. When the firing of several nerve fibers is summed up, this leads to a pattern, where maximum firing happens at the peak values of the sound signal (Figure 2.7).

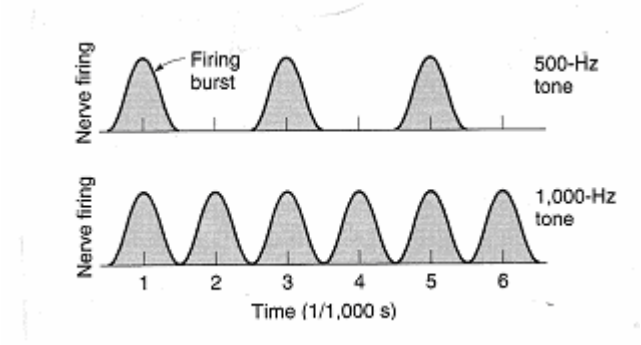

Figure 2.7. Pattern of firing caused by phase locking (adapted from Goldstein, 2002)

2.3. Human brains and central nervous system

Although the cerebral hemispheres (more about them later) are the final stage in the sensory processing, some processing happens already before them. An important processing and especially distribution "station" of sensory information is the *thalamus*, forming *diencephalon* (or *between-brain*) with the *hypothalamus.* The thalamus is located deep inside the brains (Figure 2.8). Almost all the sensory information going to the cerebral cortex is first being processed and distributed by the thalamus; the exception from this rule is the *olfactory system*, which functions so,

that neural signals coming from the nose connect directly to the cerebral cortex (Goldstein, 2002)..

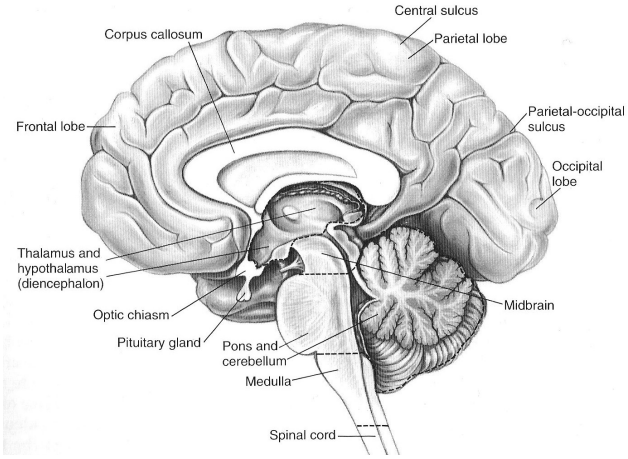

Figure 2.8. Figure shows the location of thalamus and hypothalamus (together forming the diencephalons) and other structures in brains. Adapted from Kandel et al., 1991

The most distinct part of human brains and, on the larger scale, central nervous system (CNS) are *cerebral hemispheres* (for a more detailed description of the CNS, look e.g. Kandel et al., 1991). Cerebral hemispheres consist of the *cerebral cortex*, the white matter beneath cerebral cortex (cerebral cortex consists of grey matter) and three nuclei lying inside cerebral cortex: the *amygdala*, the *basal ganglia* and the *hippocampal formation*. Two hemispheres are separated from each other by interhemispheric fissure, but are connected to each other by *corpus callosum* and other smaller commissures (Kandel et al., 1991). The essential part for processing audio, visual and audiovisual information is the cerebral cortex, although preliminary processing of information coming to these modalities takes place earlier (e.g. in thalamus, as mentioned before).

The cerebral cortex can be divided in to four *lobes*, which each have their own function (Figure 2.9). The lobes are called *frontal, parietal, occipital* and *temporal lobe* according to the cranial bones lying over them. There are, in addition to these four lobes, two other areas in the cortex. The *insular cortex* is located at the medial wall of the lateral sulcus. The *limbic lobe* is located beneath the outer layer of cortex, and its functions are related to learning, memory and emotions. Lobes consist of primary, secondary and tertiary sensory and motor areas, and association areas, which combine information from different sensory cortices, and are also responsible

for higher cognitive functions (language, thinking, emotions etc.). However, one must point out that, according to recent studies (e.g. Laurienti et al., 2002), information integration from other senses may happen already at unimodal areas. Laurienti et al. noticed that visual stimulus causes deactivation, among other temporal areas, in the Brodmann area 41, which is part of the primary auditory cortex. Thus, it would seem that visual input feeds to that area, either directly or by first going to a multisensory area, where the signal is fed back to another unimodal area.

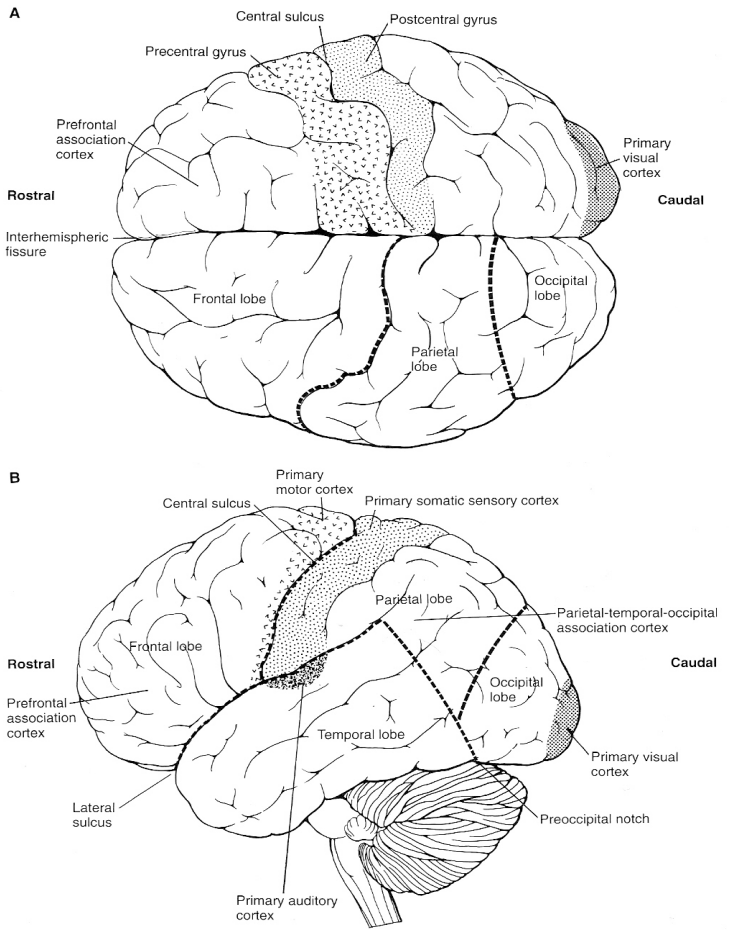

Figure 2.9. Overview of the cerebral cortex (adapted from Kandel et al., 1991).

Primary auditory cortex is located at the temporal lobe, as is secondary auditory cortex and an area directly linked to secondary auditory cortex, Wernicke's area. Other areas located on this lobe include visual-temporal areas, and Brodmann area 38, which deals with emotions (see Figure 2.10) and olfactionary area hidden in interhemispheric fissure.

O*ccipital lobe* has cortical areas involved with vision: *primary visual cortex (V1), V2, V3, V3a, V4, VP, MT and MST.* There is more information about the occipital lobe in the Chapter 2.3.3, in the section “Visual areas”.

Areas located on *parietal lobe* include visual-parietal areas and primary somatosensory cortex. Areas located on *frontal lobe* include motor areas, areas responsible for higher cognitive functions, frontal eye fields and Broca's area (responsible for motoric production of speech). Areas responsible for cognition spread also to parietal and temporal areas located in interhemispheric fissure, as can be seen in Figure 2.10.

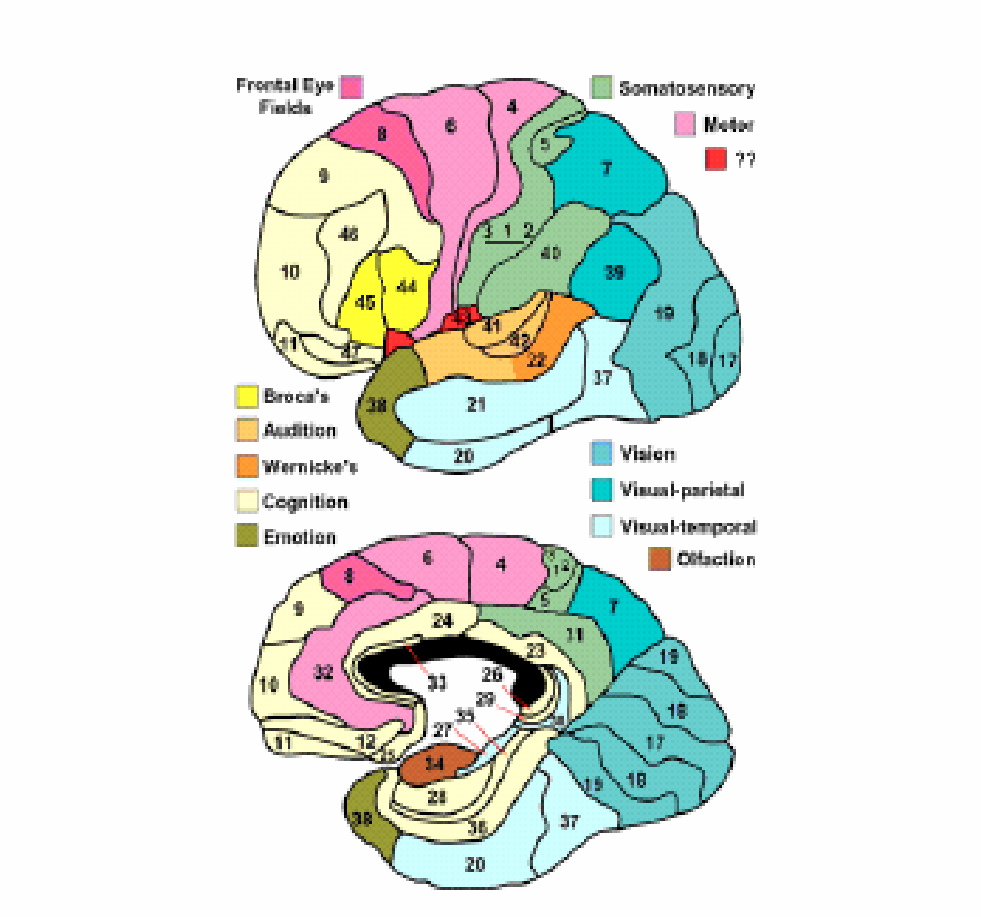

Figure 2.10. Functional and Brodmann areas of the brain (Dubin, 2001)

A brief description of the terminology used in neuroscience is appropriate here, for it helps to understand, what brain areas are described later in this thesis. Figure 2.11 shows the terminology which is used to describe directions and planes in brains. The somewhat complex terminology is due to the fact, that there is an approximately 120° angle between the forebrain and the brainstem; thus, what means e.g. “towards nose” (rostral) in the CNS, is approximately towards nose in the forebrains, and approximately towards top of the head in the brainstem, as can be seen from figure 2.11 b).

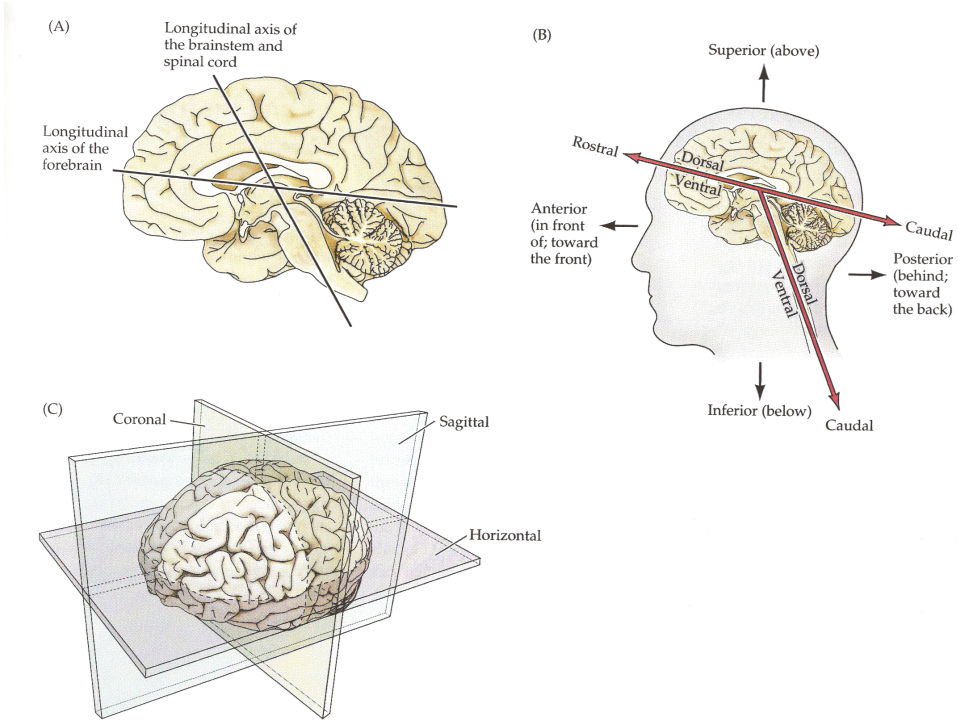

Figure 2.11. Part a) shows the anatomical reason for somewhat complex terminology describing directions in brains. Because humans have evolved to stand upright, this has led to bending of the CNS. Part b) shows the terminology used when describing locations in the brains. Terms *anterior*, *posterior*, *superior* and *inferior* mean that these areas are positioned so compared to the longitudal axes of body; this means, that these directions are same for the forebrain and brainstem. In contrast, terms *dorsal*, *ventral*, *rostral* and *caudal* refer to how areas are positioned compared to the longitudal axes of the *CNS*, so when moving from the brainstem to forebrain, there occurs a fore mentioned bending in directions. It can be seen from the picture, what these directions are for brainstem and for forebrain (dorsal is towards the top of the head in the forebrain etc.), although it may not be obvious from the picture, that rostral is towards the top of the head for brainstem. Part c) shows terminology used, when plains in brains are being talked about (e.g. in the context of brain imaging). (Adapted from Purves et al., 2001)

2.3.1 Electric signals in the brain

Brains operation as a "cognitive processor" is based on the firing of *neurons* in the brains. Neurons consist of three parts: (1) a cell body, (2) dendrites and (3) an axon, or nerve fiber (see Figure 2.12). Cell body contains a nucleus and other structures, which keep the cell alive. Dendrites are responsible for picking up firing from other neurons (or in some cases, from e.g. sensory receptor cells). Axons are responsible for sending neurons signal to other nerve cells (or e.g. directly to muscles). Multiple axons together form a nerve, e.g. in the optic nerve, there is about one million axons.

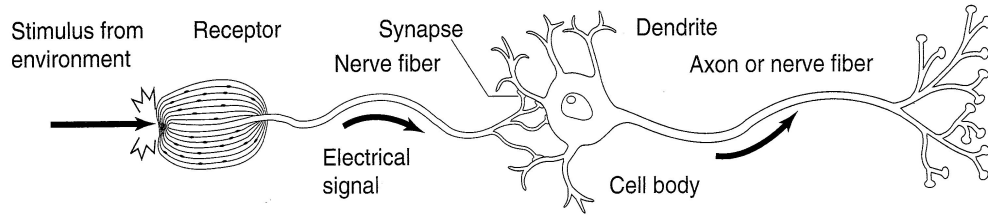

Figure 2.12. A picture of two neurons. The left one is a *receptor cell* and the right one is a typical neuron. The left one is an example of a neuron that is directly connected to the environment, receiving stimuli from it (thus the term receptor) (Adapted from Goldstein, 2002).

Neurons work, in a simplified manner, in the following way. When the neuron is excited strongly enough, it triggers an *action potential*, and this electrical impulse is then send forward to another neuron. The strength of action potential is always the same. However, neurons have a way of signaling the strength of their stimulation. The more the neuron is stimulated, the higher is the frequency with which it triggers the action potentials. The stimulation of the neuron depends on how many *excitatory* and *inhibitory* synapses are being stimulated. As their name suggest, stimulation of *excitatory* synapses *increases* the stimulation of the neuron, and stimulation of *inhibitory* synapses *decreases* the stimulation of the neuron. Whether a synapse is excitatory or inhibitory is determined by what kind of a potential emerges from the stimulation of the synapse. These activations, which are changes in cell membranes potential, are called *postsynaptic potentials.* In more detail, this is how the neuron codes its stimulation. The neuron has a *spontaneous firing rate;* it slowly fires action potentials without any stimulation. After it receives enough excitatory stimulation, a *threshold value* is exceeded, and the neuron starts to increase its firing rate. After that, the neurons firing rate increases, until it reaches a *saturation* (because of the

anatomy of the neuron), after which the firing rate doesn't increase. If the neuron receives more inhibitory than excitatory stimulation, then its firing rate starts to decrease. In this case, the saturation is reached, when the neuron stops firing

2.3.2 Auditory pathway

Auditory pathway starts from cochlea (Figure 2.14). From there electric audio signals are carried along nerve fibers to cochlear nucleus in the brain stem. Leaving cochlear nucleus, nerve fibers go to inferior colliculus, which is located in the midbrain. After inferior colliculus, nerve fibers go to medial geniculate nucleus of thalamus. Finally, after thalamus, nerve fibers end to primary audio cortex. During the path from inner ear to audio cortex, there is interconnection between right and left side pathways: from ventral cochlear nucleus, there is a connection to the opposite sides inferior colliculus; from dorsal nuclear cochleus, there is a connection to the inferior colliculus, and contralateral inferior colliculuses are connected to each other (Figure 2.14). However, auditory pathway has several parallel streams, and some of them bypass inferior colliculus and reach the auditory thalamus directly (Purves, 2001).

The processing of sounds with a particular significance starts already as early as inferior colliculus. More precisely said, many neurons in the inferior colliculus respond only to frequency-modulated sounds, and others respond only to sounds of specific duration (Purves, 2001). Those sounds are typical components of biologically relevant sounds, which in humans naturally includes speech.

Medial geniculate complex (MGC) may be the first location in the auditory pathway, which is selective for combinations of frequencies. Also, MGC may be sensitive for specific time differences between frequencies. These kinds of properties have been found from echolocating bats MGCs and might be present also in humans, to serve e.g. the processing of speech, but this is not known (Purves, 2001). MGC has several divisions. *Ventral division* functions as a major thalamocortical relay (that is, it is a major relay between thalamus and cortex). *Dorsal* and *medial divisions* are organized like a belt around the ventral division. In rhesus monkeys (Rauschecker et al., 1997), different parts of MGC project to different parts of the auditory cortex. The ventral part of the MGC projects to both primary auditory cortex (A1) and rostral area (R).

Other areas of auditory cortex, such as caudomedial area (CM), receive input only from the dorsal and medial parts of the medial geniculate nucleus. Figure 2.13 shows the respective areas (and others) in macaque monkey. The results from monkey experiments are of significance in the understanding of human brain functions, because of close relation between other primates and humans. In fact, studies have shown, that it is possible to establish a direct correspondence between areas at the lateral region of human STG and areas of nonhuman primates (e.g. A1, R, ML (medial lateral belt) etc.).

The auditory system has also efferent, feedback pathways in addition to the afferent, feed forward pathways described in this chapter (Kandel et al., 1991). Auditory cortex, as other parts of the cortex, is divided in to several layers. Layer IV works as an input layer, whereas layer V projects back to medial geniculate nucleus and layer VI projects back to inferior colliculus. Inferior colliculus sends feedback to cochlear nucleus. There is a cluster of cells near the superior olivary complex giving rise to the efferent *olivocochlear bundle*, which terminates in the cochlea (either directly on the hair cells or afferent fibers innervating them). Figure 2.15 shows a schematic presentation of these feedback connections. It is possible, that these feedback connections are important for regulating attention to particular sounds.

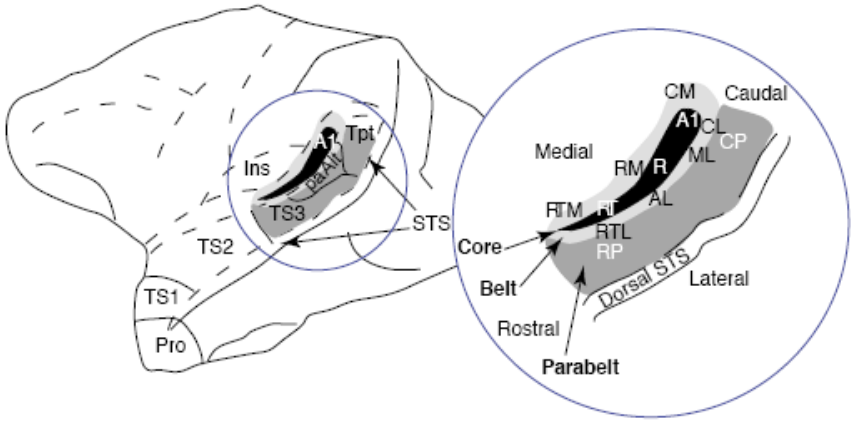

Figure 2.13. Figure shows the location of rostral area (R), caudomedial area (CM), primary auditory cortex (A1) and medial-lateral belt (ML) in macaque monkey (adapted from Scott et al., 2003)

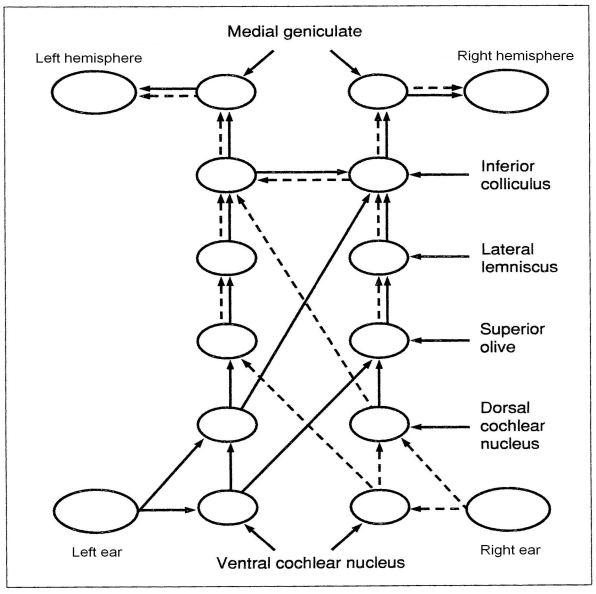

Figure 2.14. Auditory pathway with its crucial interconnections (adapted and modified from Karjalainen, 1999).

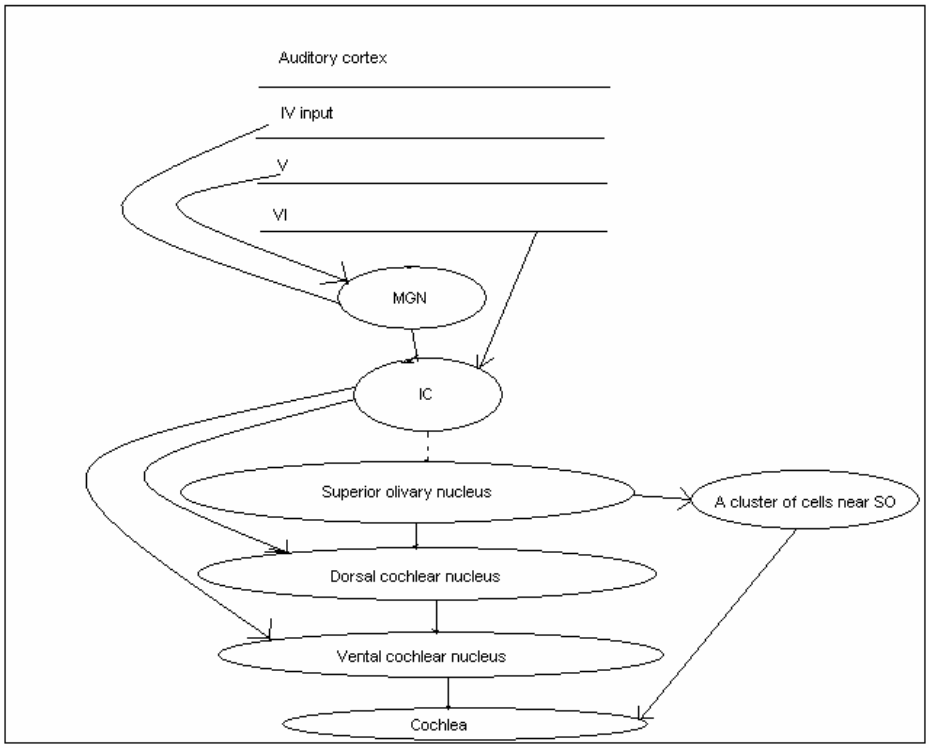

Figure 2.15. Figure shows the efferent connections in the auditory stream. Simple, unified lines (i.e. not arrows) show, that areas are next to each other in the auditory stream. Dashed line between superior olivary nucleus and IC means, that these areas are not consecutive in the auditory stream (lateral lemniscus is located between them). Arrows show the efferent connections in the auditory stream. Figure by author, based on Kandel et al., 1991, and personal communication with Jääskeläinen.

2.3.3 Audiovisual and speech areas in the brain

Multimodal integration is a general function of the nervous system, and so it is also with audiovisual signals, including speech perception. Where and how this integration occurs is not completely clear, although there are some well known brain areas for audiovisual integration.

Visual areas

Visual information is processed first (in cortical areas) in the *occipital lobe*. First the visual information arrives in the cortex to V1 or *striate cortex* as this area is also known. The name *striate cortex* comes from the white stripes (striate = striped) created by nerve fibers that run through it. The striate cortex has six different layers, and consists of several types of neurons, responding to different levels of stimuli (some respond to bars with certain orientation, others to moving corners etc.).

From striate cortex, visual processing proceeds to exstrastriate cortex. Exstrastriate cortex, literally, means areas outside striate cortex. Processing of visual information indeed takes place on all the lobes. Two separate streams begin from the striate cortex; *ventral stream* or *pathway*, which goes from the occipital lobe to the temporal lobe, and *dorsal stream* or *pathway*, which goes from occipital lobe to the parietal lobe. Ventral pathway is also called “what pathway”, because areas in the temporal lobe are responsible for recognition of objects. Dorsal pathway is also called “how pathway” by some scholars (traditionally “where” pathway, but new studies have shown, that this term is somewhat inaccurate (Goldstein, 2002), because areas in the parietal lobe are responsible for detection of place and movement of objects, and also taking appropriate action, e.g. picking up an object). This two pathway model is not the only one that exists; it has been proposed, that brain processes vision in three or possibly more parallel pathways (Kandel, 1991).

Functionality of some visual areas is relatively known, while others are less well known. The exstrastriate cortex is divided into *modules* each dedicated for a specific function (e.g. perception of motion); this quality is called *modularity*.

Auditory areas

Nerve fibers from thalamus, which are related to hearing, arrive at *primary auditory cortex*. Primary auditory cortex is located on the superior temporal gyrus in the temporal lobe. A distinction can be made between primary auditory cortex and other surrounding auditory areas or belt areas. Belt areas are not very well understood, although studies from these areas have been made.

Primary auditory cortex is organized in a tonotopic way; frequencies next to each other are located next to each other. Thus PAC (primary auditory cortex) preserves the tonotopic organization of the cochlea. PAC receives point-to-point input from the ventral division of the medial geniculate complex (thus preserving the tonotopic organization). The belt areas of auditory cortex receive a more diffuse input from the belt areas of the medial geniculate complex, and thus have a less precise tonotopic organization. Area A1 (part of PAC) already responds to species specific calls; this has been confirmed in the case of monkeys, and it could be assumed, that this is also the case with humans. However, selectivity to species specific calls is stronger, when moving up in the auditory hierarchy (towards belt and parabelt areas). At least in cats, spatially tuned neurons (that is, neurons that are sensitive to the direction of the sound) are already found at A1, although the final processing of spatial information takes place elsewhere (Rauschecker, 1998).

PAC is also arranged binaurally in to *stripes.* The neurons in one stripe are excited by both ears, while the neurons in other stripe are excited by one ear and inhibited by the other ear. These stripes alternate in a manner of type1-type2-type1-type2 etc. pattern.

Beyond PAC the processing of sounds is less well known. These areas are thought to be responsible for higher order processing of more complex sounds, especially natural sounds. Some of these areas are specialized for processing combinations of frequencies; others are specialized for processing modulation of amplitude and frequency.

Studies made with both nonhuman primates and humans indicate that the processing of auditory signal is to some extent analogous to processing of other sensory systems (Rauschecker, 1998). That is, the hierarchy of cortical processing is comparable to other modalities. The auditory system, analogous to visual system, can be divided in two pathways (this is analogical to visual system, which also has two pathways for information processing): *dorsal stream* is for processing of auditory spatial information and *ventral stream* is for processing auditory patterns, including speech and music. Studies done with primates would also suggest that auditory cortex is divided in to core areas, belt areas and parabelt areas, each cluster of areas representing higher place in hierarchy of auditory information processing. In macaque monkeys, the number of core areas is 2-3, and there are several belt and parabelt areas. Figure 2.16 shows how these are, based on several studies, connected to each other in macaque monkeys. Also, as mentioned earlier, studies have shown that it is possible to find areas in human cortex matching areas of nonhuman primates' audio areas.

Lateral belt areas in macaque monkeys respond to wide band sound stimuli so, that they are most responsive to stimuli of both certain center frequency and certain bandwidth. Also, lateral belt areas neurons are selective for directions and rates of frequency modulation. This FM selectivity is found throughout the auditory pathways, but is more pronounced in the lateral belt than in other areas (Rauschecker, 1998). Lateral belt neurons are also selective to *species specific calls*, in a way, which can't be explained by mere frequency tuning. Instead, these neurons are selective to combinations of frequencies and temporal order of the signals (e.g. two complex sounds evoke a response only when played in correct order).

Studies done with macaque monkeys indicate, that communication calls (in humans this would include speech) are processed in the anterior and lateral parts of STG. Studies done with humans suggest, that processing of phonemes takes place in the superior temporal region of the brains (Rauschecker, 1998).

The processing of music and speech is lateralized in human brains. There is more about the lateralization of speech in the chapter "Areas responsible for speech and language". Music is lateralized so, that pitch and timbre are presented predominantly

in the right cortex, whereas rhythm sounds are presented more on the left cortex (Rauschecker, 1998).

Especially important area for speech is the *Wernicke's area* lying at the posterior part of secondary auditory cortex, posterior to the primary auditory cortex (Figure 2.10). Based on studies about *aphasias* (or "disturbances of comprehension and formulation of language" (Axer et al., 2000), it is evident that Wernicke's area is important in putting together objects or ideas and the words that signify them. When a brain damage (due to e.g. stroke) in this area happens, these capabilities are compromised, and a person suffering from this aphasia produces fluent, grammatically correct speech, but it has no sensible meaning (this speech is described as so called "word salad"). It has to be noted, that some recent studies show, that Wernicke's area wouldn't have such a prominent role in speech as has been thought before (Gazzanica et al., 2002; Binder et al. 2000). In these studies, superior temporal gyrus (STG), part of which Wernicke's area is, didn't show different kind of activations for speech sounds compared to non-speech sounds (e.g. tones). Gazzanica et al. mention, that this could partly be explained by the fact that in the case of Wernicke's aphasia, when Wernicke (neurologist, who discovered Wernicke's aphasia, named after him) made his original discoveries, other areas beside Wernicke's area had been damaged in the patients.

It is worth mentioning, that although auditory cortex is traditionally thought of as a unisensory area, recent studies (Calvert et. al., 1997) show that this area is actually activated by right kind of visual stimulus (speech or pseudospeech).

Parietal cortex, although not an auditory area, seems to be the final location of auditory spatial processing. Previously, it was thought, that this location was responsible only for visual spatial processing (Rauschecker, 1998).

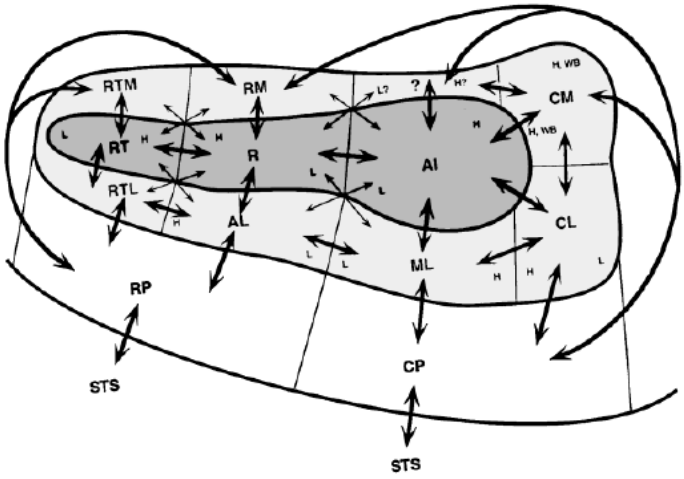

Figure 2.16. A schematic picture of auditory areas and their connections in macaque monkey. Dark grey area matches core areas, light grey belt areas, and white parabelt areas. Heavy arrows indicate strong connections and weaker arrows indicate weaker connections between areas. H means, that area (and the relative location, where the letter is) is responsive for high frequency stimuli, L means low-frequency stimuli and WB means wide band stimuli. Abbreviations stand for the following areas: RT = rostrotemporal; R = rostral area (primary auditory); AI = auditory area I (primary auditory); RTM = medial rostrotemporal auditory belt; RM = rostromedial region; CM = caudomedial auditory belt; RTL = lateral rostrotemporal auditory belt; AL = anterior lateral auditory belt; ML = middle lateral auditory belt; CL = caudal lateral auditory belt; RP = rostral auditory parabelt; CP = caudal auditory parabelt and STS = superior temporal sulcus (adapted from Hackett et al., 1998)

Audiovisual integration areas

Areas responsible for audiovisual integration are less well known then respective unimodal areas, especially primary sensory cortices. However, there are studies identifying these areas.

Superior temporal sulcus (Beuchamp et al., 2004; Callan et al., 2004; Macaluso et al., 2004 and Raij et al., 2000) and middle temporal gyrus (Beuchamp et al., 2004 and Callan et al., 2004) seem to be crucial places for audiovisual integration, although the evidence is not conclusive. Speech integration is centered predominantly to the left cortex (Callan et al., 2004) at the prementioned areas, although Raij et al. report activation in the right STS with audiovisual integration of letters. Some form of audiovisual integration is also possibly found at other parts of the brain, e.g. inferior parietal lobule (Macaluso et. al., 2004).

Prefrontal cortex may be an important place for audiovisual integration (Rauschecker, 1998). Tracer studies show, that injections to belt areas lead to labeling in prefrontal cortex, showing, that belt areas are connected to prefrontal areas. It is to be assumed, that visual-auditory associations are initially formed in these areas.

Studies done with several different species of mammals show that the bordering areas in brains between unisensory areas are important for multisensory integration of unisensory areas, which are connected to them (Wallace et. al, 2004). The importance of these bordering areas in multimodal integration is well known for humans also. An important association area is the parietal-temporal-occipital association cortex, which is located at the junction of the lobes for which it is named. This area is concerned with higher perceptual functions considering the three sensory areas it is bordering (that is, vision, hearing, and somatic sensation). However, it must be pointed out, that at least with the experiment carried by Wallace, few multisensory neurons were present also in the unisensory areas.

Areas responsible for speech and language

Naturally speech comprehension requires, when we are talking about “regular” speech and e.g. not about signing language or lip-reading, that the auditory system is not damaged. However, brain has areas which are distinctively responsible for speech. Damage to these areas may leave auditory system and motor system of the mouth intact but severely disturb speech perception and production.

With large majority of people (97%; Purves et. al, 2001), language functions are located predominantly in the left hemisphere. With the rest of the people, these functions are usually located in the right hemisphere and with few individuals, in both hemispheres. Right side dominance of language is much more common with left handed than right handed people, although a large majority of left handed have left hemisphere dominance in language functions. While left hemisphere is more important in the comprehension and production of syntax, lexicon and semantics of speech, right hemisphere is more important in the emotional coloring of the words (that is to say, which tone of voice was used in a conversation etc.). Damage to areas

on the right side which match Broca's and Wernicke's areas on the left side leads to disturbances in understanding and producing normal emotional and tonal components of speech; these disturbances are called aprosodias.

Figure 2.09 and Figure 2.10 show areas which are important for production and understanding of the speech. Primary areas (Figure 2.09) are not specialized for speech, but are nonetheless important in it for obvious reasons (e.g. primary auditory cortex is required to hear anything in the first place), whereas Broca's area and Wernicke's area (Figure 2.10) are areas specialized in speech. In addition to the pre mentioned areas, also association sensory and motor areas are important in processing language.

Wernicke's area and its role in speech production and comprehension was discussed earlier, when auditory cortex was being discussed. Whereas Wernicke's area is important to the comprehension of speech, Broca's area is important in the production of speech. Damage to this area leads to Broca's aphasia, which symptoms are disturbances in the syntax, grammatics, structure and general fluency of speech. However, the person can still understand speech and can express himself somewhat sensibly, although with difficulties. Table 2.1. summarizes the effects of and differences between Wernicke's and Broca's aphasia.

Broca's and Wernicke's areas were discovered already in the late 19th century. Since then, more knowledge of the brains have been obtained, and it is now obvious that also other areas are responsible for language, and damage to these areas also results in language deficits, although they are more subtle than Broca's or Wernicke's aphasias.

Table 2.1. Table summarizes differences between Broca's and Wernicke's aphasia (adapted from Purves, 2001).

	Broca's aphasia	**Wernicke's aphasia**
	Halting speech	Fluent speech
	Repetitive (perseveration)	Little repetition
	Disordered syntax	Syntax adequate
	Disordered grammar	Grammar adequate
	Disordered structure of individual words	Contrived or inappropriate words

2.4 Speech

2.4.1 Development of speech and audiovisual integration

There is a wide consensus of the stages that occur during the first two years of speech development in infants (Kuhl et. al., 1996). List of these stages (from Kuhl. et. al., 1996) is as follows: reflexive phonation (0-2 months), reflexive or vegetative sounds (e.g. crying) predominate; cooing (1-4 months), infants produce quasivocalic sounds resembling vowels; expansion (3-8 months), clear, fully resonant vowels and a wide variety of new sounds (e.g. screams) occur; canonical babbling (5-10 months), infants produce consonant-vowel syllables (e.g. mama) and, finally, meaningful speech (10-18 months), infants mix meaningful speech and babbling.

Somewhat different division of speech development in infants has been introduced in the book "The child's path to spoken language" (Locke, 1993), although it has to be noted, that this division is originally introduced in a study from as early as year 1980. This division has six stages (from Locke, 1993), and they are as follows. Phonation stage (0-1 months): Nondistress sounds (this excludes e.g. crying) are characterised by an open vocal tract and lack of oral closure and linguistic and mandibular movements. GOO Stage (2-3 months): Crude syllables appear, initiated by closures perceptibly resembling voiced velar stops (e.g. [g]). Expansion Stage (4-6 months): Vocal behaviour diversifies; vocal like sounds start to emerge, and a variety of less speech-like sounds, like squealing. At this stage, marginal babbling may be present. Reduplicated babbling stage (7-10 months): Onset of babbling meaning that the infant produces well-formed syllables. Syllables are in the form of consonant-vowel, and are produced repetitively (like dadada). Variegated babbling stage (11-12 months): Babbling diversifies, and now multisyllabic strings include several kinds of syllables (e.g. [daba]). It has been proposed, that aforementioned two babbling stages overlap, and may actually constitute a single stage.

Already at the infancy both the face and the voice of a speaker (usually mother) work together in the infants phonetic learning (Locke, 1993). There are some indications

that audiovisual integration of speech occurs already at the early infancy (even at the age of 4 moths), but these results are somewhat unclear, and in any case, the integration effects are not as strong as with adults (Desjardins et. al., 2004). It has been proposed, that the development of audiovisual speech integration in fact goes beyond the age of 12 (Hockley et. al., 1994). More precisely, with increasing age, the influence of auditory part of the speech decreases (when dealing with audiovisual speech), but the influence of the visual part and the integration of audiovisual information increases with age.

It would seem that infants (3-months old in this case) brains show a partial preference of mother tongue over other languages, although adults show preference for mother tongue in brain areas, in which infants do not (e.g. STS). The results of this particular study (Dehaene-Lambert et. al., 2002) indicate that infants have already learned something about the prosody of their native language at the age of three months.

2.4.2 Audiovisual integration of speech

There is no clear consensus on how and where in the brains audiovisual integration of speech takes place. Instead, there are several competing models.

Models can be divided roughly in two categories, according to the assumed level of integration and the stage of information processing, at which the integration takes place (Möttönen, 2004); (1) The early integration models: Audiovisual integration occurs before phonetic categorization. (2) The late integration models: acoustic and visual speech information is processed separately up to the phonetic level, where the integration takes place. Below is a list of different models (adapted from Möttönen, 2004, except the section about motor theory).

According to a *fusion model* (Robert-Ribes et al., 1998), the brains combine information from different modalities in a way, that the more reliable modality in a given context is the dominant modality (instead of audio being the dominant modality, as is assumed by some other models). In any case, this leads to a result that audio-visual detection is always at least as efficient and with noisy signals, more

efficient than audio or visual detection alone. For more detailed analysis, see (Robert-Ribes et al., 1998).

Direct realist theory (Fowler, 1996) assumes that vocal tract gestures are detected from acoustic signal, and these gestures are in the core of speech perception. This contrasts with *acoustic theories* of speech (Diehl and Kluender, 1989), which postulate that acoustic features are the objects of speech perception, not articulatory gestures.

In the *motor theory of speech perception (*Liberman and Mattingly, 1985), there is a special language module, which is responsible for both speech perception and production. According to motor theory, both speech perception and production are inherently motoric. When the speech is being perceived, it happens automatically and effortlessly with aforementioned module that is designed for it. This module detects from the speech signal the intended gestures of the speaker in the vocal tract (not actual gestures, because there is considerable overlapping in gestures from one phoneme to the next).

According to *Fuzzy Logical Model of Perception (FLMP)* (Massaro, 1999), speech perception is a form of pattern recognition, analogical (to some extent) to such tasks as recognition of faces. According to FLMP, in recognition (in general) brain uses multimodal perception in an optimal matter; by a statistically optimal integration rule. Also other modalities beside vision and hearing are used, and also higher order knowledge (e.g. when recognising meaningful sentences). FLMP is a late integration model, having four different stages (where integration is the second): 1) evaluation, 2) integration 3) assessment and 4) response selection.

TRACE is an interactive activation model. In TRACE it is assumed that a large number of simple units are connected to each other with excitatory and inhibitory connections. Parallel units are connected to each others with inhibitory connections. There is thus competition between these units. There are three levels of units, which are connected consecutively to each other: features, phonemes, and finally words. Features are connected to phonemes, and phonemes are connected to words. These connections are reciprocal and excitatory. TRACE is originally concerned only with

spoken language, but this model can and has been expanded to cover audiovisual speech. TRACE makes good predictions in multitude of different experiments, but is computationally extremely expensive and therefore unrealistic (considering even brains huge capacity).

2.4.3 Production, characteristics and recognition of speech

As mentioned in this thesis, Brocas area in the brains is crucial for the production of speech, and naturally motor cortex must function properly. According to motor theory of speech, there is a special module in the brains, which is responsible for both *production* and *recognition* of speech (Liberman and Mattingly, 1985). From the motor cortex, then, neural impulses are fed to *muscles* responsible for speech production.

The *vocal organs* are a complex system responsible for the acoustical production of speech. Figure 2.17 shows a schematic presentation of the vocal tract and functional roles of its different parts.

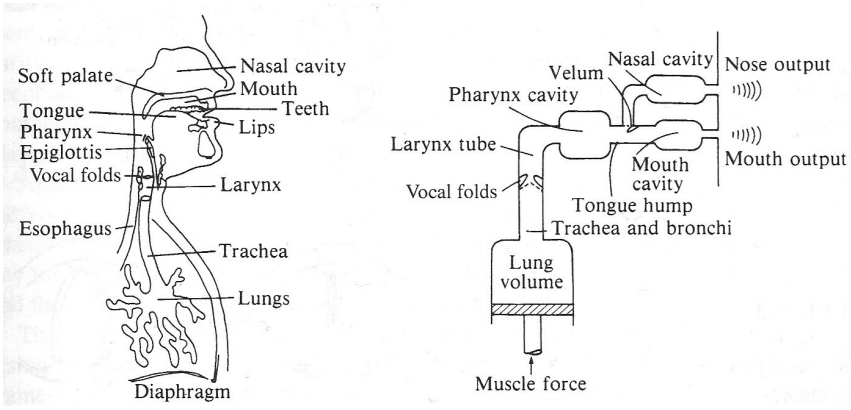

Figure 2.17. Figure shows a schematic presentation of the vocal organs and functional roles of its different parts (adapted from Rossing, 2002).

The functioning of vocal tract as a whole is rather complex. A schematic figure of a mechanical model of the vocal organs is presented in Figure 2.18 (Karjalainen, 1999). The lungs work as a pressure source, from which the air goes to the larynx. At larynx are the vocal folds, which can be moved by muscles attached to them. There is a crack between the vocal folds, which is called glottis. The size of this crack can be changed. This system at larynx is responsible for the production of *voiced* sounds. When airflow from the lungs goes through the glottis, the vocal folds begin to

vibrate. This is also called *phonation.* The frequency of the resonance is the fundamental frequency of the speech (e.g. for males approximately 120 Hz).

After the pressure pulses leave glottis, they arrive at pharynx. From there, the pulses travel to oral cavity and/or to nasal cavity. The shape of these cavities and the fact, whether these cavities (or which one of them) are open or not, shapes the sounds. The effect of these phonation organs is called *articulation*.

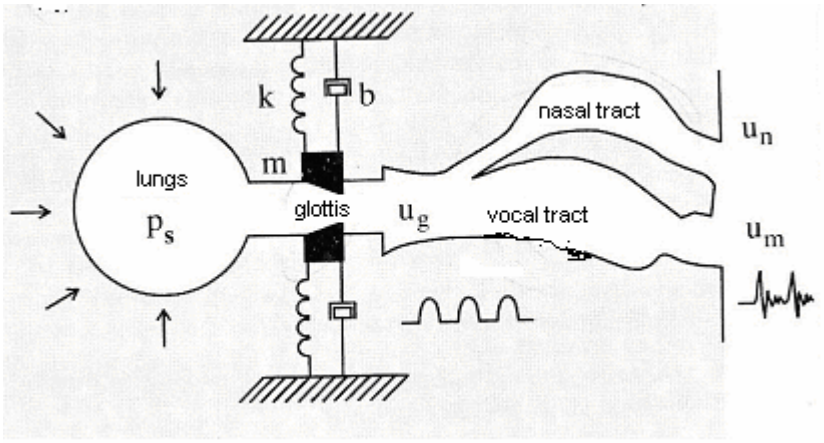

Figure 2.18. A schematic picture of a mechanical model of the lungs. The lungs work as a pressures source. The lungs use either the glottis as an "oscillator", or the constriction part of the tract as a noise generator, in the sound production. Together the vocal and nasal tracts form an "acoustic resonator", which "modulates" the spectral structure of the produced speech (adapted from Karjalainen, 1999).

The acoustic tube formed by larynx, pharynx and oral cavity is called the *vocal tract*, and changing the oral cavity to nasal cavity, one gets the *nasal tract*. The vocal and nasal tracts function as modifiers of the glottal impulse. The vocal tract causes resonances to the sound. These resonances are called *formants.* A formant means in practice, that acoustic energy is concentrated around some frequency. The position and movements of the tongue affect the properties of the tract, as do the movements of the lips and the jaw.

The division of different speech sounds is rather complex, so details are leaved out in this thesis. However, it is in order to go through some basic concepts here. Different speech sounds can be, and humans do this automatically, categorized as different phonemes. A phoneme (lets say /a/) always sounds the same (in a sense, that is; it

always sounds like /a/, although of course the height of the voice might differ, which sounds different etc.), although the acoustical signal may vary considerably. There are cues, based on which humans can differiate different phonemes from each other, although they are not fully understood; more about this later. Phonemes can be divided in to two categories: vowels and consonants. Vowels are phonemes that are always voiced (i.e. they are produced with the vocal folds in vibration). Vowels in Finnish language are as follows: /a, e, i, o, u, y, ä, ö/. Consonants may be either voiced on unvoiced. Consonants in the Finnish language are as follows: original Finnish consonants are /d, h, j, k, l, m, n, ŋ, p, s, t, v/ and consonants, which have come from other languages, are /b, f, g/ (Karjalainen, 1999).

Vowels consist of distinct formants. These formants play a crucial role in the detection of vowels. Vowels have four to five formants, but the first two to three formants are usually sufficient in the recognition of the vowels. Also, vowels can be recognized, under certain conditions, even when the two lowest formants are missing. In normal speech there are thus multiple acoustic cues aiding the recognition of vowel sounds, making it possible to detect vowels with the presence of distortion and interference. An example of this distortion is a speech, which is played at a faster speed than it was recorded (so called ”duck talk”). This speech, although unnatural, is still understandable, although naturally all the formant frequencies have been increased. So, somehow we can ”scale” the speech automatically; how this is done is, however, not well understood (Rossing, 2002). And even if the formants are not scaled to each other the same way (this is a difference between the speech of adults and that of children), we can still recognize, which vowel is being spoken.

2.5 Magnetoencephalography

Magnetoencephalography (or MEG) measures the magnetic field generated by the brains. *Neurons* in the brain are connected to each other, forming a highly complex network (at least 10^{10} neurons in the cortex, forming 10^{14} connections (Hämäläinen et. al., 1993)). Information between these neurons is being transmitted via *electric*

current. This current produces an *electric field,* which in turn produces a *magnetic field,* which is being measured.

The relationship between electrical current and magnetic field can be derived from the quasistatic approximation of Maxwell's Equations, when measuring magnetic field outside the head. This means, that in the calculation of **E** and **B,** $\partial\vec{E}/\partial t$ and $\partial\vec{B}/\partial t$ can be ignored as source terms (Hämäläinen et. al., 1993). This is because neuromagnetism generally deals with frequencies that are below 100Hz (Hämäläinen et. al., 1993), so the signal doesn't change too rapidly for the quasistatic approximation. Also, permeability of the tissue in head is that of the free space, $\mu = \mu_0$. Maxwell equations are as follows:

$$\nabla \cdot \vec{E} = \rho / \varepsilon_{0,} \tag{1}$$

$$\nabla \times \vec{E} = -\partial\vec{B}/\partial t \ , \tag{2}$$

$$\nabla \cdot \vec{B} = 0 \ , \tag{3}$$

$$\nabla \times \vec{B} = \mu_0(\vec{J} + \varepsilon_0 \partial\vec{E}/\partial t) \, , \tag{4}$$

Where $\vec{E}$ is electric field $[V/m]$, ρ is free electric charge density $[V/m^3]$, ε_0 is permittivity of free space [8,854*10^-12 F/m], $\vec{B}$ is magnetic flux density [T or Wb/m^2], t is time [s], μ_0 is permeability of free space [$4\pi \times 10^{-7}\ H/m$] and $\vec{J}$ is current density [A/m^2].

When we take in to consideration the quasistatic approximation, we get the following formula from (2):

$$\nabla \times \vec{E} = 0 \tag{5}$$

so $\nabla \times \vec{E}$ doesn't exist (in reality, it has a value, but its contribution to **E** is so small, that it can be ignored)

We get the following formula from (4)

$$\nabla \times \vec{B} = \mu_0 \vec{J} \quad (6)$$

MEG measures mostly *postsynaptic signals* from the nerve cells. MEG also measures mostly sources, which are tangential to the surface of the brain. This means, that MEG mostly measures signals coming from *fissures.*

The magnetic field surrounding the head is picked up using SQUIDs (superconducting quantum interference device), which are sensitive detectors of magnetic flux. SQUIDs were first introduced in the late 1960s by James Zimmerman. In practice, multiple SQUIDs are used to measure signals from the brain, and the subject is brought to as close proximity to the SQUIDs as possible. This is done by e.g. putting a subject into a measurement device, where he/she is seated and then his/her head is placed to a "cup", inside which the SQUIDs are (Figure 2.19). Several SQUIDS are used, because when the current distribution inside the head is to be determined, the magnetic field has to be sampled from several locations and preferably simultaneously (Hari, 1998). The SQUIDs have to be kept in a cold environment, in practice in liquid helium (at -269°C). Superconducting flux transformers couple the magnetic field into the SQUID sensors. There are multiple kinds of transformers available, and they have different kinds of properties. Of the transformers, a simple magnetometer is the most sensitive to signals coming from the brains, but it is also most sensitive to noise.

First-order gradiometer is a more elaborate transformer, containing a compensation coil, which is wound in the direction opposite to the pickup coil. This arrangement decreases the influence of distant disturbances, so the output of the first order gradiometer is mostly determined by the nearby neuronal source. This kind of coil picks up amplitude of the radial field component B_r. The planar gradiometer measures the tangential derivative ($\partial B_r / \partial x$ or $\partial B_r / \partial y$). Planar gradiometer is better at localizing the source's place along the surface of the cortex, whereas axial gradiometer is better in localizing source's depth.

Figure 2.19. A subject in a MEG device (adapted from http://www.elekta.com/healthcare_international_functional_mapping.php)

Detection of the head position relative to the sensors is essential for the measurements. The head position can be determined by placing three or four small wire loops on specific spots on the scalp. Then magnetometers pick up the field pattern produced by currents led through the loop.

The problem posed in MEG is the fact that electrical currents in the brain have to be deducted from the measured magnetic field. This is the so called *electromagnetic inverse problem.* When a current distribution is known, then the magnetic field can be calculated from it. But when the magnetic field is known, there is no unique solution to this problem (calculating the current distribution). Therefore source models (e.g. current dipoles) or special estimation techniques have to be used to interpret the data (Hämäläinen et al., 1993). The current dipole is a widely used concept in neuromagnetism.

It is useful to divide the current density **J** in to two components (this is useful in calculating the current dipole). *Return* or *volume current*

$$\vec{J}^{v}(\vec{r}) = \sigma(\vec{r})\vec{E}(\vec{r}) \qquad (7)$$

is passive. Everything else is the primary current $\vec{J}^P$. We get the following formula for the entire current

$$\vec{J}(\vec{r}) = \vec{J}^{\rho}(\vec{r}) + \sigma(\vec{r})\vec{E}(\vec{r}) = \vec{J}^{p}(\vec{r}) - \sigma(\vec{r})\nabla V(\vec{r}) \qquad (8)$$

Here, $\sigma(\vec{r})$ is the macroscopic conductivity $[S/m]$ which assumes that cortex is a homogenous conductor. From the equation (8) it can be seen that neural activity gives rise to *primary current* mainly inside or in the vicinity of the cell, whereas *volume current* spreads everywhere in the tissue. If the primary activity is located, then the source of activation is located. The primary current can be approximated with the following formula:

$$\vec{J}^{P}(\vec{r}) = \vec{Q}\delta(\vec{r} - \vec{r}_Q) \qquad (9)$$

where $\delta(\vec{r})$ is the Dirac delta function.

Acquiring the final result using current dipole model is, though, quite a complex task, and if more realistic conductor models are used (the conductor is not assumed to be spherically symmetrical), than also numerical solutions have to be used. Description of the calculation of the source distribution can be found from Hämäläinen *et al.*

Magnetic fields measured using MEG are extremely week, several orders of magnitude weaker than earths magnetic field (Hari, 1998, p. 1108, see Table 2.2). Also other sources in the urban and laboratory environment cause disturbances in the measured magnetic field (trains, elevators etc.). Therefore MEG measurements are usually done in a magnetically shielded room. Metallic or other magnetic objects (e.g. digital watches) must be kept outside the room, because if brought inside the shielded area, they cause large disturbances in the SQUIDs, especially if they are brought to close distance from the sensors. Magnetically shielded room attenuates the external magnetic field, and is usually made of several layers of aluminum and μ-metal. Measurements can also be done without the shielding room, if special compensation techniques are available, but better results are obtained using shielding than compensation.

Table 2.2. Orders of Magnitude of different Magnetic Fields (in femtoteslas) (adapted from Hari, 1998)

Magnetic resonance imaging	1 000 000 000 000 000 (=1 T)
Steady magnetic field of the earth	100 000 000 000
Magnetocardiogram	100 000
Cereblar alpha rhythm	1000
Cereblar evoked response	100
Sensitivity of magnetometers	10
Noise within a shielded room	1

MEG can be used to measure auditory responses from the temporal cortex (location of auditory cortices). Speakers can't naturally be brought to a shielded room, since they produce a strong magnetic field. Instead, audio stimuli can be produced with electroacoustic transformers placed outside the shielded room, which are then connected to the subject through plastic tubes. When a subject is exposed to audio stimuli, his/her brain responses to it: these responses are called Auditory Evoked Fields (or AEFs). Earliest cortical responses can be seen within 19-20 ms from the stimulus onset. Middle latency responses are seen in about 30 ms time (this time varies somewhat between individuals). These are followed by responses around 50, 100 and 200 ms (P50m, N100m and P200m). In these terms, P means that EEG (EEG = electroencelophalography; measures electric field at the scalp, and deducts sources in brains based on that) measurement of the same signal from the top of the head is positive, N that it is negative, and m refers to "magnetic" (Figure 2.20).

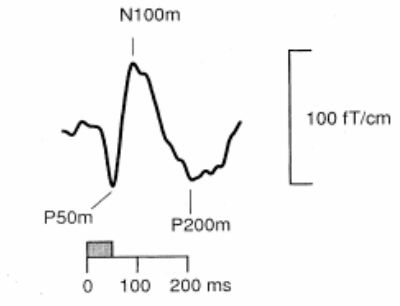

Figure 2.20. A typical magnetic response, measured from the proximity of subjects auditory cortex. In this case, the stimulus used was a 50 ms tone, and 150 single responses were averaged. The P50m, N100m and P200m responses are clearly detectable from the picture. (Adapted from Hämäläinen et. al., 1993)

MEG is not limited to measuring auditory responses from the brains. When responses from the somatosensory cortex are being monitored, these responses are called somatosensory evoked fields or SEFs (analogically to Auditory Evoked Fields or AEFs). When responses from visual cortex are being monitored, these responses are called visual evoked fields or VEFs.

2.6 Purpose of the study and specific hypotheses

The purpose of the study was to show, that the visual system (i.e. what is seen) sensitizes the auditory system so, that when the visual stimulus is visual speech and the auditory stimuli are sine sweeps, then the activation differs from the situation, than when the visual stimulus is a still face, and the auditory stimulus are sine sweeps. This sensitizing should either increase or decrease the activation in the auditory cortex. It was assumed, that this effect would be observable at the auditory cortex, but it origins might be either at the auditory cortex or at the corticofugal connections.

3. Methods

3.1 Subjects

11 healthy, voluntary, right handed subjects participated in the experiment (2 females, 9 males). Subjects were chosen to be right handed partly because right handed people have a smaller portion of people (in percentages), whose language functions are not left lateralized in the brains, compared to left handed population. The age of subjects ranged between 22-32 years (mean±stdev=25.9±3.0). All had normal hearing (self reported), no neural diseases (self reported) and either normal vision (self reported) or corrected-to-normal vision. Three of the subjects had to be discarded from the final results due to various reasons, which led to poor data quality, so the final number of the subjects in the analysis was 8 (6 males, 2 females, age 22-32 years, mean±stdev=26.4±3.3).

.

3.2 Used stimuli

The experiment used audio and visual stimuli. The audio stimuli were 6 different kinds of sine sweep sounds, all lasting 50 ms. Beginning and ending frequencies of the sweeps were the following: 200-700 Hz (F1), 400-1800 Hz (F2a), 1000-1800 Hz (F2b), 1600-1800 Hz (F2c), 2200-1800 Hz (F2d) and 2800-1800 Hz (F2e). The range of F1 is approximately 1.81 octaves and the range of F2a is approximately 2.17 octaves. The approximate range of the remaining sine sweeps are as follows (in octaves): F2b 0.848; F2c 0.170; F2d -0.290 and F2e -0.637. First sound matched a simplified version of the syllables /ba/ and /ga/ first formant transitions, and other sounds had been interpolated between the simplified version of second formant transition matching /ba/ (400-1800 Hz), and the simplified version of second formant transition matching /ga/ (2800-1800 Hz). Three different video stimuli were used: 1) a person was pronouncing /ba/, 2) the same person was pronouncing /ga/ or 3) a still picture from the person was shown. Each video stimulus lasted for 1312 ms.

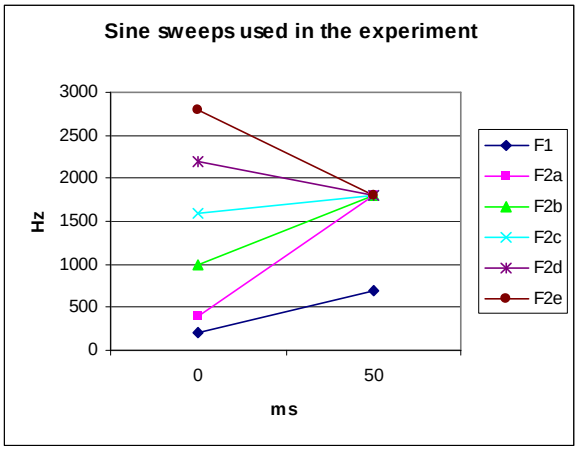

Figure 3.1. The plots show beginning and ending frequencies for the different stimuli used.

3.3 Proceeding of the experiment

Sounds were presented to subjects via earplugs (manufacturer Etymotic), and a video was presented to a "wall" via video projector. The stimulation was audiovisual in nature; video clips and sounds were simultaneously presented to subjects. Sounds were presented in a random order, but similar sounds never came in succession. Inter stimulus interval (ISI), from onset to onset, for sounds was varied randomly between 990-997 ms. ISI between different video clips was 100-200 ms. Videos were presented in blocks (one condition was repeated) that lasted 20-40 s (the duration being randomized). Order of the blocks was randomized in the following way: each three conditions (that is, /ba/, /ga/ and still-condition) come first in a random order, then they come again in random order, but so, that the first one of this three-block block isn't the same condition as the last one in the previous three-block block, etc., so the conditions don't come in succession. Figure 3.2 clarifies the order of video blocks.

The experiment lasted for about 45 minutes; the aim was to obtain approximately 100 fast artifact-free epochs for each category. The test was done in about five minute's periods; after each period, a short pause was kept. This was done so that subjects wouldn't get fatigued. Subjects were instructed to lift their finger each time the video block changed; these finger lifts were then recorded to a log.

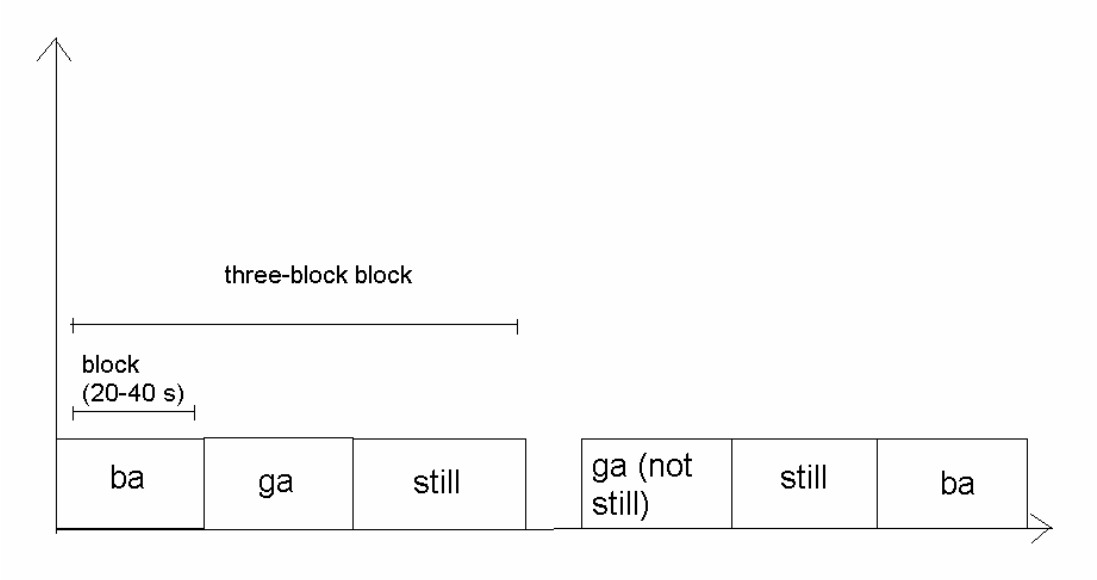

Figure 3.2. Figure shows an example, how two three-block blocks might be organized. One three-block block shows all different situations in random order, but so, that the first video stimuli in a three-block block isn't the same one as the last video stimuli in the previous three-block block (so that the same condition doesn't come up consecutively).

3.4 Equipment

The neural responses were recorded using a MEG device. The MEG device was located at Helsinki University of Technology, Low Temperature Lab. The device was a 306 channel *Neuromag Vectorview* SQUID neuromagnetometer, with 102 sensor elements in a helmet array (two orthogonal planar gradiometers and one magnetometer in each element). The equipment is located in a magnetically shielded room, which attenuates magnetic fields from outside the room. The recording and analysis software is also provided by Neuromag Ltd. All channels were used in a recording, although results from the temporal lobe areas were searched. EOG electrodes were used to detect blinking artifacts, which were then omitted from the final results automatically. During the recording, MEG data was filtered with a 0.1-172 Hz passband. Segments with over 3000 fT/cm (MEG) or ± 75 µV amplitude were automatically rejected. The number of segments averaged varied individually quite a lot due to various reasons.

3.5. MEG analysis

First the raw MEG data was handled using a Matlab script which calculated the averages over each situation (3 video stimuli*6 sounds = 18 situations) for each individual separately. These results were then first handled in two separate ways. Dipole fitting procedure was used, and also simple vector sum calculations were used as to confirm the dipole fitting results. In the end, it was decided, that only results from vector sum calculations would be used in this study. Thus, I have omitted the description about what was done with dipole fitting results (it is sufficient to say, that results were similar, although not equal, with vector sum calculations).

A Matlab script was originally used to calculate the vector sums for the N100m component from connected gradiometer channels. Calculation of vector sums is commonly used method in interpreting MEG measurement results (for a study, in which this method is used, see e.g. Ahveninen et al., 2000). This script did a "passband" filtering (actually first a low-pass and then a high-pass filtering) and a baseline correction from the filtered signal, before the vector sums (vector sum = sqrt((amplitude of the first channel)^2 + (amplitude of the second channel)^2) were calculated for connected channels. Then a Matlab script was used to pick maximum channels from the left and right side, and this script also picked up the maximum activation and its latency (the N100 peak was looked for, so the script did a search on the vicinity of that time). Maximum vector sums obtained this way differed significantly from those, which were obtained using Neuromag. Because of the prementioned problem, the vector sums were calculated using filters in the Neuromag, but so, that the maximum channel which was obtained with the Matlab scripts, was looked. However, in some cases, where the channel pair obtained with Matlab script was obviously wrong, results obtained with automatic script was ignored, and different channels and/or activation times were chosen.

It was planned that vector sum calculations would be done also to P200m peaks and again the latencies and amplitudes of the observations would be measured. One could assume that the P200m peak would actually be more prominent to show modulation

results, since at least, when McGurk effect was studied (Sams et. al., 1991), an integration effect was observable 200-300 ms from the stimulus. The situation between the study by Sams et al. and the present study is similar enough to assume that there might be some correlation between the results (if the assumed modulation effect in the present study exists in the first place). However, the P200m responses were very weak with some subjects, so there wasn't enough good data to do analysis with this response.

Preliminary handling of the vector sums was done with Excel. Averages across all the subjects for different situations were calculated, both for the magnitude and the latency of the activation. Also standard errors of means were calculated for the averages. Standard errors of means were relatively large, and thus the potentially observed effects are not reliable, since differences between different situations fit with the range of SEM (standard error of mean).

Data was analysed using repeated measures ANOVA (Arnold and Milton, 1995); Statistica software was used to perform the analysis. Two different kinds of ANOVA analysis were done: in the first one, the video clip was the first independent variable and the formant was the second independent variable. Left and right hemispheres were analyzed separately. The latencies of activation peaks were analyzed, as well as activation peak amplitudes. This means, that in the first analysis, altogether (2 hemispheres) * (2 different values (times & activation values)) = 4 different situations were analyzed. Activation peaks and their latencies were dependent variables. In the second ANOVA analysis, there were three independent variables: video clip, formant and hemisphere. In this case, 2 different situations were analyzed (2 different values, amplitudes and latencies), since hemispheres were independent variables. In the first ANOVA analysis, the effects within formants, within video clips, and interaction effect between formants and video clips were looked for. In the second ANOVA analysis, the effects within formant, within video clips, within hemispheres, and interaction effects between video clips and formants, between video clips and hemispheres, between formants and hemispheres and, finally, between all three independent variables were looked for.

For potentially meaningful effects, contrast analysis was performed using Scheffé test. This test is a rather conservative one, so it doesn't show significant differences between different means as often as other, less conservative tests like Duncan test (Winer, 1962).

Contrast analysis was done for all the conditions, where hemisphere was not a variable (4 conditions = 2 hemispheres * 2 measured variables (latency or amplitude)) so, that analysis was done separately for each formant sweep, and still situation was compared against the combination of ba and ga situation. Also, contrast analysis was done so, that always an individual value from a visual condition was compared against the matching value from another visual condition (e.g. left amplitudes, /ba/-F1 is compared against /ga/-F1). In this case, the results are actually identical to that, if paired two tail t-test to compare the averages would be used.

The behavioral results (subjects responses, i.e. finger liftings, to visual stimuli) were studied from the log-files. Hit rate was calculated; hit rate = hits/(hits + misses + false alarms + too slow responses (> 3s)). From these hit percents it was decided, which subjects were "good" (meaning, that they had a high enough hit percent to assume, that they had paid attention to the task and had understood the task correctly). Also, from the hits that weren't too slow (t < 3s), average response time, standard deviation and standard error of means (SEM) were calculated for each "good" subject individually and these values were also calculated over all "good" subjects. There was no exact limit decided for what was a "good" subject, but separating "good" subjects from "bad" subjects was easy in the end, since differences in the hit percent were so large between these two groups. It was decided *post hoc*, that the statistical calculations that were done for all eight subjects, would be done separately for all the good subjects (here, N = 4, more about this in the chapter "4. Results").

4. Results

Figures 4.1-4.4 show results for maximum channel activations (N = 8).

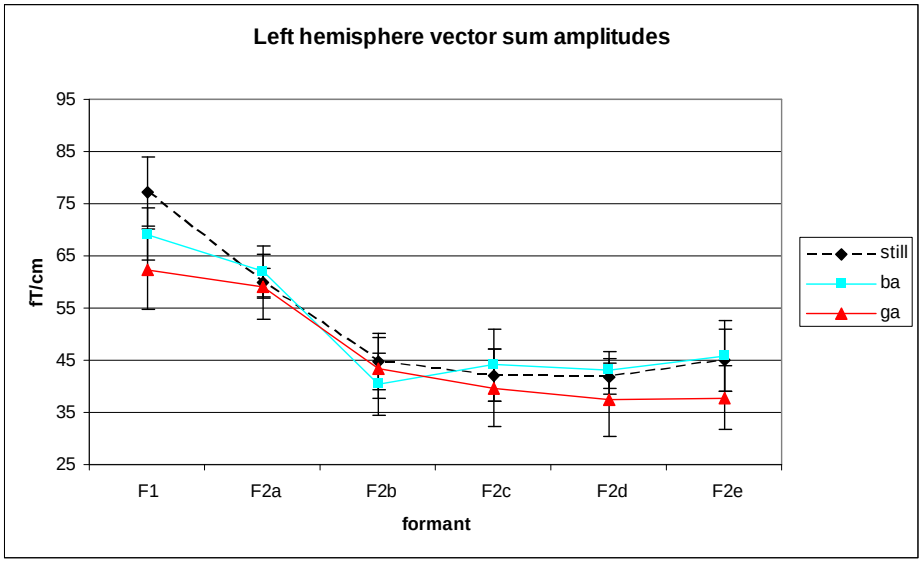

Figure 4.1. The maximum vector sum amplitudes for the left hemisphere.

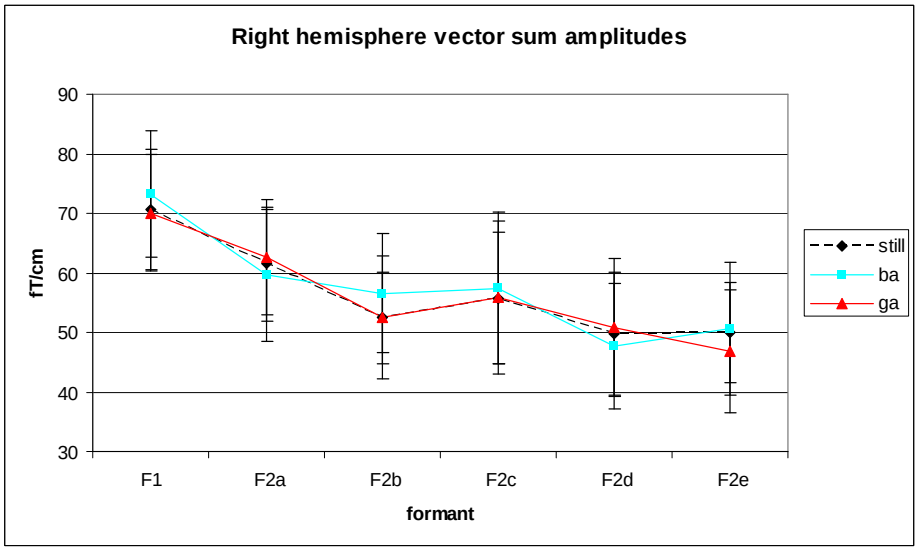

Figure 4.2. The maximum vector sum amplitudes for the right hemisphere.

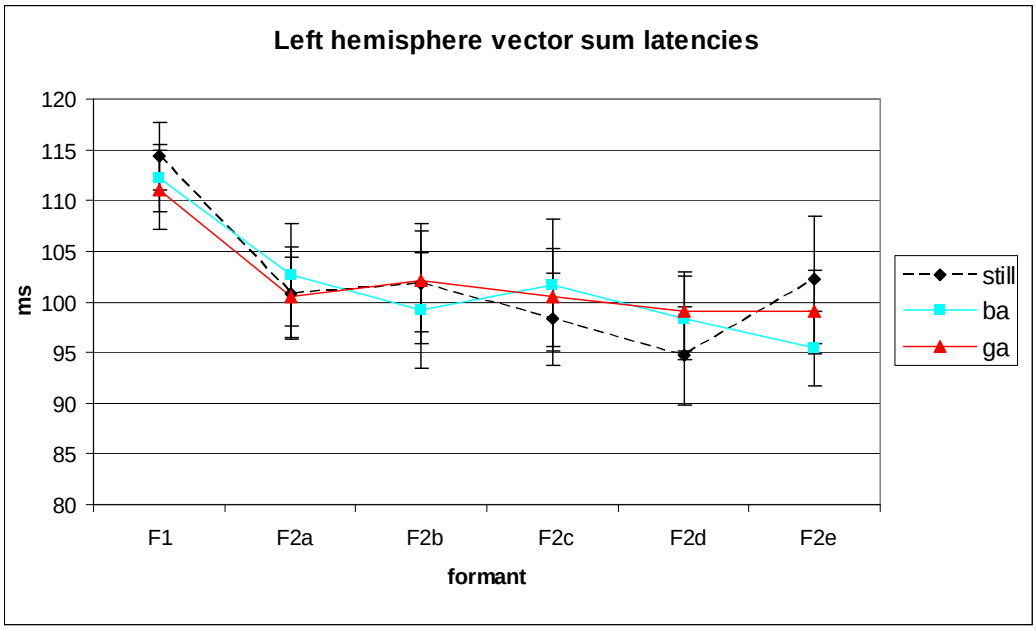

Figure 4.3. The latencies of maximum responses from the left hemisphere. The Y-axis shows time from the onset of the sound.

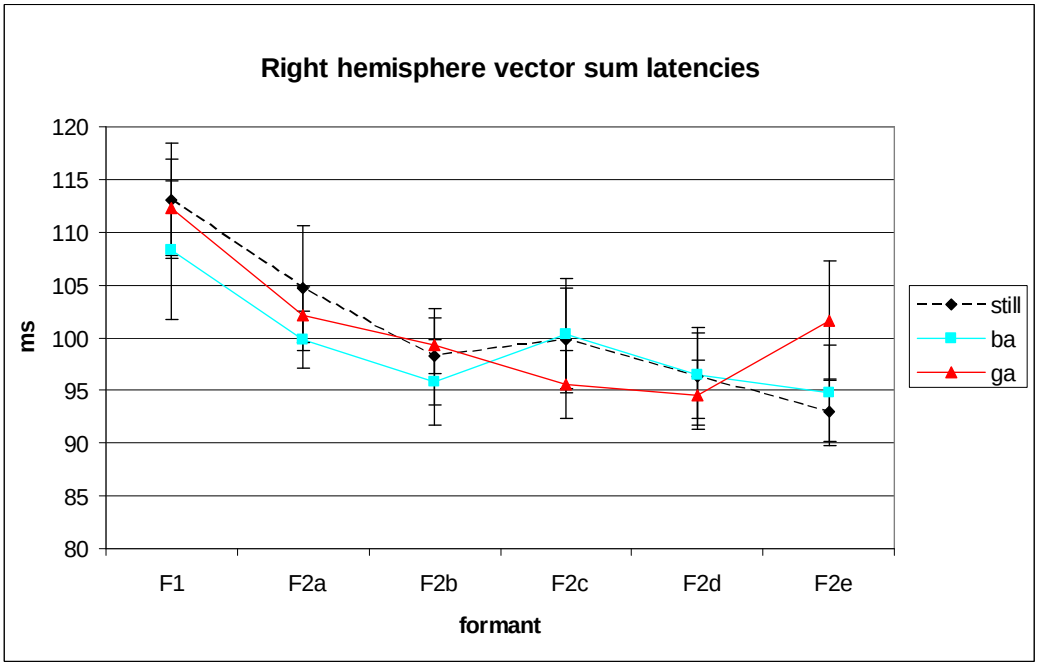

Figure 4.4. The latencies of maximum responses from the right hemisphere.

ANOVA showed that the only statistically significant effects were within formants effects with vector sums. Table 4.1 shows F- and p-values for different conditions, for the hypotheses that different formant sweeps cause different kind of activation in the brains, when left and right hemispheres are examined separately, and Table 4.2. shows similar values, when hemisphere is one of the independent variables. As can be seen from the charts, in every situation, a formant effect was present with p-value $p < 0.01$, and excluding the result from right hemisphere latencies, $p < 0.005$ in every situation.

Table 4.1. F- and p-values for a formant effect in different conditions, when hemispheres are examined separately.

	F(5,35)	P
Left amplitude	21.7	7.7E-10
Right amplitude	4.28	0.00385
Left latencies	8.66	2.04E-05
Right latencies	4.06	0.00518

Table 4.2. F- and p-values for a formant effect in different conditions, when hemisphere is an independent variable.

	F(5,35)	P
Maximum amplitudes	14.9	7.86E-08
Maximum latencies	8.17	3.43E-05

Results from Scheffé test (all six formants were compared with each other) are summarized in tables 4.3 and 4.4 (two tables are used because of lack of space), which shows p-values for the assumption, that F1 differs from other formants and that F2a differs from other formants. Only these combinations are presented in the chart, since at no time did the activations of formants F2b, F2c, F2d and F2e differ from each other significantly. $p < 0.05$ has been used as a significant value, because this is a traditional limit for a meaningful result. This is a rather arbitrary limit, leaving out e.g. a result 0.0546 (left maximum channel amplitude, still situation, F1 vs. F2a), and including e.g. a result 0.0477 (left maximum channel amplitude, still situation, F2a vs. F2c). The results show, that when the amplitude is measured, at the left side the activation matching F1 always significantly differs from other activations. F2a, in addition to differing from F1, also differs most of the time from other formants (9 times out of 12). When looking at the amplitude of activation at the right side of the brain, F1 significantly differs from other activations occasionally.

When looking at the latency of maximum activations at the left side of the brain, then only F1 differs from other formants. When looking at the latency of maximum activations at the right side of the brain, then again only F1 differs from other formants. These results fit with the general trend that can be intuitively detected looking at the figures drawn from the averages: when looking at the amplitude of activation, the amplitude drops from F1 to F2a, and again drops from F2a to F2b, after which it remains relatively constant. This effect is apparently much more prominent on the left than on the right side (where it is practically non existent), which would indicate, that there is some sort of lateralization present. However, ANOVA didn't show a lateralization effect, so that can be just coincidence. When looking at the latency of activation, the latency drops from F1 to F2a, after which it remains relatively constant.

Table 4.3. Table shows the p-values for formant pairs including F1. Significant p-values ($p < 0.05$) are highlighted with red color

		F1 vs. F2a	F1 vs. F2b	F1 vs. F2c	F1 vs. F2d	F1 vs. F2e
Left amplitudes	Still	0.0546	2.56E-05	5.89E-06	4.93E-06	2.98E-05
	Ba	0.788	5.89E-05	0.000517	0.000284	0.00131
	Ga	0.985	0.00553	0.000521	0.000126	0.000159
Right amplitudes	Still	0.800	0.131	0.315	0.0543	0.0596
	Ba	0.619	0.395	0.452	0.0463	0.105
	Ga	0.874	0.104	0.289	0.0589	0.0120
Left latencies	Still	0.0204	0.0392	0.00344	0.000209	0.0479
	Ba	0.425	0.120	0.312	0.0863	0.0191
	Ga	0.00197	0.0124	0.00192	0.000368	0.000345
Right latencies	Still	0.737	0.149	0.249	0.0742	0.0175
	Ba	0.835	0.492	0.870	0.548	0.402
	Ga	0.272	0.0732	0.00885	0.00478	0.220

Table 4.4. Table shows the p-values for pairs including F2a (excluding pair F1F2a, which was in table 4.3). Significant p-values ($p < 0.05$) are highlighted with red color. Pairs not including F1 or F21 were omitted from the tables, since these pairs didn't show significant results (that is, $p > 0.05$ always).

		F2a vs. F2b	F2a vs. F2c	F2a vs. F2d	F2a vs. F2e
Left amplitudes	Still	0.135	0.0477	0.0416	0.149
	Ba	0.00367	0.0250	0.0150	0.0533
	Ga	0.0357	0.00415	0.00107	0.00135
Right amplitudes	Still	0.808	0.967	0.578	0.603
	Ba	0.999	1.00	0.729	0.898
	Ga	0.654	0.914	0.498	0.187
Left latencies	Still	1.00	0.990	0.667	1.00
	Ba	0.983	1.00	0.960	0.710
	Ga	0.990	1	0.995	0.994
Right latencies	Still	0.887	0.963	0.730	0.380
	Ba	0.993	1.00	0.997	0.980
	Ga	0.989	0.714	0.582	1.00

The first contrast analysis showed that visual stimulus affected the amplitude in some cases, but not usually. Below are charts of the p-values obtained from contrast analysis. In each situation, one of the three visual stimuli was always compared against the two other stimuli. Only one contrast analysis (visual /ga/ was compared against visual /ba/ and still) showed significant results (Table 4.5), so results from other analysis have been omitted here. In that case, when looking at the activation amplitudes from the left hemisphere, with both sweeps F1 and F2e, the activation in brains associated with /ga/ is significantly smaller than that associated with /ba/ and still.

Table 4.5. Table shows different p-values for contrast analysis, when visual /ga/-condition is compared against visual /ba/- and still-conditions. Significant p-values ($p < 0.05$) are highlighted with red color.

	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.0240	0.671	0.703	0.276	0.275	0.000358
Right hemisphere amplitudes	0.564	0.570	0.494	0.773	0.425	0.144
Left hemisphere latencies	0.331	0.343	0.546	0.723	0.203	0.871
Right hemisphere latencies	0.500	0.958	0.551	0.0779	0.495	0.0701

The second contrast analysis also showed that visual stimulus affected the amplitude occasionally. Only results from those tests, which showed significant results, are showed here; in this case, the results from analysis, where still-condition was compared against /ba/-condition, were omitted. Table 4.6 shows results, when still-condition is compared against /ga/-condition. Here, the visual effect is present, when

looking at amplitudes from the left cortex, with sine sweeps F1 and F2e. The activation associated with still is significantly stronger than what is associated with /ga/. Table 4.7 shows results, when /ba/-condition is compared against /ga/-condition. Here, the visual effect is present once, when looking at amplitudes from the left cortex with sine sweep F2e. The amplitude associated with /ba/ is significantly stronger than that associated with /ga/.

Table 4.6. Table shows different p-values for contrast analysis, when visual still-condition is compared against visual /ga/-condition. Significant values ($p < 0.05$) are highlighted with red color

	F1	F2a	F2b	F2c	F2d	F2e
Left amplitudes	0.0364	0.846	0.620	0.644	0.345	0.0118
Right amplitudes	0.749	0.751	0.993	0.969	0.739	0.425
Left latencies	0.176	0.659	0.898	0.0876	0.133	0.351
Right latencies	0.627	0.533	0.811	0.0866	0.581	0.0934

Table 4.7. Table shows different p-values for contrast analysis, when visual /ba/-condition is compared against visual /ga/-condition. Significant values ($p < 0.05$) are highlighted with red color.

	F1	F2a	F2b	F2c	F2d	F2e
Left amplitudes	0.104	0.582	0.390	0.0828	0.247	0.0059
Right amplitudes	0.507	0.512	0.246	0.548	0.363	0.0826
Left latencies	0.608	0.350	0.461	0.660	0.711	0.199
Right latencies	0.363	0.149	0.379	0.117	0.708	0.0645

As was mentioned in the Chapter 3, statistical methods that were applied for the eight subjects, were separately applied for the four good subjects. Separanting the ”good” subjects from the ”bad” subjects was rather easy: with good subjects, the hit percents were as follows: 85.3, 85.1, 90.7, and 97.4%. With ”bad” subjects, the hit percents were as follows: 0 (meaning that there were no responses), 59.7, 58.1, and 52.4%. The Table 4.8 shows hit percentages, average response times, standard deviations of response times, and standard errors of means of response times for all the ”good” subjects individually and together.

Table 4.8. Table shows following values for all the ”good” subjects (N = 4) invidually and together: hit percentage, and average, standard deviation, and standard error of means of response latencies.

	Subject 1	Subject 2	Subject 3	Subject 4	All subjects
Hit percentage	85.3	85.1	90.7	97.4	89.8
Averages (s)	1.32	1.18	1.29	0.892	1.16
Standard deviations (s)	0.438	0.485	0.353	0.207	0.414
Standard error of means (s)	0.055	0.065	0.043	0.024	0.026

Figures 4.5-4.8 show preliminary results for maximum channel activations for all "good" subjects (N = 4).

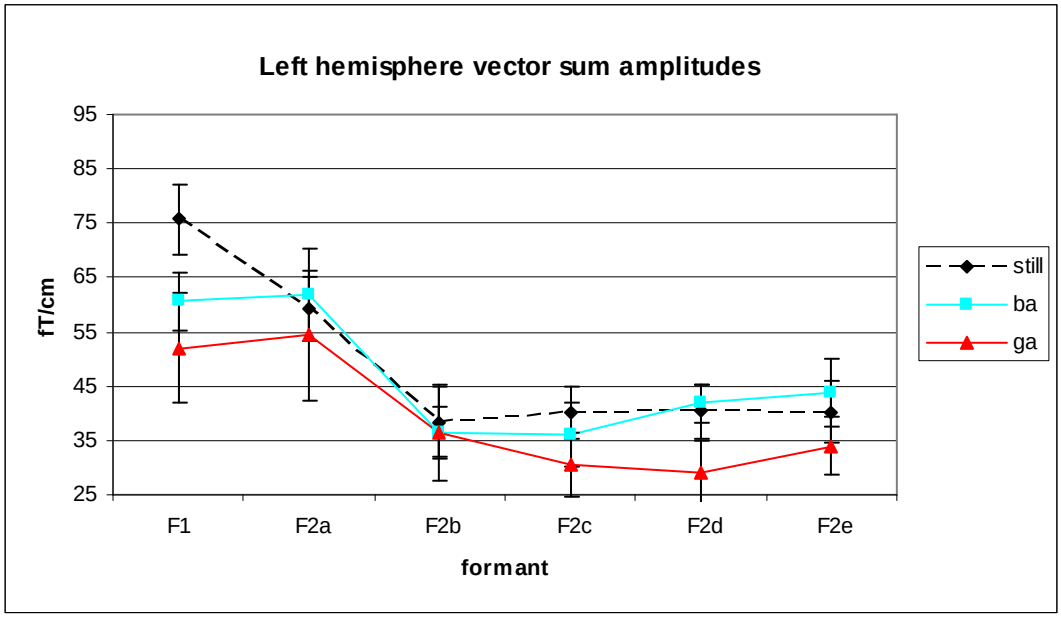

Figure 4.5. The maximum amplitudes for the left hemisphere for "good" subjects.

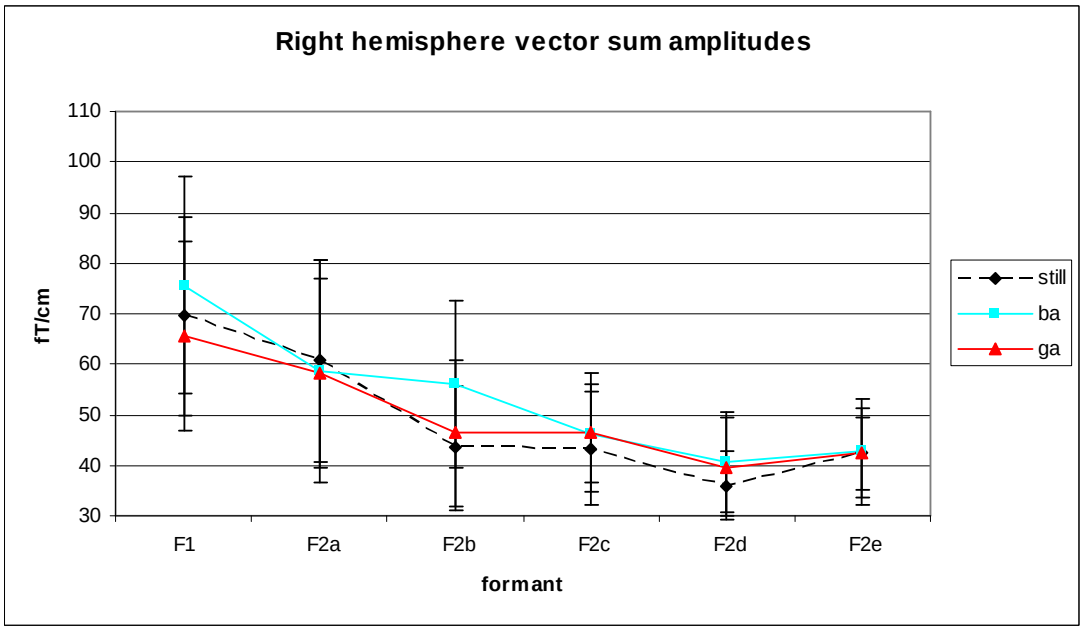

Figure 4.6. The maximum amplitudes for the right hemisphere for "good" subjects.

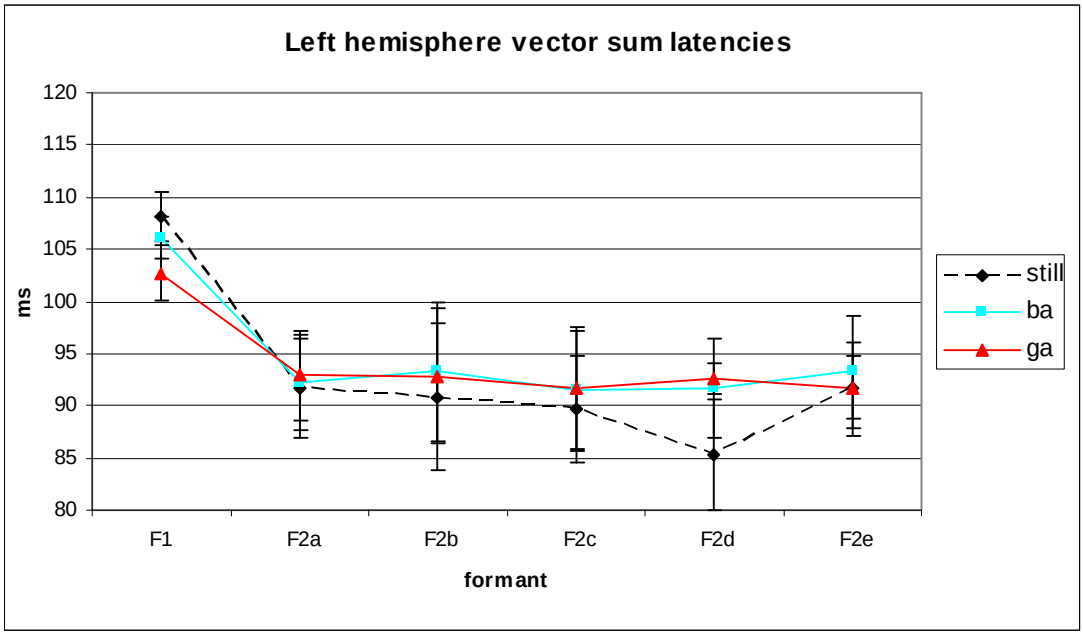

Figure 4.7. The latencies from the left hemisphere for "good" subjects. The Y-axis shows time from the onset of the sound.

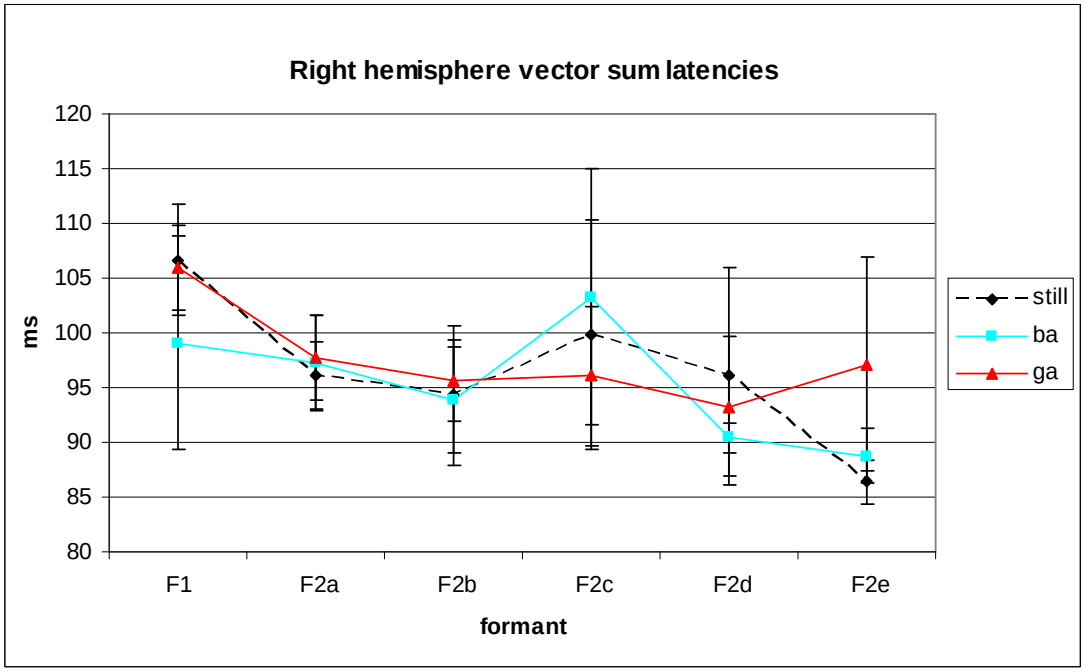

Figure 4.8. The latencies from the right hemisphere for "good" subjects

For “good” subjects, ANOVA showed significant effects for formants, for different visual stimuli, and a significant interaction effect between auditory and visual stimuli. For the condition, where hemispheres are examined separately, Table 4.9 shows the formant effect, Table 4.10 shows the visual effect, and Table 4.11 shows the interaction effect.

Table 4.9. F- and p-values for formant effect in different conditions, when hemispheres are examined separately.

	F(5, 15)	P
Left amplitudes	14.9	2.29E-05
Right amplitudes	3.88	0.0188
Left latencies	7.46	0.00107
Right latencies	1.51	0.244

Table 4.10. F- and p-values for visual effect in different conditions, when hemispheres are examined separately.

	F(2, 6)	P
Left amplitudes	6.09	0.0359
Right amplitudes	1.256	0.350
Left latencies	1.27	0.346
Right latencies	0.527	0.615

Table 4.11. F- and p-values for interaction effect (visual*auditory) in different conditions, when hemispheres are examined separately.

	(10, 30)	p
Left amplitudes	2.183	0.0481
Right amplitudes	1.21	0.326
Left latencies	1.64	0.144
Right latencies	0.977	0.483

For the condition, where hemisphere is one of the independent variables, Table 4.12 shows the formant effect. There were no other effects present, when hemisphere was an independent variable, so no other results are presented here.

Table 4.12. F- and p-values for formant effect in different conditions, when hemispheres are independent variable.

	F(5, 15)	P
Amplitudes	11.9	8.61E-05
Latencies	4.28	0.0129

Tables 4.13 and 4.14 show the results from Scheffé test for all ”good” subjects, testing the assumption, that there are significant differences in activations between different sine sweeps. The results are similar to those obtained with all eight subjects.

Table 4.13. Table shows the p-values for formant pairs including F1 for the group of ”good” subjects. Significant p-values ($p < 0.05$) are highlighted with red color.

		F1 vs. F2a	F1 vs. F2b	F1 vs. F2c	F1 vs. F2d	F1 vs. F2e
Left vector sum amplitude	Still	0.150	0.000233	0.000393	0.000421	0.000402
	Ba	1.00	0.0338	0.0288	0.1366	0.215
	Ga	0.999	0.272	0.0614	0.0419	0.155
Right vector sum amplitude	Still	0.946	0.135	0.131	0.0325	0.112
	Ba	0.765	0.652	0.243	0.113	0.151
	Ga	0.975	0.414	0.432	0.145	0.233
Left vector sum latency	Still	0.0393	0.0278	0.0170	0.00300	0.0405
	Ba	0.0579	0.0887	0.0421	0.0436	0.0887
	Ga	0.0708	0.0678	0.0342	0.0594	0.0326
Right vector sum latency	Still	0.803	0.682	0.961	0.800	0.198
	Ba	1.00	0.995	0.998	0.953	0.906
	Ga	0.694	0.470	0.509	0.260	0.630

Table 4.14. Table shows the p-values for formant pairs including F2a (excluding the pair F1-F2a, which is presented in the previous table) for the group of ”good” subjects. Significant p-values ($p < 0.05$) are highlighted with red color

		F2a vs. F2b	F2a vs. F2c	F2a vs. F2d	F2a vs. F2e
Left vector sum amplitude	Still	0.0387	0.0666	0.0713	0.0681
	Ba	0.0249	0.0212	0.103	0.166
	Ga	0.160	0.0326	0.0220	0.0865
Right vector sum amplitude	Still	0.510	0.501	0.171	0.450
	Ba	1.00	0.923	0.718	0.806
	Ga	0.832	0.847	0.446	0.613
Left vector sum latency	Still	1.00	0.998	0.794	1
	Ba	1.00	1.00	1.00	1.00
	Ga	1.00	0.999	1.00	0.999
Right vector sum latency	Still	1.00	0.998	1.00	0.849
	Ba	0.999	0.990	0.983	0.957
	Ga	0.999	1.00	0.964	1.00

The first contrast analysis done with ”good” subjects showed, that visual stimulus affected the amplitude of responses occasionally (and latency once). There are some differences compared to the results obtained from all eight subjects. This time,

amplitudes in still-situation differed from those in /ba/- and /ga/-situations. Also, latency of brain responses was once affected by visual stimulus (still-situation vs. /ba/- and /ga/-situation, sweep F2b). When /ga/-situation was compared against /ba/- and /still/-situation, amplitudes from the left hemipshere were only affected with the sweep F2e and not with sweep F1 (which was the case with all eight subjects).

Tables 4.15 and 4.16 show the results from the first contrast analysis done with ”good” subjects (results obtained from the analysis, when /ba/-condition was compared against still- and /ga/-condition are omitted here, because there were no significant results). It can be seen from the Table 4.15, that when visual still-condition is compared against visual /ba/- and /ga/-conditions, then visual effect is occasionally present in the left hemisphere, but not in the right one. In the case of left hemisphere amplitudes, the amplitude caused by visual still is stronger than amplitude caused by /ba/ and /ga/ observed together with sine sweeps F2c and F2d (actually the amplitude caused by /ba/ is slightly stronger than that caused by still with F2d). In the case of left hemiphere latencies, the latency of still is shorter than that of /ba/ and /ga/ with sine sweep F2b. It can be seen from the Table 4.16, that when visual /ga/-condition is compared against visual still- and /ba/-conditions, visual effect is present once, when looking at amplitudes from the left hemisphere, with sine sweep F2e. In this case, the amplitude caused by /ga/ is significantly weaker than those caused by still- and /ba/.

Table 4.15. Table shows different p-values for contrast analysis done with all ”good” subjects, when visual still-condition is compared against visual /ba/- and /ga/-conditions. Significant p-values ($p < 0.05$) are highlighted with red color.

still vs. others	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.0830	0.726	0.724	0.00893	0.0179	0.804
Right hemisphere amplitudes	0.410	0.324	0.169	0.596	0.520	0.961
Left hemisphere latencies	0.171	0.187	0.0291	0.199	0.228	0.756
Right hemisphere latencies	0.375	0.528	0.506	0.945	0.613	0.282

Table 4.16. Table shows different p-values for contrast analysis done with all ”good” subjects, when visual /ga/-condition is compared against visual still- and /ba/-conditions. Significant p-values ($p < 0.05$) are highlighted with red color

ga vs. others	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.0857	0.399	0.698	0.106	0.0502	0.00196
Right hemisphere amplitudes	0.149	0.799	0.396	0.549	0.799	0.869
Left hemisphere latencies	0.340	0.360	0.218	0.494	0.227	0.354
Right hemisphere latencies	0.538	0.333	0.557	0.310	0.993	0.281

The second contrast analysis done with all the “good” subjects gave, in short, the following kind of results: visual effect was present only on the left cortex, occasionally, and with both amplitudes and latencies. Exact results can be seen from tables 4.17., 4.18., and 4.19. Table 4.17 shows results from constrast analysis, where visual still-condition was compared against /ba/-condition. Here, visual effect is present once; when looking at latencies from the left hemisphere, with sine sweep F2b. The latency of still is conciderably shorter than that of /ba/. Table 4.18 shows results from contrast analysis, where visual still-condition is compared against /ga/-condition. Here, the visual effect is present with left hemisphere amplitudes with sine sweeps F2c and F2d, and with left hemisphere latencies with sine sweep F2b. The amplitude caused by still is conciderably stronger than that caused by /ga/. In the case of latencies, the latency caused by /ga/ is conciderably shorter than that caused by still. Table 4.19 shows results from constrast analysis, where visual /ba/-condition was compared against visual visual /ga/-condition. Here, the visual effect is present once, with left hemipshere amplitudes with sine sweeps F2e. The amplitude caused by /ba/ is conciderably stronger than that caused by /ga/.

Table 4.17. Table shows different p-values for contrast analysis done with all ”good” subjects, when visual still-condition is compared against /ba/-condition. Significant p-values ($p < 0.05$) are highlighted with red color.

still vs. ba	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.0978	0.513	0.780	0.175	0.586	0.621
Right hemisphere amplitudes	0.213	0.367	0.107	0.651	0.558	0.952
Left hemisphere latencies	0.120	0.556	0.0416	0.226	0.301	0.580
Right hemisphere latencies	0.417	0.699	0.773	0.102	0.593	0.319

Table 4.18. Table shows different p-values for contrast analysis done with all ”good” subjects, when visual still-condition is compared against /ga/-condition. Significant p-values ($p < 0.05$) are highlighted with red color.

still vs. ga	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.0808	0.467	0.674	0.0305	0.0287	0.187
Right hemisphere amplitudes	0.107	0.621	0.533	0.555	0.574	0.982
Left hemisphere latencies	0.261	0.176	0.0408	0.282	0.195	0.920
Right hemisphere latencies	0.613	0.345	0.680	0.437	0.688	0.281

Table 4.19 Table shows different p-values for contrast analysis done with all ”good” subjects, when visual /ba/-condition is compared against /ga/-condition. Significant p-values ($p < 0.05$) are highlighted with red color

ba vs. ga	F1	F2a	F2b	F2c	F2d	F2e
Left hemisphere amplitudes	0.138	0.369	0.974	0.290	0.0777	0.0435
Right hemisphere amplitudes	0.168	0.955	0.128	0.908	0.881	0.930
Left hemisphere latencies	0.453	0.628	0.536	0.940	0.761	0.298
Right hemisphere latencies	0.472	0.785	0.403	0.237	0.654	0.283

5. Discussion

5.1. Summary of the results and some general thoughts

The control condition (still face) was such, that possible effects caused by pronunciation *per se* cannot be differentiated from general effects caused by movement of the stimulus mouth, facial features and head (although the head was relatively still in the video clips), although such effect as early as 100 ms from the sound onset, at the temporal cortex, is doubtful. But then again, this experiment had two different kinds of conditions, which were being studied (face pronouncing /ga/ and face pronouncing /ba/) and any differences on the activations *between* these two conditions most likely had to do with pronunciation (since both conditions naturally included mouth, facial and head movements).

The results would indicate that there may be a weak interaction effect between formant transition like sine sweeps and visual stimuli. When all the eight subjects were analyzed, this effect wasn't present, but when all four "good" subjects were tested, there was an interaction effect present with the amplitudes of brain activations from the left hemisphere. However, looking at the plot from left hemisphere activations (Figure 4.5), the exact nature of the possible interaction effect remains unclear. No lateralization effect was found.

The lack of a strong and easily identifiable interaction effect is somewhat surprising, when comparing this result to related experiments. In an auditory-visual modulation situation, when the auditory signals are sine tones (in this case, sine tones with frequencies of 125, 250, 500, 1000, 2000, 4000 and 8000 Hz, duration 50 ms in all cases) and visual stimuli are videos of a person pronouncing different vowels (in this case, Finnish vowels /y/, /a/, /o/ and /i/), the visual stimuli modulated the activation on the auditory cortex (Kauramäki, 2006). More precisely, in this MEG study there were three visual conditions; 1) video stimuli showing articulations, 2) control task with an oval with changing shape, in front of a mouth in a still face, was shown, and 3) a still face was shown. The results were as follows: in the left hemisphere, visual speech modulated the activations on the auditory cortex so that they were smaller

than with still face or control task. In the right hemisphere, the changing oval modulated activations so that they were larger than with still face or visual speech. There were no significant differences in the latencies of activations between different conditions. When the auditory stimuli are actual speech, and visual stimuli are visual speech, (Wassenhove et. al., 2005), then the effects are such that when observing auditory-visual stimulus, compared to auditory stimulus alone, then the ERP amplitudes are smaller and their latencies shorter. So, based on these two studies, it would be assumed that when visual speech is combined with sine sweeps, there would be a coherent modulation effect present. Possible explanation to the lack of a clear modulation effect, in addition to the noisy data, could be that the task in present experiment (lift a finger, when one visual condition changes to another kind) was so simple, that the subjects didn't have to perform lip reading, and thus modulation effects would be smaller than when subjects consciously do lip reading.

The reasons behind the detectable formant effects are not clear. This may be partly due to fact that the F1 and F2a pass through a larger range of frequencies. However, as can be easily seen, F2a has a larger range than F1, and yet the activation caused by F1 is larger than activation caused by F2a, so the differences in activations can not completely be explained with the range of the sweeps. Hearing thresholds are not a reason in this case, since in the range 200-2800 Hz (frequency range of the stimuli) audibility curves and equal loudness curves are rather flat, and actually at the low end, they are higher (Figure 2.1), meaning that if they would affect the results, activations matching F1 should actually be smaller than activations matching other formants. This effect, however, is rather small, and shouldn't affect the results considerably, anyway. Another possible explanation for the formant effect is that in the brain areas, which react most strongly to sine sweeps, there are more neurons dedicated to sweeps approximately matching F1 and F2a than to the other sweeps used in this experiment, but this is purely speculation.

When studying results from the formant comparisons more accurately, one notices, that in the case of all eight subjects, F2a response differs significantly from F2e, when visual stimulus is /ga/ (in the case of left vector sum), but not when the visual stimulus is /ba/. Comparison of this formant like pair is interesting, because sweep F2a matches the formant transition of the syllable ba, and sweep F2e matches

formant transition of the syllable ga. So some sort of asymmetrical effect may be present here, but this is uncertain, since 1) ANOVA showed interaction effect between audio and visual stimuli only with four "good" subjects and 2) the difference between activations was only present with all eight subjects. This effect might be present, however, if less noisy data, with a larger number of subjects (thus increasing statistical reliability), all of which would show "good" behavioral responses, would be used. F1 seems to differ most of the time from other formant glides, although the amplitudes differ significantly more often on the left side than on the right side, but as mentioned before, this lateralization effect is unreliable, since ANOVA didn't show any kind of lateralization effect.

Visual effect in this study is present only occasionally. Summarizing the results obtained both from all eight and from four "good" subjects, following can be observed: visual effect is only present in the left hemisphere, and is occasionally present with amplitudes (11 times out of 80), and very rarely with latencies (3 times out of 80). No consistent effects are present; the exact details of the effects are presented in Tables 4.5-4.7 and 4.15-4.19.

It is worth noting, that neurons, which are most sensitive to frequency modulated signal, are located in the belt areas, which are higher in hierarchy than primary auditory cortex (Rauschecker, 1998).

5.2 Suggestions for further studies

The formant effect in this study is a rather interesting phenomena and further studies are needed to explain, what causes it, or does it even exist for certainty. A study using fMRI could reveal the exact location of this effect, which can be located roughly to the auditory cortex based on the present MEG study.

The used sound stimuli in the experiment were highly simplified from natural formant transitions: natural formant transitions have much broader bandwidth than sine sweeps. The integration effect could be present with this kind of more elaborate stimuli. Also sine wave speech could be used as stimulus, and 2 different kinds of conditions could be used: 1) subjects wouldn't recognize sine wave speech as speech

and 2) subjects would recognize sine wave speech as speech. Usually subjects don't recognize sine wave speech as speech unless they are instructed in recognizing it. Those subjects, that do recognize it as speech without tutoring, could be omitted from the condition 1 by testing subjects before the actual experiment. In condition 2, subjects would be taught to recognize sine wave speech. Then results from the 2 different conditions could be compared separately and against each other.

One interesting possibility would be to *simultaneously* play several formant transitions like sine sweeps. These signals would be in a sense sine wave speech, except for the fact, that their length would be only 50ms and the signals would include only the formant transition part. So the signal would be a bit like in the present experiment, except, that multiple sweeps would be played simultaneously (matching F1, F2, and possibly also F3 and F4). This would preserve a simplified version of the frequencies which are present in the formant transition (the maximum spots would be approximately the same), but the spectrum would have spikes at the maximum spots, and almost no energy between the frequencies with maximum energy. This would be interesting, because the information, which vocal is in question, is (speaking simplified) somehow transmitted with the *relationships* of the formants, or at least this is one of the most important cues in detecting the vowel. Using this kind of stimuli might thus lead to a modulation effect, because information of the relative frequencies (F1/F0, F2/F1 etc.) would be present in the signal. However, as mentioned, this modulation effect would be expected to be present also in this study, since when other, similar studies are observed, this effect is present. One of these studies actually includes *simpler* audio signals then the present study (Kauramäki, 2006).

Also single or many simultaneous natural formant transitions could be played, and it could be observed, whether an integration effect would be present, if the auditory stimulus would have a normal bandwidth.

6. Bibliography

Ahveninen, J., Kähkönen, S., Tiitinen, H., Pekkonen, E., Huttunen, J., Kaakkola, S., Ilmoniemi, R.J., and Jääskeläinen, I.P. (2000). Suppression of transient 40-Hz auditory response by haloperidol suggest modulation of human selective attention by dopaminen receptors. *Neuroscience Letters,* 292(1): 29-32.

Arniold, J.C. and Milton, J.S. (1995). *Introduction to Probability and Statistics.* McGraw-Hill, Inc., 3rd edition: 535-584

Axer, H., Jantzen, J., Berks, G., Südfeld, D., and Keyserlingk, D.G.v. (2000). *The Aphasia Database on the Web: Description of a Model for Problems of Classification in Medicine.* ESIT 2000, 14-15 September 2000, Aachen, Germany

Beuchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004). Integration of Auditory and Visual information about objects in Superior Temporal Sulcus. *Neuron,* 41: 809-823

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., and Possing, E.T. (2000). Human temporal lobe activation by Speech and Nonspeech sounds. *Cerebral cortex,* 10(5): 512-528.

Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., and Vatikiotis-Bateson, E. (2004). Multisensory Integration Sites Identified by Perception of Spatial Wavelet Filtered Visual Speech Gesture Information. *Journal of Cognitive Neuroscience* 16(5): 805-816.

Calvert, G.A., Bullmore, E., Brammer, M., Cambell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., and David, A.S. (1997) Activation of Auditory Cortex During Silent Lipreading. *SCIENCE* 276: 593-596

Dahaene-Lambertz, G., Dahaene, S., and Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *SCIENCE* 298: 2013-2015

Desjardins, R.N. and Werker, J.F. (2004). Is the Integration of Heard and Seen Speech Mandatory for Infants? *Developmental psychobiology* 45(4): 187-203

Diehl, R.L. and Kluender, K.R. (1989) On the Objects of Speech Perception. *Ecological Psychology* 1:121-144

Dubin, M. (2001). Brodmann areas. Retrieved from http://users.tkk.fi/~jkaurama/dippa/dippa_FINAL.pdf

Eimer, M. and Driver, J. (2001). Crossmodal links in endogenous and exogenous spatial attention: evidence from event-related brain potential studies. *Neuroscience and Biobehavioral Reviews* 25: 497-511

Fowler, C.A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America* 99:1730-1741.

Gazzanica, M.S., Ivry, R.B., and Mangun, G.R. (2002). *Cognitive Neuroscience.* W. W. Norton & Company, Inc., 2nd edition.

Goldstein, E.B. (2002). *Sensation and perception.* Wadsworth Group, 6th edition

Hackett, T.A., Stepnievska, I., and Kaas, J.H. (1998). Subdivisions of Auditory Cortex and Ipsilateral Cortical Connections of the Paranbelt Auditory Cortex in Macaque Monkeys. *The Journal of Comparative Neurology,* 394: 475-495.

Hari, R. (1998). Magnetoencephalography as a Tool of Clinical Neurophysiology. In Niedermeyer, E. and Lopes da Silva, F., editors, *Electroencephalography: basic principles, clinical applications, and related fields*, chapter 60, pages 1107–1134. Lippincott Williams and Wilkins, 4th edition.

Hockley, N.S. and Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *The Journal of the Acoustical Society of America*, 96(5): 3309.

Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J. and Lounasmaa, O.V. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2): 413-497.

Kandel, E.R., Schwartz, J.H., and Jessell, T.M. (1991). *Principles of Neural Science.* Elsevier Science Publishing Co., Inc.; 3rd edition

Karjalainen, Matti (1999). *Kommunikaatioakustiikka.* Otamedia Oy.

Kauramäki, Jaakko (2006). Human auditory cortex is tuned by lip-reading: an MEG study. Personal communication, article to be published

Kuhl, P.K. and Meltzoff, A.N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100(4): 2425-2438.

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

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

Locke, J.L. (1993). *The Child's Path to Spoken Language.* The President and Fellows of Harvard College.

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage*, 21: 725-732.

Massaro, D.W. (1999). Speechreading: illusion or window into pattern recognition. *Trends in Cognitive Sciences*, Vol. 3, No. 8: 310-317

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

Möttönen, R. (2004): *Cortical mechanisms of seeing and hearing speech*, Helsinki University of Technology Laboratory of Computational Engineering Publications

Ojanen, V. (2005): *Neurocognitive mechanisms of audiovisual speech perception*, Helsinki University of Technology Laboratory of Computational Engineering, Technical Report B-49

Purves, D., Augustine, G.J., and Fitzpatrick, D. (2001). *Neuroscience.* Sinauer Associates, Inc., 2nd edition

Raij, T., Uutela, K., and Hari, R. (2000). Audiovisual integration of letters in the Human Brain. *Neuron*, 28: 617-625

Rauschecker, J.P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8: 516-521

Rauschecker, J.P., Tian, B., Pons, T., and Mishkin, M. (1997). Serial and Parallel Processing in Rhesus Monkey Auditory Cortex. *The Journal of Comparative Neurology*, 382: 89-103.

Reisberg, D., McLean, J., and Goldfield, A. (1987). *Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli.* In: Hearing by Eye: The Psychology of Lip-reading (Dodd B, Cambell R, eds), pp 97-113. London: Lawrence Erlbaum Associates

Robert-Ribes J., Schwartz J.L., Lallouache T., and Escudier P. (1998) Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual

identification of French oral vowels in noise. *The Journal of the Acoustical Society of America,* 103:3677-3689

Rossing, T.D., Moore, F.R., and Wheeler, P.A. (2002). *The Science of Sound.* Pearson Education, Inc., 3rd edition.

Scott, S.K. and Johnsrude, I.S. (2003). The neuroanatomical and functional organization of speech perception. *TRENDS in Neurosciences,* 26(2): 100-107.

Sumby, Wh. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America* 26:212-215

Wallace, M.T., Ramachandran, R., and Stein, B.E. (2004). A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences of the United States of America,* 101(7): 2167-2172

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

Winer, B.J. (1962). *Statistical Principles in Experimental Design.* McGraw-Hill, Inc.

Appendix A

Presentation scripts used for study

A.1 Code for visual stimulus

A.1.1 visual.sce

```
pcl_file = "visual_program.pcl";

TEMPLATE  "visual_2.sce";


array {

   trial ba;

   trial ga;

   trial stillface;
}trialss;

array {
   trial ba;

   trial ga;

   trial stillface;
}trialss2;
```

A.1.2 visual_2.sce

```
 #Adapted partially from av_template_vis.sce
   #Modified by Juuso Tujunen
   write_codes = true;
   default_trial_type = fixed;
   scenario_type = trials;


   default_picture_duration = 32;
   active_buttons = 1;
   button_codes = 0;
                           pulse_width = 20; #copied from somewhere
               response_matching = simple_matching;
begin;
   picture {
      bitmap { filename = "ba/ba0.bmp"; };
      x = 0; y = 0;
      } default;
   LOOP $i 41;

               picture {
                  bitmap { filename = "ba/ba$i.bmp"; };
                  x = 0; y = 0;
                  } "ba$i";

               picture { bitmap { filename = "ga/ga$i.bmp"; }; x = 0; y = 0; } "ga$i";

   ENDLOOP;

trial {

   stimulus_event {
      nothing {};
                           port_code = 1;
   } eventba;

   picture ba0;
   time = 0;
   code = "videonalku_ba";
   picture ba1;
   time = 32;
   picture ba2;
   time = 64;
   picture ba3;
   time = 96;
   picture ba4;
   time = 128;
   picture ba5;
   time = 160;
   picture ba6;
   time = 192;
   picture ba7;
   time = 224;
   picture ba8;
   time = 256;
   picture ba9;
   time = 288;
   picture ba10;
   time = 320;
   picture ba11;
   time = 352;
   picture ba12;
   time = 384;
   picture ba13;
   time = 416;
   picture ba14;
   time = 448;
   picture ba15;
   time = 480;
   picture ba16;
   time = 512;
   picture ba17;
   time = 544;
   picture ba18;
   time = 576;
```

```
   picture ba19;
   time = 608;
   picture ba20;
   time = 640;
   picture ba21;
   time = 672;
   picture ba22;
   time = 704;
   picture ba23;
   time = 736;
   picture ba24;
   time = 768;
   picture ba25;
   time = 800;
   picture ba26;
   time = 832;
   picture ba27;
   time = 864;
   picture ba28;
   time = 896;
   picture ba29;
   time = 928;
   picture ba30;
   time = 960;
   picture ba31;
   time = 992;
   picture ba32;
   time = 1024;
   picture ba33;
   time = 1056;
   picture ba34;
   time = 1088;
   picture ba35;
   time = 1120;
   picture ba36;
   time = 1152;
   picture ba37;
   time = 1184;
   picture ba38;
   time = 1216;
   picture ba39;
   time = 1248;
   picture ba40;
   time = 1280;

   code = "trialinloppu_ba";
} ba;

   trial {


     stimulus_event {
      nothing {};
              port_code = 2;
#target_button=1;
   } eventga;
   picture ga0;
   time = 0;
   code = "videonalku_ga";
   picture ga1;
   time = 32;
   picture ga2;
   time = 64;
   picture ga3;
   time = 96;
   picture ga4;
   time = 128;
   picture ga5;
   time = 160;
   picture ga6;
   time = 192;
   picture ga7;
   time = 224;
   picture ga8;
   time = 256;
   picture ga9;
   time = 288;
   picture ga10;
   time = 320;
   picture ga11;
   time = 352;
   picture ga12;
   time = 384;
   picture ga13;
   time = 416;
   picture ga14;
```

```
   time = 448;
   picture ga15;
   time = 480;
   picture ga16;
   time = 512;
   picture ga17;
   time = 544;
   picture ga18;
   time = 576;
   picture ga19;
   time = 608;
   picture ga20;
   time = 640;
   picture ga21;
   time = 672;
   picture ga22;
   time = 704;
   picture ga23;
   time = 736;
   picture ga24;
   time = 768;
   picture ga25;
   time = 800;
   picture ga26;
   time = 832;
   picture ga27;
   time = 864;
   picture ga28;
   time = 896;
   picture ga29;
   time = 928;
   picture ga30;
   time = 960;
   picture ga31;
   time = 992;
   picture ga32;
   time = 1024;
   picture ga33;
   time = 1056;
   picture ga34;
   time = 1088;
   picture ga35;
   time = 1120;
   picture ga36;
   time = 1152;
   picture ga37;
   time = 1184;
   picture ga38;
   time = 1216;
   picture ga39;
   time = 1248;
   picture ga40;
   time = 1280;
  code = "trialinloppu_ga";
} ga;

  trial {
      stimulus_event {
      nothing {};
              port_code = 0;
#target_button=1;
   } eventstillface;
   picture ba0;
   time = 0;
   code = "videonalku_stillface";
   picture ba0;
   time = 32;
   picture ba0;
   time = 64;
   picture ba0;
   time = 96;
   picture ba0;
   time = 128;
   picture ba0;
   time = 160;
   picture ba0;
   time = 192;
   picture ba0;
   time = 224;
   picture ba0;
   time = 256;
   picture ba0;
   time = 288;
   picture ba0;
   time = 320;
   picture ba0;
   time = 352;
```

```
    picture ba0;
    time = 384;
    picture ba0;
    time = 416;
    picture ba0;
    time = 448;
    picture ba0;
    time = 480;
    picture ba0;
    time = 512;
    picture ba0;
    time = 544;
    picture ba0;
    time = 576;
    picture ba0;
    time = 608;
    picture ba0;
    time = 640;
    picture ba0;
    time = 672;
    picture ba0;
    time = 704;
    picture ba0;
    time = 736;
    picture ba0;
    time = 768;
    picture ba0;
    time = 800;
    picture ba0;
    time = 832;
    picture ba0;
    time = 864;
    picture ba0;
    time = 896;
    picture ba0;
    time = 928;
    picture ba0;
    time = 960;
    picture ba0;
    time = 992;
    picture ba0;
    time = 1024;
    picture ba0;
    time = 1056;
    picture ba0;
    time = 1088;
    picture ba0;
    time = 1120;
    picture ba0;
    time = 1152;
    picture ba0;
    time = 1184;
    picture ba0;
    time = 1216;
    picture ba0;
    time = 1248;
    picture ba0;
    time = 1280;

    code = “trialinloppu_stillface”;
} stillface;
```

A.1.3 visual_program.pcl

```
#by Juuso Tujunen

#Initial time set at 0
loop until clock.time() >= 0 begin end;

#Here the code goes through ba-ga-still combinations
loop
int i = 0
until i > 900
begin

#Here the code organices the sets raondomly so, that the first
#videoclip of the n:th set isn't the same as the last videoclip
#of the n-1:th set (so one doesn't show the same videoclip
#consecutively)
if (i > 0) then
   trialss2 = trialss;
   loop
   trialss.shuffle();
   until
   trialss[1] != trialss2[3]
   begin
   trialss.shuffle();
   end;

else
   trialss.shuffle();
end;

#Here one goes through the randomized sets of ba, ga and stillface
loop
int j = 1
until j > 3

begin

   #Here one loops through a set a randomized time
   loop
   int l = clock.time() + random(20000, 40000);
   bool done = false;
   until done
   begin
   if clock.time() > l then
   done = true;
   end; #end of loop

   trialss[j].present();

   int m = random(100, 200);

   end;
   j = j + 1;
end;
   i = i + 1;
end;
```

A.2 Code for audio stimulus

A.2.1 phonmod.sce

```
# Original author Jaakko Kauramäki
#Modified by Juuso Tujunen
scenario = "Effect of attention on neural tuning";
# in this phase 1 answer buttons
active_buttons=1;
button_codes=128;
target_button_codes=128;
write_codes=true; # write all codes to parallel port (for EEG acquisition)
pulse_width=20; # seems to be ok
response_matching = simple_matching; # don't stop trial on answer, there is one
"correct" answer
default_monitor_sounds = false; # by default don't stop sounds
pcl_file = "phonmod.pcl"; # read volume info from file (att_tone.txt)

$att_tone=0.0; # attenuate 20dB by default


begin;

array{
sound { wavefile { filename = "formantti1_uus2.wav"; }; attenuation = $att_tone;}
tone1;
sound { wavefile { filename = "formantti21_uus2.wav"; }; attenuation = $att_tone;}
tone2;
sound { wavefile { filename = "formantti22_uus2.wav"; }; attenuation = $att_tone;}
tone3;
sound { wavefile { filename = "formantti23_uus2.wav"; }; attenuation = $att_tone;}
tone4;
sound { wavefile { filename = "formantti24_uus2.wav"; }; attenuation = $att_tone;}
tone5;
sound { wavefile { filename = "formantti25_uus2.wav"; }; attenuation = $att_tone;}
tone6;

} tones;

picture { } default;
picture { bitmap { filename = "fixation3.bmp"; }; x=0; y=0;} fixation;

# show default pic
trial {
   monitor_sounds = false;
   trial_duration = 1;
   trial_type = fixed;

   picture fixation;
};

# empty main trial (idea copied from Presentation help,
# files stimulus_event.pcl and stimulus_event.sce)
trial {

   trial_type=fixed;

   stimulus_event {
      nothing {};
   target_button=1;
   } event1;


} main_trial;
```

A.2.2 phonmod.pcl

```
#Original author Jaakko Kauramäki
#modified by Juuso Tujunen
int MIN_TONES=150*10; # minimum number of each tone
#array <int> port_codes[tones.count()]={1,2,4,8,16,32,3,5,6,7,9};
array <int> port_codes[tones.count()] = {1,2,4,3,5,6};
array <int> tone_count[tones.count()];
array <int> tone_order[tones.count()*(MIN_TONES+2)];

# create the tone_order[] array
# (repeat each tone MIN_TONES times)

#Looping, untill the initial time is 0
loop until clock.time() >= 0 begin end;

loop
   int i=1;
   int ndx=1;
until
   i>(MIN_TONES+2)
begin
   loop
      int j=1;
   until
      j>tones.count()
   begin
      tone_order[ndx]=j;
      j=j+1;
      ndx=ndx+1;
   end;
   i=i+1;
end;

# shuffle the order
tone_order.shuffle();

#Program randomices the stimuli so, that similar tones don't
#come up consecutevily


loop
  int i=2;
  int tone_n;
  int last_tone;
until
  i>tones.count()*(MIN_TONES+1)
begin
   tone_n = tone_order[i];


   if (tone_order[i] == tone_order[i-1]) then

      loop
         int ok=0;
         int ofs=1;
      until
         ok==1
      begin
         if(tone_order[i+ofs] != tone_order[i-1]) then

            int tmp=tone_order[i+ofs];
            tone_order[i+ofs]=tone_order[i];
            tone_order[i]=tmp;
            ok=1;
         else
            ofs=ofs+1;
         end; # end if
         if ofs>tones.count() then
           ok=1;
         end; # end if
      end; # end loop
   end; #ens if
   i=i+1;
end;

loop
  int i=tones.count()*(MIN_TONES+1);
  int tone_n;
  int last_tone;
until
  i>tones.count()*(MIN_TONES+2)
```

```
begin
   loop
      int ok=0
   until
       ok==1
   begin
       tone_n = random( 1, tones.count() );

   if (tone_n != last_tone) then
         ok=1;
       else
         ok=0;
       end;
   end;
   tone_order[i]=tone_n;
   last_tone=tone_n;
   i=i+1;
end;



loop
   int i = 1;
   int tone_n;  # index of the currently presented tone
   int last_tone; # index of the last presented tone
until
# comment extra lines away from below.. i.e. if
# tones.count() is only 5, comment out lines
# pointing to tone_count[6] and above
   (tone_count[1]>=MIN_TONES) && (tone_count[2]>=MIN_TONES) &&
   (tone_count[3]>=MIN_TONES) && (tone_count[4]>=MIN_TONES) &&
   (tone_count[5]>=MIN_TONES) && (tone_count[6]>=MIN_TONES) #&&
   #(tone_count[7]>=MIN_TONES) && (tone_count[8]>=MIN_TONES) &&
   #(tone_count[9]>=MIN_TONES) && (tone_count[10]>=MIN_TONES)
begin
  tone_n=tone_order[i];

   event1.set_stimulus( tones[tone_n] );
   event1.set_target_button( 1 );
   event1.set_event_code( "tone " + string( tone_n ) );
   event1.set_port_code( port_codes[tone_n] );

   #Below is the setting for the duration of sounds, which
   #varies randomly between 990-997 ms

   int m = random(990, 997);
   main_trial.set_duration(m);
   main_trial.present();

   last_tone=tone_n;
   tone_count[last_tone]=tone_count[last_tone]+1;

   i = i + 1
end;

loop
  int i=1
until
  i>tones.count()
begin
  term.print("count["+string(i)+"]="+string(tone_count[i])+"\n");
  i=i+1;
end;
```