Prior experience shapes speech perception: a behavioural and neuroimaging perspective

Maria Hakonen
Prior experience shapes speech perception: a behavioural and neuroimaging perspective

Maria Hakonen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at Auditorium F239a of the Department of Neuroscience and Biomedical Engineering on 2nd February 2020 at 12 noon.

Aalto University
School of Science
Department of Neuroscience and Biomedical Engineering
Brain and Mind Laboratory
Supervising professors
Professor Iiro Jääskeläinen, Aalto University, Finland
Professor Mikko Sams, Aalto University, Finland

Thesis advisors
Docent Hannu Tiihonen, Aalto University, Finland
Docent Patrick May, Lancaster University, England
Professor Iiro Jääskeläinen, Aalto University, Finland
Docent Miika Koskinen, University of Helsinki, Finland; Aalto University, Finland

Preliminary examiners
Professor Sonja Kotz, Maastricht University, German
Dr. Pia Rämä, Paris Descartes University, France

Opponent
Viljami Salmela, University of Helsinki, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 18/2020

© 2020 Maria Hakonen

ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2020

Finland
Abstract

Recent developments in neuroimaging and computational methods have allowed us to measure brain responses to naturalistic stimuli such as words, sentences, and narratives. This has increased our understanding of how the human brain process connected speech. However, it is still unclear how the brain grasps the meaning of acoustically distorted speech. Further, the question of why an interpretation of the same narrative can vary between individuals has remained unknown. This thesis illuminated these issues by studying how prior experience affects speech perception.

The first studies of the thesis demonstrated that initially unintelligible, acoustically distorted speech stimuli can be rendered intelligible by presenting them after only a single exposure to their respective intact counterparts. The intelligibility of sentences increased more than that of words, and vowels remained unintelligible. The contrast between the magnetoencephalography response to the distorted vowels presented before versus after the intact vowels revealed enhanced source current density in the auditory cortex and surrounding areas at the latencies of 130–160 ms from stimulus onset. The corresponding contrast between the first and second presentations of distorted sentences in functional magnetic imaging (fMRI) revealed modulations in responses generated in the primary auditory cortex and surrounding areas as well as in several extralinguistic brain areas that have been associated with memory and executive functions in previous studies. The final study showed that when subjects share a cultural family background, and therefore presumably have also accumulated more similar experience throughout their lifetime, this is reflected in the interpretation and neural processing of spoken narratives.

In sum, the results of this thesis show that prior experience can dramatically increase speech intelligibility in acoustically adverse conditions, and lexico-semantic information in long-term memory seems to be important in this process. The results provide further evidence that the activity in the auditory areas is not only modulated by auditory information but also by prior experience. Moreover, it seems that experience accumulated throughout the lifetime is reflected in speech processing. The results of this thesis increase knowledge of speech processing in acoustically suboptimal conditions. The results may also help to overcome potential challenges in mutual understanding between individuals from different cultural backgrounds.
Viimeaikainen kehitys laskennallisissa- ja neurokuvantamismenetelmissä on mahdollistanut aivovaloisten keraamisen aiempaa luonnonmukaisemmille ärskykeille, kuten sanoille, lauseille ja tarinoille. Tämä on lisännyt ymmärrystämmeksi, miten ihmisaivot prosessoivat puhetta. On kuitenkin epäselvää, kuinka pystymme ymmärtämään äänenlaadultaan heikkenneytä puhetta. Epäselvää on myös se, miksi eri henkilöt usein tulkitsevat saman tarinan hyvin eri tavoin. Tämän vääntönäkön tavoitteena on valottaa näitä kysymyksiä tutkimalla aiemman kokemuksen vaikutusta puheen havaitsemiseen.

Käytyväntymistä mittaavat kokeet osoittivat, että äänenlaadultaan heikkenneytään äänenlaadultaan paremmin, kun ne esitetään samojen, mutta äänenlaadultaan normaalien äänenlaaduk Changes in normaalien vokaalien tunnistettavuus ei parantunut lainkaan. Kun äänenlaadultaan heikkenneytään vokaalit esitetään äänenlaadultaan normaalien vokaalien jälkeen, ne herättivät voimakkaammat magnetokenfelografi-aivakuvauksen ja sen ympäristössä 130–160 ms:n latenssella vokaalin alkamisesta. Toiminnallisessa magnetettikuvausen aiemmin esitetynä äänenlaadultaan normaalit läuseet heijastivat äänenlaadultaan heikkenneytjen läuseiden herättämiä aivovaloisiin useilla aivoalueilla, jotka on liitetty aiemmessa tutkimuksessa musiikkiin ja toimeenpanepiin prosesseihin. Viimeinen tutkimus osoitti, että koehenkilön perheen kulttuuritausta heijastuu seiihen, kuinka henkilö puhuttua tarinaa, että tarinan synnyttämiä aivovaloisiin.

Acknowledgements

The research for this thesis was conducted at the Department of Neuroscience and Biomedical Engineering, Aalto University School of Science. The work was financially supported by the Academy of Finland, the Emil Aaltonen Foundation, the Jane and Aatos Erkko Foundation, and the Medical Research Council (United Kingdom).

This work would not have been possible to do without the support and guidance that I received from many people. In the beginning, I would like to say thanks to my supervising professors Mikko Sams and Iiro Jääskeläinen for their support during this research project and for the opportunity to work with highly talented people at the Brain and Mind (BML) laboratory.

I want to express my deepest gratitude to my excellent advisors Docent Hannu Tiitinen, Docent Patrick May, Prof. Iiro Jääskeläinen and Docent Miika Koskinen for the opportunity to work and finish my doctoral thesis under their guidance. Without their help and patient encouragement, the completion of this thesis would not have been possible. I highly appreciate the all the time and effort they put in this thesis project. I could not have imagined having better advisors and mentors for my study.

I would like to extend my gratitude to all the colleagues with whom I had the pleasure to work with during these years. I present my special thanks for the valuable and educative collaboration to all co-authors, Jussi Alho, Prof. Paavo Alku, Emma Jokinen, Arsi Ikäheimonen, Annika Hultén, Janne Kauuttonen, Prof. Fa-Hsuan Lin and Anastasia Lowe. Warm thanks also go to the past and current BML members for their practical and mental support as well as nice overall atmosphere. I am also thankful for Juha Montonen and Suvi Lehto from BIOMAG laboratory, Mia Illman from Aalto University as well as Toni Auranen, Marita Kattelus and Tuomas Tolvanen from Advanced Magnetic Resonance Imaging Centre of Aalto University for their practical assistance in collecting the data. I also greatly appreciate the technical support I got from our IT team. Further, I would like to thank all my test subjects for their participation and patience.

Prof. Sonja Kotz from Maastricht University and Dr. Pia Rämä from University of Paris Descartes are greatly acknowledged for pre-examining this dissertation and giving encouraging feedback.

Last but not the least, I would like to thank my parents for their support throughout this thesis work and my life in general. Your constant...
encouragement, patience, optimism and advice was more valuable than you could ever imagine.

Espoo, December 2019

Maria Hakonen
## Contents

Acknowledgements ........................................................................................................... i
List of Abbreviations ......................................................................................................... v
List of Publications ............................................................................................................ vii
Author’s Contribution ....................................................................................................... ix

1. Introduction .................................................................................................................. 1
   1.1 Perceiving speech ..................................................................................................... 1
       1.1.1 Miracle of speech perception ........................................................................... 1
       1.1.2 Stages of processing in speech perception .................................................... 2
   1.2 Neural correlates of speech perception .................................................................... 3
       1.2.1 The peripheral auditory system ....................................................................... 3
       1.2.2 Time windows of speech processing in the auditory cortex ......................... 4
       1.2.3 Cortical brain structures involved in speech processing ............................... 5
       1.3 Prior experience in speech perception ................................................................. 7
       1.3.1 Speech perception and memory systems ......................................................... 7
       1.3.2 Prior experience increases speech intelligibility in acoustically suboptimal conditions .................................................................................................................. 9
       1.3.3 Prior experience shapes interpretation of narratives ..................................... 11
       1.3.4 Culture shapes human cognition ..................................................................... 12
   1.4 Aims of the study ..................................................................................................... 13

2. Materials and methods ................................................................................................. 15
   2.1 Subjects ................................................................................................................... 15
   2.2 Stimulus materials .................................................................................................. 15
   2.3 Magnetoencephalography ..................................................................................... 16
       2.3.1 Acquisition of MEG data ................................................................................. 17
       2.3.2 Preprocessing of MEG data ............................................................................ 18
       2.3.3 Sensor-level analysis ...................................................................................... 18
       2.3.4 Source localization ......................................................................................... 19
   2.4 Magnetic resonance imaging .................................................................................. 20
   2.5 Functional magnetic resonance imaging ............................................................... 21
       2.5.1 Increasing sampling-rate of fMRI ................................................................. 22
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG</td>
<td>angular gyrus</td>
</tr>
<tr>
<td>BOLD</td>
<td>blood-oxygen-level dependent</td>
</tr>
<tr>
<td>CBC</td>
<td>cerebellar cortex</td>
</tr>
<tr>
<td>CC</td>
<td>cingulate cortex</td>
</tr>
<tr>
<td>D-I-D</td>
<td>distorted-intact-distorted</td>
</tr>
<tr>
<td>dSPM</td>
<td>dynamic statistical parametric map</td>
</tr>
<tr>
<td>ECD</td>
<td>equivalent current dipole</td>
</tr>
<tr>
<td>EEG</td>
<td>electroencephalography</td>
</tr>
<tr>
<td>EPI</td>
<td>echo planar imaging</td>
</tr>
<tr>
<td>ERF</td>
<td>event related field</td>
</tr>
<tr>
<td>ERP</td>
<td>event related response</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional magnetic resonance imaging</td>
</tr>
<tr>
<td>FOC</td>
<td>frontal operculum</td>
</tr>
<tr>
<td>FP</td>
<td>frontal pole</td>
</tr>
<tr>
<td>GLM</td>
<td>general linear model</td>
</tr>
<tr>
<td>HG</td>
<td>Heschl’s gyrus</td>
</tr>
<tr>
<td>HRF</td>
<td>hemodynamic response function</td>
</tr>
<tr>
<td>IAT</td>
<td>implicit association test</td>
</tr>
<tr>
<td>ICA</td>
<td>independent component analysis</td>
</tr>
<tr>
<td>IFG</td>
<td>inferior frontal gyrus</td>
</tr>
<tr>
<td>Ins</td>
<td>insula</td>
</tr>
<tr>
<td>ISC</td>
<td>inter-subject correlation</td>
</tr>
<tr>
<td>LG</td>
<td>lingual gyrus</td>
</tr>
</tbody>
</table>
LOC  lateral occipital cortex
MEG  magnetoencephalography
MFG  middle frontal gyrus
MNE  minimum norm estimate
MRI  magnetic resonance imaging
MTG  middle temporal gyrus
OFC  orbitofrontal cortex
OP   occipital pole
OTC  occipitotemporal cortex
OTFC ocipitotemporal fusiform cortex
PCG  precentral gyrus
PCun precuneus
PET  positron emission topography
PHG  parahippocampal gyrus
PT   planum temporale
ROI  region of interest
SMG  supramarginal gyrus
smsInI simultaneous multislice inverse imaging
SPL  superior parietal lobue,
STG  superior temporal gyrus
STS  superior temporal sulcs
TE   echo time
TR   repetition time
tSSS temporal Signal-Space Separation
USQ  uniform scalar quantification
List of Publications

This doctoral dissertation consists of a summary and the following two journal publications as well as one submitted manuscript which are referred to in the text by their roman numerals.


Author’s Contribution

**Study I:** Previous exposure to intact speech increases intelligibility of its digitally degraded counterpart as a function of stimulus complexity.

The candidate planned experiments with co-authors, prepared stimulus material, recruited subjects, acquired and analyzed behavioral and MEG data as well as wrote the first version of the manuscript and edited the final version based on the input from the other authors. Assistance for the acquisition and analysis of the behavioral data was received from Emma Jokinen.

**Study II:** Predictive processing increases intelligibility of acoustically distorted speech: Behavioral and neural correlates.

The candidate planned experiments with co-authors, recruited the subjects, acquired and analyzed behavioral and fMRI data as well as wrote the first version of the manuscript and edited the final version based on the input from the other authors. Assistance for the acquisition and analysis of the behavioral data was received from Emma Jokinen.

**Study III:** Processing of a spoken narrative in the human brain is shaped by family cultural background.

The candidate planned the experiments with co-authors, participated in the preparation of the stimulus material, recruited half of the subjects, acquired the fMRI data and behavioral data, analyzed the fMRI data and the data from the background questionnaire and implicit association test as well as wrote the first version of the manuscript and edited the submitted version based on the input from the other authors. The candidate supervised the analysis of the data from the word-association experiment. The candidate also pioneered the use of a novel experimental ultrafast fMRI sequence and determined the optimal measurement protocol (e.g. the length of the measurements and measurement parameters) with her advisor and the personnel of Advance Magnetic Resonance Imaging center. The candidate examined the optimal preprocessing of the ultrafast fMRI data and implemented the preprocessing scripts.
1. Introduction

1.1 Perceiving speech

1.1.1 Miracle of speech perception

The ability to understand spoken language is an indissociable feature of humanity. No other species have such a versatile communication system that allows expressing an infinite number of ideas, feelings, and thoughts using a finite number of auditory symbols. Indeed, language is also referred to as a window into the human mind (Pinker, 2007). Speech is essential for effective human interaction and cooperation, and it is one of the main factors that allowed the emergence of complex human societies.

Comprehension of spoken language is so natural and effortless to us that most of us never consider how miraculous an ability it indeed is. The speech signal contains thoughts of the speaker that he/she has encoded into a series of air vibrations by organs within his/her throat and mouth (Pisoni and Remez, 2005). Speech perception is a complex process where the message of the speaker is decoded from these air vibrations – as small as the diameter of an atom (Gazzaniga et al., 2002) – and transformed into meaning in the brain of the listener. This process is remarkably robust: In real world situations, the speech signal is never “clean” but comes in many variations, such as different accents and voices, and it is usually embedded in noise (Mattys et al., 2012; Peelle, 2018). Yet, our brains are very good at dealing with this kind of messy signal. Indeed, the ability of our mental speech recognition system to grasp the meaning of spoken language outperforms even the most sophisticated artificial speech recognition systems (Petkar, 2016).

The secret behind our capability to grasp the meaning of spoken language may lie in our ability to predict the most plausible interpretation of the speech signal using information acquired through prior experience (Davis et al., 2005; Giraud et al., 2004; Hannemann et al., 2007; Peelle, 2018; Rönnberg et al., 2010; Tiitinen et al., 2012). Thus, speech perception relies not only on the incoming, often ambiguous auditory input but also on what we remember from the past. The aim of this thesis is to shed light on the question of how our past experience shapes the way we understand spoken language.
1.1.2 Stages of processing in speech perception

Speech perception begins by identification and categorization of phonemes, the smallest units of spoken language that affect the meaning of an utterance (e.g. cap vs. tap, Gazzaniga et al., 2002). While in written language, spaces and punctuation marks serve as physical boundaries between linguistic units, the speech signal consists of continuous fluctuations of air pressure, and therefore lacks corresponding physical segmentation. Instead, in spoken language, the information about the boundaries of linguistic units is encoded into several acoustic features such as the fundamental frequency of the utterance and the duration and intensity of the phonetic elements (Cutler et al., 1997). Especially low-frequency information of the amplitude envelope in the 4–8 Hz band is important in speech segmentation (Peelle and Davis, 2012). The combination of these acoustic features creates different prosodic variables such as pitch, duration, loudness, and rhythm of speech which allow us to identify phonemes, words, and sentences from the continuous stream of spoken language (Cutler et al., 1997).

Mapping the auditory information of the speech signal onto mental representations of words starts already after a few milliseconds from the time the speech signal arrives to the ear (Gazzaniga et al., 2002). In the field of psycholinguistics, it has been proposed that semantic and syntactic information as well as the details of word forms (spelling and sound pattern) are stored in a mental lexicon. Several psycholinguistic models have been developed to describe in which format the lexical information is stored and how it is accessed (Gabriel, 2015). The majority of these models agree that the transformation from acoustical signal to linguistic meaning consists of several hierarchically organized processing stages. It is also generally accepted that there are many different levels of representations between the speech signal and its semantic meaning (e.g., acoustic features, phones, phonemes, morphemes, and words), although the number of levels and the units represented in each level are still under debate. One of the central findings is that the mental semantic representation is constructed incrementally, that is, the interpretation of the word does not start only after the end of the word (Gabriel, 2015). Instead, as a word evolves in time, it activates all word candidates in parallel in the mental lexicon that are consistent with the incoming speech input up to that time point. The number of candidates decreases when more perceptual information becomes available and this results finally in the selection of the mental representation that best matches the input.

However, many aspects of the speech perception process, such as the direction of the flow of the information between the processing stages and the format of stored prelexical and lexical representations, are still under investigation and vary between models of spoken word recognition. These models can be divided into modular or bottom-up models and interactive or top-down models depending on the direction of the information flow (Gazzaniga et al., 2002). The modular models are data-driven, allowing only unidirectional flow of information from the lower- to higher-level processing
stages. Thus, contextual information cannot influence the processes of lexical access or lexical selection but is integrated to bottom-up information only at the high-level processing stages. In contrast, interactive models propose that there are feedback connections between each consecutive stage of speech processing which allow the higher-order processing levels to modulate the processing at the lower levels. The details of the different models of spoken word recognition are beyond the scope of this thesis but can be found in the review article by Weber and Scharenborg (2012).

1.2 Neural correlates of speech perception

1.2.1 The peripheral auditory system

Sound waves enter the auditory system via the outer ear where they travel through the ear canal causing the ear drum at the end of the canal to vibrate (Impairments et al., 2004; Qing and Mao-li, 2009). In the middle ear, the ear drum vibrations are amplified by three tiny bones, the malleus, incus, and stapes, into the vibrations of the oval window, a membrane which separates the middle ear from the inner ear. The vibrations of the oval window elicit pressure waves in the fluid of the cochlea in the inner ear. The cochlea is a fluid-filled bony snail-shaped chamber where the pressure waves propagate from the base to the apex of the cochlea (Robles and Ruggero, 2001). The movement of the fluid stimulates the hair receptors at the organ of corti (Schwander et al., 2010), a specialized sensory epithelium along the length of the cochlea. As a result, the hair receptors send action potentials to the auditory cortex bilaterally via auditory pathway which consists of the cochlear nucleus and the superior olive in the brain stem, the inferior colliculus in the midbrain and the medial geniculate in the thalamus (Fig. 1, Qing and Mao-li, 2009). These areas of the precortical auditory system are tonotopically organized (Mann and Kelley, 2011). Tonotopic organization of an area means that each cell is tuned to a characteristic frequency, and that the distance in frequency between two sounds is mapped onto the physical distance between the respective cells tuned to these frequencies. Apart from analyzing frequency information, the precortical auditory system extracts information on the location of sound sources (Risoud et al., 2018).
1.2.2 Time windows of speech processing in the auditory cortex

Intracrrotical measurements in humans have demonstrated that, in primary auditory cortex, activation elicited by sounds commences at about 10 ms post-stimulus (Liégeois-Chauvel et al., 1991) and evolves then into the so-called middle-latency responses of the event-related potential (ERP) of the electroencephalogram (EEG, Liégeois-Chauvel et al., 1994). This is followed by the P1 response, peaking at around 50 ms, and the N1 response, a prominent landmark feature of the ERP with a peak latency around 100 ms (Näätänen and Picton, 1987). Magnetoencephalography (MEG) studies have shown that the N1m, the magnetic equivalent to the N1, represents activation in the superior temporal cortex bilaterally (Hari, 1990), predominantly in belt and parabelt (i.e., secondary) areas of auditory cortex (Inui et al., 2006). The N1/N1m response is sensitive to many acoustic-phonetic properties of speech such as frequency (Mäkelä et al., 2002), intonation (Mäkelä et al., 2004), formant transitions (Mäkelä et al., 2005), periodic structure of vowels (Alku et al., 2001; Yrttiaho et al., 2008) and phonetic features of consonants (Obleser et al., 2006). The amplitude and latency of the N1m response differs between speech and acoustically matched nonspeech stimuli (Tiitinen et al., 1999), especially in the left hemisphere (Parviainen et al., 2005), suggesting that the human brain is sensitive to acoustic-phonetic features of speech already at 100 ms after stimulus onset.

In the time interval of 150–200 ms from stimulus onset, acoustic stimuli typically elicit another transient response in EEG and MEG recordings referred to as the P2 or P2m response, respectively (Crowley and Colrain, 2004). Only a few studies have investigated the role of P2/P2m
response in speech perception, and they have shown P2 to be sensitive to sound intensity, pitch, and acoustic structure. Interestingly, a study by Tiitinen et al. (2012) revealed that the P2m may also reflect that auditory cortex and surrounding areas are sensitive to speech intelligibility.

During the presentation of words and sentences, the N1m response is followed by a sustained field that persists during the stimulus and is generated in the superior temporal cortex. It has been suggested that the sustained field reflects the processing of sound periodicity (Gutschalk et al., 2002; Keceli et al., 2012). Also, Tiitinen et al. (2012) showed that when distorted sentences are intelligible, they elicit a stronger sustained field than when the subject is unable to understand them. These fields are generated posterior to the primary auditory cortex, and their functional significance is still unclear.

1.2.3  **Cortical brain structures involved in speech processing**

The evidence obtained from lesion and neuroimaging studies in humans and intracortical neurophysiological studies in nonhuman primates suggest a dual-pathway model of speech comprehension in the human brain (Fig. 2, DeWitt and Rauschecker, 2012; Hickok, 2009; Hickok and Poeppel, 2007, 2004, 2000; Rauschecker, 2012; Specht, 2014). According to this model, the cortical processing of speech starts from spectrotemporal analysis of the speech signal in the primary auditory cortex. Subsequently, the processing of phonological information takes place in the middle-posterior superior temporal sulcus (STS). These initial stages of speech processing are largely bilateral, although, the processing of phonological properties may be slightly pronounced in the left hemisphere. After these initial stages, speech processing diverges into the ventral and dorsal streams presumably in the middle and posterior STS (Hickok and Poeppel, 2007), although it is still controversial where the ventral and dorsal streams exactly diverge (Specht, 2014). It is also unclear which brain structures should be included in each stream. According to a recent study (Fridriksson et al., 2016), the most plausible candidates for the ventral stream are the lateral temporal lobe structures extending to the inferior parietal lobe and the inferior frontal lobe. The dorsal stream is strongly left-lateralized and most likely involves anterior speech areas, including pars opercularis and premotor areas, from where it extends to the posterior supramarginal gyrus and straddles the edge of the notch of the Sylvian fissure at the boundary of the parietal and temporal lobes (area Spt, Fridriksson et al., 2016).
Figure 2. The dual stream model of the language processing (Hickok and Poeppel, 2007). Dorsal parts of the bilateral STG are proposed to process spectrotemporal information (shaded with green) and posterior half of the bilateral STG phonological information (shaded with yellow) of speech. Regions shaded red represent the ventral stream which is bilateral with a weak left-hemispheric bias. Regions shaded blue represent the dorsal stream which is left-lateralized. Abbreviations: Ins = insula, iFG = inferior frontal gyrus, MTG = middle temporal gyrus, ITS = inferior temporal sulcus, dSTG = dorsal superior temporal gyrus, a = anterior, p = posterior.

The ventral stream, also referred to as the “what” stream, is responsible for speech comprehension (Hickok, 2009; Hickok and Poeppel, 2007; Specht, 2014). The left-hemispheric ventral stream process phonetic, phonological, lexical, and semantic information whereas the right-hemispheric ventral stream is more involved in voice identification and the processing of prosody. In the left-hemispheric ventral stream, the auditory and phonetic processing is pronounced in the posterior parts of the temporal lobe whereas semantic and sentence processing predominantly takes place in the anterior parts of the temporal lobe. Moreover, speech processing, especially in the STS and anterior parts of the temporal lobe, is increasingly leftward lateralized during lexicos-semantic processing as compared with auditory-phonetic processing (Peelle, 2012; Specht, 2013). The middle and posterior STS contribute to speech comprehension and have been shown to be sensitive to speech intelligibility whereas anterior parts process mainly phonological and sub-lexical information (Peelle et al., 2010). The processing hierarchy in the ventral stream includes inter-hemispheric interactions as well as feed-forward and feedback connections and is also modulated by higher-order cognitive processes, such as attention and expectations. The dorsal stream, also referred to as the “where” stream, is responsible for sound localization, mapping acoustic signals to articulatory networks, phonological and long-term memory, and speech production (Hickok, 2009; Hickok and Poeppel, 2007; Rauschecker, 2012; Specht, 2014). It is still unclear to which extent the dorsal stream is involved in speech perception (Specht, 2014), but recent studies suggest that it may be important in speech comprehension in acoustically suboptimal conditions (Hervais-Adelman et al., 2012; Peelle, 2018). In addition of the brain areas included in the ventral and dorsal streams, many neuroimaging studies of speech comprehension have found activity in the
angular gyrus, supramarginal gyrus, basal ganglia, anterior insula, and cerebellum (for a review, see Specht, 2014). Although these brain areas are clearly involved in the speech comprehension processes, their inclusion into the model of speech comprehension is still under debate since they have also several nonlinguistic functions.

Recent developments in neuroimaging and data analysis techniques as well as increased computational resources have made it possible to collect and analyze larger and more complex datasets. This has resulted in a rapid increase in neuroimaging studies that use spoken narratives as stimuli (Hamilton and Huth, 2018; Huth et al., 2016; Lerner et al., 2011; Whitney et al., 2009; Yarkoni et al., 2008). Narratives provide an ecologically more valid way to study speech comprehension than do isolated speech stimuli. Compared to the comprehension of isolated words or sentences, narrative comprehension results in more widespread activation that extends into frontal and parietal areas, as well as into the precuneus and cingulate cortex (Huth et al., 2016; Lerner et al., 2011; Whitney et al., 2009; Yarkoni et al., 2008) and is also bilateral (for a review, see Jung-Beeman, 2005). These results indicate that focusing on responses to isolated speech stimuli provides only a limited window into speech processing in the human brain and, thus, more research using natural language as stimulation is needed to uncover how the brain integrates information over several sentences capturing the meaning of narratives that can stretch out over minutes and hours.

1.3 Prior experience in speech perception

1.3.1 Speech perception and memory systems

Because speech unfolds over time, speech perception requires integration of the incoming information into the representation of the past input. The role of memory and other cognitive skills in speech perception has recently become a central topic of research, as exemplified by the emergence of the new fields of cognitive hearing science (Arlinger et al., 2009) and auditory cognitive science (Holt and Lotto, 2008). The classical model of human memory contains three systems: sensory memory, short-term memory, and long-term memory (Atkinson and Shiffrin, 1968; Gazzaniga et al., 2002). These systems have been assumed to differ from each other in the duration and capacity of the information they support. However, it should be emphasized that the organization of memory is still under debate. For example, some researchers have proposed that short-term memory is equivalent to activated long-term memory (Cowan, 1999, 1988; Oberauer, 2009, 2002).

Sensory memory is the first stage of the storage of sensory input. In the case of auditory processing, it is also known as echoic memory, and it contains sensory-based representations of auditory information situated in the auditory cortex on the supratemporal plane. Further, it has large capacity and does not require attentional controlled processing (Alain et al., 1998; Gazzaniga et al., 2002). Representations in echoic memory have the shortest
lifetimes, with estimates varying from one second or less (Guttman and Julesz, 1963) to up to 4 seconds (Darwin et al., 1972) and even 10 seconds (Eriksen and Johnson, 1964; Sams et al., 1993).

Short-term memory differs from echoic memory in that it has a smaller capacity, but the lifetimes of the memory traces are longer, from several seconds to minutes (Gazzaniga et al., 2002). Moreover, while echoic memory is automatic, short-term memory requires attentional processing. Working memory is often used as a synonym for short-term memory. However, some researchers have posited these systems to be separate whereby short-term memory only serves as a storage of information whereas working memory allows also for the manipulation of the stored information (Chai et al., 2018; Cowan, 2008). Working memory seems to be important in segmentation of a speech into words (Jacquemot and Scott, 2006) as well as in speech comprehension under acoustically suboptimal conditions (Dryden et al., 2017; Füllgrabe and Rosen, 2016; Rönnberg et al., 2010; Zekveld et al., 2012).

Long-term memory encodes more permanent knowledge and skills (Gazzaniga et al., 2002). It contains models of the world, also referred to as schemas, that we learn through experience during our lifetime and that enable understanding of the world around us (Bar, 2009, 2007). Long-term memory is divided into declarative or explicit memory system and nondeclarative or implicit memory system based on the characteristics of information stored in them (Squire and Dede, 2015). Declarative memory allows conscious recollection of events and facts, whereas nondeclarative memory includes skill-based information that is accessed without consciousness (Squire and Dede, 2015). The classical division of declarative memory is into episodic memory for our personal history and semantic memory for world knowledge. Neuroimaging studies have identified several brain areas that are associated with semantic memory (for a meta-analysis, see Binder et al., 2009). These areas include the posterior inferior parietal lobe (IPL), the middle temporal gyrus (MTG), the fusiform and parahippocampal gyri, the dorsomedial prefrontal cortex, the inferior frontal gyrus (IFG), the ventromedial prefrontal cortex, and the posterior cingulate gyrus.

Priming is an interesting phenomenon of nondeclarative memory that refers to increased speed or accuracy of the perception of a stimulus after prior exposure to the same or related stimulus (Henson, 2003; Squire and Dede, 2015). Priming occurs, for example, when a word is recognized more quickly when it is presented after a semantically related word (e.g. the word “bread” is recognized faster following the word “backer” than following the word “nurse”), or when words with some missing letters are completed more likely with previously presented words than novel words. Importantly, priming does not require conscious or explicit recollection of the previously presented items. The evolutionary significance of priming is presumably more efficient and faster perception in familiar environments where things are typically encountered more often than once. In neuroimaging studies, priming typically results in reduced responses in the brain areas relevant to the task.
1.3.2 Prior experience increases speech intelligibility in acoustically suboptimal conditions

In everyday situations, speech intelligibility can be decreased for a number of reasons. These include accent, disfluencies (e.g. repairs, restarts and fillers), disorders (e.g. dysarthria, apraxia of speech, cleft palate), competing signals in the environment (e.g. noise, background babble), as well as distortions originated from the communication channel (e.g. filtering of speech on a telephone, Mattys et al., 2012). Receiver limitations such as sensorineural hearing impairments, incomplete knowledge of language (i.e. non-native listeners) and neurological deficits of language function (e.g. auditory agnosia, cortical deafness, Wernicke’s aphasia) may also decrease speech intelligibility (Mattys et al., 2012). The acoustic distortions of the speech signal can be divided into internal and external (Mattys et al., 2012; Miettinen, 2015). In internal distortion, the structure of the speech signal itself is degraded typically by the properties of the communicating channel or by compression of digital sound. Internal distortions decrease speech intelligibility more than external distortions. These, in turn, occur when the speech signal is degraded by other, external sounds which result in energetic or informational masking. In energetic masking, the masking auditory signal contains energy in the same frequency bands as the speech signal. This results in interference already in the basilar membrane of the cochlea and therefore decreases the availability of speech cues. In informational masking, the target and the masker can be acoustically dissimilar and, on this basis, can be easily segregated from one another. However, the masker interferes with the processing of the target for the reason that they activate the same late-stage cognitive processes. This happens, for example, when the masker is also a speech signal.

Prior experience and contextual information can improve speech comprehension in acoustically suboptimal conditions. For example, an acoustically distorted sentence is easier to recognize when it precedes a story that set the context for the sentence than when the sentence is presented in isolation (Drager and Reichle, 2001). Similarly, a word is easier to recognize within a sentence than when it is presented in isolation (Miller et al., 1951). The human language system can also learn to grasp the content of accented (Bent and Bradlow, 2003) or even severely distorted speech (Davis et al., 2005) by dynamically adapting to process the speech signal optimally. The improvements in speech comprehension obtained through adaptive learning can be long-lasting. For example, listeners exposed to time-compressed speech outperform naïve listeners even one year after their original exposure to time-compressed speech (Altmann and Duncan, 1993). The ability to comprehend speech in acoustically suboptimal conditions also correlates with working memory capacity as measured with the reading span test (Akeroyd, 2008; Rönnberg et al., 2008; Rudner et al., 2011; Zekveld et al., 2012). This likely reflects that speech understanding is supported by information stored in memory through prior experience.

Studying the neural basis of top-down mechanisms such as memory in speech comprehension has been challenging because a direct
comparison between brain responses to intact speech stimuli and acoustically degraded speech stimuli or speech stimuli masked with noise reflect not only brain mechanisms subserving speech comprehension but also the acoustic differences between the stimuli. In this case, it becomes difficult to tell whether the differences in the brain responses to intelligible and unintelligible speech reflect the acoustic structure of the stimulus or top-down mechanisms (e.g. memory or attention) subserving speech comprehension. To overcome this problem, Tiitinen et al. (2012) and Hannemann et al. (2007) used an experimental paradigm that basis on the priming effect (see 1.3.1). In this paradigm, acoustically distorted, initially unintelligible speech stimuli were rendered intelligible for the subject by presenting exactly the same stimuli again, but this time exposing the subject first to a single presentation of the acoustically intact, intelligible counterparts of the stimuli. In the following, we refer to this paradigm as a “D-I-D” (i.e. distorted-intact-distorted) paradigm. The D-I-D paradigm is remarkable, because it allows for rapid manipulations of the intelligibility of a signal without changing the signal’s acoustic properties. This provides a unique possibility to investigate in a highly controlled way how top-down processes, such as memory, contribute to speech comprehension. Using the D-I-D paradigm with sentences, Tiitinen et al. (2012) showed that increased intelligibility was reflected as enhanced responses in the auditory cortex as well as in regions anterior and posterior to this area in the time range of the P2m response (>200 ms). In the time range of the sustained field (> 300 ms), the areas posterior to the auditory cortex showed enhanced activity with increased speech intelligibility. Hannemann et al. (2007) found improved intelligibility to be associated with increased gamma band activity in the left temporal regions at around 350 ms.

Griaud et al. (2004) used fMRI recordings and a related paradigm where initially unintelligible noise-vocoded sentences became intelligible as a result of perceptual training during lengthy sessions where the subject was exposed to repetitive presentations of distorted speech stimuli paired with their acoustically intact counterparts. The results suggested that speech comprehension is reflected as increased fMRI-responses in the bilateral medial and inferior temporal regions, and that both auditory search and speech comprehension are associated with increased fMRI-responses in posterior superior temporal cortex. Further, attention increased fMRI-responses in the dorsal part of Broca’s area, and auditory attention and comprehension interacted in bilateral insula, the anterior cingulate, and the right medial frontal cortex. Möttönen et al. (2005) used a similar approach as Griaud and colleagues to show that the left posterior STS elicited enhanced activity in the post- vs. pre-training session.

More recently, a number of studies have manipulated the intelligibility of noise-vocoded speech stimuli by presenting them immediately after the disambiguating intact stimulus (intact speech, text or picture; Clos et al., 2014; Sohoglu et al., 2012; Wild et al., 2012; Zekveld et al., 2012). These studies have associated the left hemispheric MTG (Clos et al., 2014) and angular gyrus (AG; Clos et al., 2014; Zekveld et al., 2012) with the
comprehension of distorted speech and the left IFG with auditory search (Clos et al., 2014; Zekveld et al., 2012). Interestingly, Sohoglu et al. (2012) showed that the increase in intelligibility of noise-vocoded words obtained through matching them with written text was first reflected in the activity in the IFG and only after that in the activity of the superior temporal gyrus (STG). This was interpreted to reflect the modulation of the processing of auditory input in the STG caused by feedback connections from the IFG, which presumably stores information on the matching text. Wild et al. (2012) provided further evidence for top-down effects by revealing that activity already within the primary auditory cortex reflects intelligibility of the distorted sentences that are disambiguated using written text.

There are several issues which the above studies on the comprehension of degraded speech have left unaddressed. First, each previous study used only a single type of speech stimulus, either words or sentences. It is therefore unclear how the complexity of the message (e.g., by increasing on the vowels-words-sentences axis) affects intelligibility. Second, the D-I-D paradigm was used to investigate a limited set of event-related responses in temporal and parietal brain areas in Tiitinen et al. (2012) or oscillations in Hannemann et al. (2007). It would therefore be interesting to use the D-I-D paradigm in fMRI experiment to map out the brain regions associated with making distorted speech intelligible. While this has already been partly achieved through using immediate pairing the distorted signal with a disambiguating stimulus (Clos et al., 2014; Zekveld et al., 2012), it could be argued that the D-I-D paradigm offers stimulation conditions which are ecologically more valid. This is because disambiguating stimuli are not immediately available in real-world conditions. Therefore, it is likely that decoding distorted stimuli in the real world relies more on the rapid activation of long-term memory representations, as in the D-I-D paradigm, rather than on resources of echoic and short-term memory, as in the case of immediate disambiguation.

1.3.3 Prior experience shapes interpretation of narratives

Prior contextual information has been shown to be important in narrative interpretation (Smirnov et al., 2014; Yeshurun et al., 2017). For example, the comprehension of spoken narratives increases when the narrative is presented after a picture that matches contextually with the narrative, and this is accompanied by activity in brain areas of a fronto-parietal network and of the default mode network (Smirnov et al., 2014). In another study looking at fMRI responses during narrative listening, prior contextual information that manipulated the subjects’ beliefs was found to increase the inter-subject correlations of the brain responses in the brain areas of the mentalizing network (medial prefrontal cortex, precuneus bilateral AG), in the premotor cortex, as well as in a subset of brain areas typically activated during narrative listening (bilateral anterior IFG and orbitofrontal cortex; Yeshurun et al., 2017). However, recent studies have shown that interpretation of the narrative can also spontaneously vary between subjects (Saalasti et al., 2019; Ngyen et
al., 2019). Saalasti et al. (2019) found that the similarity in the way subjects interpreted spoken narrative predicted the inter-subject similarity of fMRI activity in the bilateral AG and SMG. Looking at the way subjects interpret a movie, Nguyen et al. (2019) found a similar prediction for a more extended set of brain areas including premotor cortex, the right AG, bilateral STG, and dorsomedial prefrontal cortex (i.e. the regions of the default mode network). These brain areas are likely to contribute to the way narratives are interpreted and to the inter-individual variations in these interpretations. A recent study also shows that fMRI responses elicited by a movie were exceptionally similar between friends in the bilateral SMG, the left inferior parietal lobe including AG, the nucleus accumbens, the amygdala, the putamen and the caudate nucleus (Parkinson et al., 2018). This similarity was decreased when the inter-subject distance in the social network increased. However, thus far, it has remained unclear whether groups of people who tend to accumulate more similar experiences during their lifetime also interpret and process narratives in a similar way.

1.3.4 Culture shapes human cognition

People from the same culture tend to share similar knowledge, experiences, beliefs, values, attitudes, and social habits. This is a societally important phenomenon as it facilitates shared understanding and fosters smooth cooperation (McPherson et al., 2001). Culture acts as a lens that guides our attention onto specific aspects of the environment and that also shapes the contents and organization of semantic memory (for reviews, see Gutchess et al., 2011; Gutchess and Indeck, 2009). For example, East Asians tend to organize information based on thematic relationships (e.g. cow-grass; Chiu, 1972; Gutchess et al., 2006; Unsworth et al., 2005) and focus more on background and contextual information (holistic thinking; Chua et al., 2005; Nisbett and Miyamoto, 2005), whereas Westerns tend to organize information based on categorical relationships (e.g. cow-chicken; Chiu, 1972; Gutchess et al., 2006; Unsworth et al., 2005) and focus more on objects (analytical thinking; Chua et al., 2005; Nisbett and Miyamoto, 2005). The investigation of the neural basis of cultural influences on cognition has recently emerged as a new field, that of cultural neuroscience (Ames and Fiske, 2010; Chiao et al., 2010). Within this field, neuroimaging studies have found cross-cultural differences, for example, in the brain activity associated with the perception of objects and colors (for a review, see Ames and Fiske, 2010) and, interestingly, also with subjective taste preferences (McClure et al., 2004) which demonstrates that the cultural influence on perception is not limited to the visual domain. Behavioral cross-cultural studies have shown that the cultural background of the listener can shape the way the listener understands the culturally specific elements in the narrative (Anderson and Barnitz, 1984). However, the way cultural background is reflected in brain activity during narrative listening has remained unexplored.
1.4 Aims of the study

The overall aim of this thesis was to study the effects of prior experience on speech perception. In the first two studies, we used the D-I-D paradigm to investigate how prior experience can dramatically increase the intelligibility of acoustically severely distorted speech. The third and final study of this thesis examined whether experience accumulated throughout life biases speech processing and interpretation. Since people who share a cultural background tend to share similar experiences, we approached this question by looking at the behavioral and fMRI responses of two groups of subjects with different cultural family backgrounds.

The specific research questions of the individual studies were:

**Study I**: How and when does brain activity index processing that is making a distorted speech stimulus intelligible? How does the complexity of the stimulus, in terms of it being a vowel sound, a word, or a sentence, affect this processing?

**Study II**: Which brain areas reflect the processes whereby a distorted sentence is interpreted to be intelligible?

**Study III**: Does family cultural background bias the perception of an auditory narrative? Does this result in individuals sharing a cultural background having increased synchrony in terms of narrative interpretations and associated brain activity?

The studies took advantage of the high temporal resolution of MEG and the high spatial resolution of fMRI. Thus, in Study I and II, we were able to investigate in which brain areas and in which time windows does prior experience modulate speech processing in the human brain. In Study I, we used MEG to study whether prior exposure to intact speech can affect the processing of its acoustically distorted counterpart even when the subject is not paying attention to the speech stimuli. Studying speech comprehension in an unattended condition would not be possible with behavioral methods. In Study II, we used fMRI to map the brain areas which involved when the subject actively makes sense of distorted speech. Finally, in Study III, fMRI made it possible to examine at which level of the speech processing hierarchy the subject’s cultural background affects the way a spoken narrative is interpreted.
2. Materials and methods

2.1 Subjects

Altogether 105 subjects participated in the experiments of this thesis (see Table 1). All subjects were fluent Finnish speakers, had normal hearing, and reported no neurological disorders. For a half of the subjects in Study III, one or both of their parents were Russian. These subjects were born in Finland or had moved to Finland in their early childhood (three subjects at the age of < 3 years old, one at the age of 8 years). They were exposed to both Finnish and Russian in the first years of their life or started to acquire Finnish at the age of three years, except one who started to acquire Finnish at the age of 8 years. They had completed Finnish primary school.

All subjects provided written informed consent. In Study I, the experiments were approved by the Ethics Committee of Helsinki University Central Hospital and in studies II and III by the Aalto University Research Ethics Committee.

Table 1. Description of the subjects participated in the studies of this thesis and the type of the experiments they participated. * MEG results neither included in publication III and nor in this thesis.

<table>
<thead>
<tr>
<th></th>
<th>Study I</th>
<th>Study II</th>
<th>Study III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>16(8)</td>
<td>16(8)</td>
<td>24(12)</td>
</tr>
<tr>
<td>Number of females</td>
<td>8</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Mean age in years</td>
<td>22.4±3.7</td>
<td>22.7±1.6</td>
<td>24.7±4.5</td>
</tr>
<tr>
<td>Left-handed</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Behavioral MEG</td>
<td>x</td>
<td>x</td>
<td>x*</td>
</tr>
<tr>
<td>Behavioral fMRI</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>fMRI</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

2.2 Stimulus materials

To study how the complexity of the message is reflected in intelligibility of distorted speech in the D-I-D paradigm in Study I (see 1.3.3 for description of the D-I-D paradigm), a stimulus material consisting of Finnish vowels (0.2 sec
in duration), words (0.31–1.40 sec in duration; mean 0.65 sec; SD 0.18), and sentences (1.7–4.6 sec in duration; mean 3.1 sec; SD 0.6 sec) was synthesized using the Bitlips TTS synthesizer (44.1-kHz sampling rate, 16-bit amplitude resolution). The vowel set comprised all eight vowels of the Finnish language (/a/, /e/, /i/, /o/, /u/, /y/, /ä/, /ö/). The sentence set included 192 Finnish sentences, each of which began with the vowel /a/, /e/, /i/, or /u/. The word set was constructed by separating the first word of each sentence in the sentence set. The distorted stimuli were created by low-pass filtering the intact stimuli with a 2.2-kHz cut-off value, resampling them at a rate of 4.41 kHz, and reducing the amplitude resolution (bit rate) of the signals through the 1-bit uniform scalar quantification (USQ) method (Gray, 1990). 1-bit USQ approximates each sample of the signal to the nearest of the two permitted levels resulting in a rectangular speech signal with new noisy harmonic frequencies. In Study II, we used the same distorted and intact sentences that were used in Study I.

In Study III, the stimulus was a 71-min fictional auditory narrative telling a story of two young men, one with a Finnish and the other with a Russian background. The objective of using the story as stimulus material was that the subjects with only a Finnish background would identify themselves with the Finnish characters and understand the Finnish-culture specific elements. Conversely, the subjects with Finnish-Russian background would identify themselves mainly with the Russian characters and understand the Russian-culture specific elements better than the subjects with only a Finnish background. In the story, the man with the Finnish background is dating a Russian woman and the man with the Russian background a Finnish woman. However, the cultural differences result in conflicts between the characters and difficulties understanding each other and, at the end of the story, the man with the Finnish background ends up in a relationship with the Finnish woman and the man with the Russian background with the Russian woman. The story includes elements specific to Finnish and Russian cultures and the characters are described as having characteristics typically associated with Finns and Russians. However, care was taken not to overemphasize these characteristics so as to present the characters as realistic persons. The narrative included both periods describing interactions between the characters and periods describing the environment without social interactions. The narrative was voice-fitted by a professional narrator and recorded at the sampling rate of 44.1 kHz.

2.3 Magnetoencephalography

MEG is a non-invasive technique for measuring the magnetic fields generated by neural activity in the brain. MEG can measure neural activity at the temporal resolution of a millisecond which makes it particularly relevant for investigations of time-evolving characteristics of speech perception (Hamalainen et al., 1993). In favourable measurement conditions, MEG can also achieve a spatial discrimination of 2–3 mm for cortical sources. To
measure the tiny magnetic fields, typically in order of 100 fT (10^-8 times weaker than earth’s magnetic field), the measurement need to be conducted in a magnetically shielded room using extremely sensitive sensors (Hamalainen et al., 1993; Hari and Puce, 2017). The most commonly used sensors are superconducting quantum interference device sensors that are cooled in supra-conducting state with liquid helium.

MEG is sensitive to the component of currents which is tangential to the scalp (Hamalainen et al., 1993; Hari and Puce, 2017). Moreover, synchronous activity of many neurons with a parallel orientation is needed to generate a magnetic field strong enough to be detectable in MEG. These conditions are met by the apical dendrites of pyramidal cells located in the cortical sulci (Fig 3). These dendrites are oriented perpendicular to the cortical surface and, therefore, in the sulci they run parallel to the scalp. Thus, their intracellular currents are tangential to the scalp and, therefore, act as the main source currents contributing to MEG signals. Given that the strength of the apical dendrite can produce 0.2 pAm, synchronous activity of 50,000 cells is required to produce a magnetic field of 10 nAm detectable by MEG (Hamalainen et al., 1993; Murakami and Okada, 2006).

![Figure 3](image)

**Figure 3.** The origin of the MEG signal. a) MEG signal originates from the currents within apical dendrites of cortical pyramidal neurons that are oriented perpendicularly to the cortical surface. Each of the pyramidal cells can be modelled as a current dipole (blue arrows). b) Electric current generates a magnetic field that curls around it in a ring. c) When ~50,000 or more pyramidal cells get activated synchronously, the electric currents in their apical dendrites collectively sum to net currents at the cortical surface. The net magnetic field generated by the net current can be detected outside of the head with SQUID-sensors.

### 2.3.1 Acquisition of MEG data

MEG-responses for Study I were acquired using a 306-channel whole-head MEG neuromagnetometer (Vectorview 4-D, Neuromag Oy, Finland) in a magnetically shielded room at BioMag Laboratory of Helsinki University Central Hospital. During the measurement, the signals were sampled at 1.2 kHz. Horizontal and vertical eye movements were monitored with electro-oculography using two pairs of electrodes. To determine the position of the
The subject’s head within the MEG helmet, four head position indicator coils were attached to the mastoids and forehead of the subject. The locations of the coils with respect to the three anatomical landmarks (nasion and two bilateral preauricular reference points) were determined using a 3-D digitizer. Before each measurement, the subject’s head position with respect to the MEG sensor array was determined by briefly feeding current to the coils. The subject was instructed to lean with the back of the head on the helmet-shaped inner surface of the MEG device and remain stationary during the measurement. The auditory stimuli were delivered to the subject’s ears as a mono signal at the sound intensity of 70 dB SPL through a pair of plastic tubes and ear pieces.

2.3.2 Preprocessing of MEG data

MEG data was first filtered using the temporal Signal-Space Separation method (Taulu et al., 2003; Taulu and Simola, 2006; tSSS, Max-Filter software, Elekta Neuromag). For analyzing transient responses, the N1m and P2m, of the event-related fields, the data was band-pass filtered at 2–30 Hz and for analyzing the sustained response of the event-related field it was low-pass filtered at 30 Hz. Thereafter, the data was epoched and the epochs were baseline-corrected by relativizing them with respect to the pre-stimulus interval of -100 ms to 0 ms. Finally, the epochs were time-locked to stimulus onsets. For the sustained field analysis, we selected the epoch responses corresponding to 112 words that exceeded 560 ms in duration. The sustained field analysis of the sentences was also restricted to the responses of the corresponding subset of 112 sentences starting with the words exceeding 560 ms in duration. The epochs exceeding 2000 ms in duration were automatically rejected from both the transient and sustained field analyses. The transient and sustained responses were averaged over post-stimulus time windows of 500 ms and 600 ms, respectively. Artifacts resulted from eye movements were reduced using independent component analysis (ICA) applied on epoched data.

2.3.3 Sensor-level analysis

Nine pairs of gradiometers centred over the left and right auditory cortices were divided into anterior, medial and posterior subsets of three gradiometer pairs each. An average was calculated over the responses measured by the pairs in each subset. Thereafter, vector sums were calculated from the averaged responses of the anterior, medial and posterior gradiometer pairs. The amplitudes of the N1m and P2m responses were determined as the local maximum response values between the time intervals of 110–170 ms and 180–300 ms, respectively. The peak latencies were determined as the time of the maximum in relation to stimulus onset.

For the sustained field analysis, the gradiometer pairs were divided into ten subsets based on their location: occipital, parietal, left and right frontal, left and right temporal, sensorimotor and occipitotemporal subsets. Averages were calculated over the magnitudes of the gradiometer-
pair vector sums in each subset. The magnitude of the sustained field was determined from each of the averaged responses as the mean of the response in the 400–560 ms time window.

2.3.4 Source localization

Determining the location of the current sources that generated the magnetic fields measured by MEG requires solving an inverse problem (Hamalainen et al., 1993; Hansen et al., 2010; Hari and Puce, 2017). This refers to determining the causes using observations of their effects. The inverse problem is ill-posed, that is, it does not have a unique solution. In the context of MEG, this means that an infinite number of different current distributions can produce the same magnetic fields measured by MEG. However, the most plausible of the current distributions can be determined by constructing computational models utilizing a priori knowledge. This model, referred to as the forward solution, gives the magnetic fields if the current sources are known. The localization of the sources of MEG signals always has some uncertainty since it depends on the model and the corresponding assumptions.

Many source modelling techniques have been developed to solve the inverse problem (Hamalainen et al., 1993; Hansen et al., 2010; Hari and Puce, 2017). Equivalent current dipole (ECD) modelling represents currents giving rise to the MEG signal as a collection of local current dipole models, the number of which is assumed to be known. The strength and location of the current dipole models are adjusted step by step until the root of the squared sum of the difference between modelled magnetic field and the measured magnetic field is minimized. ECD modelling is a popular method for localizing the sources of transient ERFs at early post-stimulus latencies when only few brain areas are expected to be activated. However, distributed source models describe better cortical activity that spreads to multiple brain regions over extended time periods or noisy data sets since they rely on minimal modelling assumptions. Distributed source models do not assume finite sources but, rather, represent brain activity as arising from a continuous current distribution (Hämäläinen and Ilmoniemi, 1994). One of the most widely used methods is minimum norm estimate (MNE) which relies on minimizing the L2-norm. MNE tend to preferentially place source activity in the superficial cortical areas, as being the closest to the sensor array, they provide the smallest norm. The resolution also drops substantially with sources in deeper in the sulci. This effect can be reduced by using MNE in combination with dynamic statistical parametric maps (dSPMs) calculated by noise-normalizing the MNE estimates. dSPM estimates can be used to determine the brain areas where the MNE amplitudes are above the noise level, and MNE estimates without noise-normalization to determine the actual currents in these areas.

In Study I, the locations of N1m and P2m sources were estimated by fitting ECDs separately in each hemisphere using a subset of 12 gradiometer pairs over each temporal area. MNE and dSPM estimates were generated separately for each subject, type of stimulus (i.e. vowels, words and sentences), and for each condition (i.e. the presentation of the intact stimuli and two
presentations of the distorted stimuli). The noise covariance matrices for MNE and dSPM estimates were calculated from the 100-ms prestimulus baseline of the data. dSPM estimates were averaged across a time window of 40 ms centred at the peaks of the N1m and P2m responses. The peak latencies were determined as the time instances when the dSPM estimates reached their maximum values within time intervals of 110–170 ms for N1m and 190–310 ms for P2m. To determine the areas where MNE estimates exceeded noise level, dSPM estimates were grand-averaged across subjects and conditions for N1m and P2m responses. For the sustained fields, mean values of the grand-averaged dSPM estimates were calculated for the time interval of 400–560 ms. The dSPM activation areas were divided into the regions of interests (ROIs) following the Desikan-Killiany-Tourville atlas. The ROIs were approximations of the brain areas since individual MRI data was not available. To obtain a time course of current strength for each ROI, the MNEs were averaged over source locations within ROI. The mean currents of N1m and P2m responses were estimated by averaging the time course of the current in each ROI over a 40-ms time window centred at the peaks of N1m and P2m responses. The peak latencies were determined as the instances when the MNE estimates reached their maxima within the intervals of 110–170 ms for the N1m and 190–310 ms for the P2m. The mean currents of the sustained fields were calculated by averaging the mean time course in each ROI over the time interval of 400–560 ms.

2.4 Magnetic resonance imaging

Magnetic resonance imaging (MRI) produces detailed 3-dimensional images of the organs and tissues within the body using a strong magnetic field and radio waves (Huettel et al., 2004). MRI uses the magnetization properties of atomic nuclei. To contribute to the MRI signal, the nuclei need to have a spin, that is, magnetic moment and magnetic torque. If the atom which has spin is placed in a magnetic field, it aligns with the magnetic field and starts to precess or wobble around the axis longitudinal to the magnetic field. Hydrogen atoms are the most commonly imaged nuclei in MRI because they have the spin and are the most abundant atoms in the human body due to their presence in the water molecule.

During the MRI measurement, the part of the subject being examined is placed in a strong magnetic field (Huettel et al., 2004). Consequently, the hydrogen atoms in the tissue take either a high-energy state in which they start to precess around the spin axis that is parallel to the magnetic field or a low-energy state where they precess around the spin axis antiparallel to the magnetic field. The stronger the magnetic field is, the more spins take the low-energy state, and the stronger net magnetization is generated along the magnetic field. During the measurement, energy is applied to the hydrogen nuclei by radio-frequency pulses adjusted to oscillate at their resonant frequency. As a result, some of the nuclei change from the low-energy state to the high-energy state. When the number of spins in the high-energy
state equals to the number of spins in the low-energy state, the longitudinal net magnetization disappears and there is only transverse magnetization (perpendicular to the main magnetic field). After the RF pulse is switched off, the spins start to change back from the higher to lower energy state emitting the same amount of energy that corresponds to the difference of energy between the two energy states. As a result, the transverse magnetization starts to disappear (transversal relaxation) and the longitudinal magnetization recovers back to its original magnitude (longitudinal relaxation). Emitted energy induces current in MRI coils generating MRI signal. The time constant which determines the time taken for the longitudinal component of the net magnetization to recover is referred to as $T_1$, and the time constant which determines the rate at which the transverse component of the net magnetization decays is referred to as $T_2$. Different tissues relax at different rates making it possible to emphasize particular tissues in the images. The time instant when the image is acquired during the relaxation process determines to which extent $T_1$ and $T_2$ will contribute to the amplitude of the MR signal, and thus, the intensity of the image. By selecting appropriate imaging parameters, different tissues will correspond to different intensities and become distinguishable in the MRI images. The most important imaging parameters that determine the contrast between tissues are repetition time (TR) of the RF-pulses that defines the time between two successive excitations and echo time (TE) that is the time between excitation RF-pulse and the peak of the MRI-signal induced in the coil. The most commonly used MRI sequences are $T_1$-weighted and $T_2$-weighted images. In $T_1$-weighted images, the intensity of the image depends on the $T_1$ value of the tissues. Conversely, in $T_2$-weighted images the intensity of the image depends on the $T_2$ value of the tissues. For example, the fat tissue has a high signal intensity on $T_1$-weighted images and a relatively low signal intensity on $T_2$-weighted images.

### 2.5 Functional magnetic resonance imaging

While structural MRI produces static images of the brain, fMRI measures a set of dynamic images that show how the brain activity changes over time (Huettel et al., 2004). However, fMRI uses pulse sequences that allow determining the changes in the amount of oxygenated blood flow. Oxygenated haemoglobin is diamagnetic, that is, it does not have unpaired electrons and a magnetic moment. In contrast, deoxygenated haemoglobin is paramagnetic, that is, it has unpaired electrons and a magnetic moment. Deoxygenated haemoglobin distorts the magnetic field shortening the time of transversal relaxation which decreases the fMRI signal. When the neural activity within a brain area increases, it consumes more oxygen and the flow of the oxygenated blood to the brain area increases resulting also in a stronger fMRI signal. Thus, fMRI detects changes in neuronal activity indirectly by measuring the changes of the oxygenated blood flow.

The signal recorded with fMRI is called the blood-oxygen-level dependent (BOLD) signal (Huettel et al., 2004). The changes in the BOLD
signal are relatively slow compared to those in the underlying neuronal activity. The BOLD signal elicited by a short stimulus is delayed from the onset of neuronal activity approximately 2 sec and peaks at 5–6 sec following the onset of neuronal activity. After peaking, the BOLD signal starts to decrease typically diving slightly under the baseline at 10 sec and returning to the baseline after approximately 30 sec. If the stimulus is extended in time, the peak is followed by a plateau of sustained signal maintained with a slightly lower amplitude than the peak. Some studies have also reported an initial dip during 1–2 sec that reflects a momentary increase of deoxygenated haemoglobin in the voxel. However, the relation between the BOLD-signal and neuronal activity is complex and the shape of the BOLD-signal varies depending on the brain area, on the properties of the stimulus, and on many confounding physiological factors (Logothetis, 2008).

2.5.1 Increasing sampling-rate of fMRI

The most commonly used fMRI sequence is the echo planar imaging (EPI) sequence that measures one slice of fMRI data after each excitation pulse (Huettel et al., 2004). Gradient fields are then switched to sample the frequency and phase encoding directions of the slice. The gradient switching limits the TR of the EPI sequence to 1–2 sec. Because the spectral power of the BOLD signal is mainly distributed below 0.1 Hz, it has been generally thought that increasing the sampling rate from the conventionally used sampling rate of 0.5 Hz (i.e. TR of 2 sec) would not provide any additional information. This conclusion stems from the Nyquist theorem which states that the sampling rate of twice the highest frequency of the signal of interest is enough to accurately reconstruct the signal of interest. However, accumulating evidence suggests that, in practice, higher sampling provides several benefits over the TRs of 1–2.5 sec, such as a more accurate removal of physiological artefacts and a more accurate determination of causal relationships between active brain areas when regional vascular and biomechanical response variability is appropriately accounted for (LeVan et al., 2018; Lin et al., 2012). Recent studies have used fast fMRI techniques to demonstrate temporal coherence between resting-state activity of brain regions (i.e. functional connectivity) above 0.2 Hz (Boubela et al., 2013; Lee et al., 2013; Lin et al., 2015; Smith-Collins et al., 2015; Trapp et al., 2018). Further, hemodynamics measured with fMRI has been found to contain oscillations related to neural activity up to 0.75 Hz which suggests that sampling the hemodynamic response at frequencies higher than the conventional sampling rate of 0.5 Hz could reveal new information about high frequency brain dynamics (Lewis et al., 2016).

In recent years, several methods have been developed to increase the sampling rate of fMRI (for reviews, see LeVan et al., 2018; Lin et al., 2012). One of these methods is dynamic inverse imaging (InI) where information in one, two, or all imaging directions is derived from the sensitivity distribution of the coil array (Lin et al., 2006). If all spatial information is derived from the coil sensitivities, the image reconstruction requires solving a 3D inverse problem (3D InI) similar to the one of source localization of MEG signals. In
contrast to MEG where the determination of source locations always requires solving a 3D inverse problem, InI can substantially decrease uncertainty by using slice selection and frequency encoding gradients to acquire the spatial information in one (2D InI) or two imaging directions (1D InI). Recently, a simultaneous multislice inverse imaging (SMSInI) method was developed that, compared to traditional InI, yields a higher spatial resolution, lower signal leakage, and a higher time-domain signal-to-noise ratio when the regularization parameter is optimized in the image reconstruction (Hsu et al., 2017).

2.5.2 Acquisition of MRI and fMRI data

All MRI and fMRI data for this thesis was acquired using a 3-tesla MRI scanner (MEGNETOM Skyra, Siemens Healthcare, Erlangen, Germany) and a 32-channel receiving head coil at the Advanced Magnetic Imaging (AMI) centre of Aalto University. During fMRI measurements, the heart-rate and respiration signals were acquired with a BIOPAC MP150 Data Acquisition system (BIOPAC System, Inc.). The stimuli were presented using Presentation software (Neurobehavioral Systems, Albany, NY, USA). MRI data was measured using a high-resolution T1-weighted Magnetization Prepared Rapid Gradient Echo (MPRAGE) pulse sequence (repetition time (TR) = 2530 ms, echo time (TE) = 3.3 ms, flip angle = 7, 256 x 256 matrix, 176 sagittal slices, 1-mm resolution).

In Study II, fMRI data was acquired using a T2*-weighted echoplanar imaging (EPI) sequence (TR = 2000 ms, TE = 30 ms, flip angle = 76°, field of view (FOV) = 220 x 220 x 220 mm³, slice thickness = 3.4 mm, 37 oblique slices acquired in ascending order covering the whole brain in plane resolution). Each run started with a measurement of six dummy scans to establish steady-state magnetization before acquiring task-related BOLD-responses.

In study III, we pioneered the use of SMSInI sequence in collecting fMRI data during listening to a 71-min narrative (Hsu et al., 2017; TR=100 ms, TE=27.5 ms, flip angle=30°, FOV=210 x 210 x 210 mm³, slice thickness = 7 mm, in-plane resolution = 5 mm x 5 mm). InI-encoding was used to recover spatial information in superior-inferior direction, frequency encoding in the anterior-posterior direction and phase encoding in the left-right direction. A 6-sec reference scan preceded each fMRI run (Hsu et al., 2017). The reference scan was measured with the same parameters as the fMRI scans, but partition-encoding steps were added after slice group excitation in InI-encoding direction. The reference scan includes information about coil sensitivities and is used as a forward operator for spatial encoding in the image reconstruction that requires solving an inverse problem along the InI-encoding direction. The inhomogeneities of the magnetic field were minimized by shimming before each reference scan.
2.5.3 Preprocessing of MRI and fMRI data

In study II, T1 and fMRI data were preprocessed with standard preprocessing pipelines. For T1 images, the preprocessing included conversion of images into Neuroimaging Informatics Technology Initiative Technology Initiative (NIfTI-1, http://nifti.nimh.nih.gov/dwg, RRID:SCR_007117) format, restoring bias field (FAST; Zhang et al., 2001) and removing nonbrain structures (Brain Extraction Tool (Smith, 2002). The preprocessing of fMRI data consisted of the conversion of images into NIfTI format, slice timing correction, motion correction (MCFLIRT; Jenkinson et al., 2002), removal of physiological artifacts by DRIFTER (Särkkä et al., 2012), removal of 6 TRs from the beginning of each measurement (used to stabilize magnetization), registration to the 2-mm Montreal Neurological Institute (MNI152) standard space (Greve and Fischl, 2010; FLIRT, http://www.nitrc.org/projects/dwiregistration/, RRID:SCR_009461), and spatial smoothing using a Gaussian kernel with 10 mm full width at half maximum. The registration to the MNI152 space was done in three steps. First, a matrix was estimated that described the transformation from structural (T1) space to MNI152 space. Second, a matrix describing the transformation from functional (fMRI) space to structural space was estimated. Third, the functional images were transformed into the MNI152 space through these two consecutive transformations.

In study III, the T1 images were reconstructed using FreeSurfer’s automatic reconstruction tool (recon -all; http://surfer.nmr.mgh.harvard.edu/) and fMRI data using the regularized sensitivity encoding (SENSE; Lin et al., 2005, 2004) algorithm with a regularization parameter of 0.005. Thereafter, the fMRI data was registered to the 3-mm MNI152 space. The registration was performed by FreeSurfer and FSL tools (Jenkinson et al., 2002; Jenkinson and Smith, 2001) using the same three registration step as in Study II. The following preprocessing steps were performed after the registration: 1) removing a period of 12.3-sec of fMRI data measured before the presentation of the narrative in each run, 2) detrending the scanner drift using a Savizky-Golay filter (order: 3, frame length: 240 sec) reducing physiological and movement artifacts with MaxCorr method (Pamilo et al., 2015), 4) temporal smoothing of the data between 0.08–4 Hz by a zero-phase band-pass filter and 5) spatial smoothing of the data by a 6-mm full-width-half-maximum Gaussian kernel. The MaxCorr method was implemented such that from the data of each subject, 10 components were removed which correlated maximally within that subject’s white matter but minimally with the other subject’s white matter.

2.5.4 General linear model analysis

General linear model (GLM) is a standard method to analyse task-related BOLD activations (Huettel et al., 2004). In GLM analysis, the measured data is modelled as a weighted sum of one or more explanatory variables and residual noise as follows:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$

The regressors $x$ correspond to the assumed time courses explaining the fMRI responses $y$, $\beta_0$ represents baseline signal intensity that is constant across the experiment, and $\varepsilon$ is residual noise in the data or error in the measurement. In practice, the regressors are created by convolving the time course of the stimulus of interest with a standard model of the hemodynamic response function (HRF). The matrix containing the regressors in its columns is referred to as a design matrix or model. The aim of GLM analysis is to find the weights $\beta_i$ that minimize the error term, that is, the difference between the modelled data (design matrix) and the measured fMRI data. The standard cost function minimized in GLM analysis is the least-squares error, that is, the squared sum of residual errors. As a result of the minimization, each regressor gets a weight that describes the contribution of that regressor in explaining the voxel time series. Large positive beta values typically indicate that the experimental condition has activated the brain area within the voxel compared to the baseline whereas large negative values have been assumed to indicate deactivation of the brain area during the experimental condition relative to baseline.

GLM was used to analyse the fMRI data in Study II. Three regressors were created, one predicting the BOLD responses to the first presentation to the distorted sentences, one to the intact sentences and one to the second presentation of the distorted sentences. For each subject, the contrasts were calculated between 1) the second distorted versus the first distorted, 2) intact versus second distorted and 3) intact versus first distorted presentations. The group-level results were calculated using the one sample $t$-test and the $t$-maps were thresholded using nonparametric permutation test (Nichols and Holmes, 2001; SnPM13, http://www.warwick.ac.uk/snpm, 10 000 random permutations, cluster-level results corrected for multiple comparisons using FEW at $p < 0.05$).

### 2.5.5 Inter-subject correlation analysis

In inter-subject correlation (ISC) analysis, voxelwise Pearson’s correlations are calculated between the BOLD time courses of each pair of the subjects (Fig. 4.) (Hasson et al., 2004; Kauppi et al., 2014). Thereafter, a correlation map where each voxel has one correlation value is calculated by averaging the pairwise correlations in each voxel. ISC analysis is especially suitable for studying BOLD responses to complex naturalistic stimuli since, in contrast to GLM analysis, it does not require a priori model of the BOLD responses to the experimental conditions.

In Study III, ISC analysis was used to investigate fMRI data measured while the subjects were listening to an auditory narrative (ISC-toolbox, https://www.nitrc.org/projects/isc-toolbox/; Kauppi et al., 2014). For each voxel and each of the 10 runs, $1128 (n \times (n - 1)/2$, where $n = 48$) ISC values were calculated across the BOLD time-courses of every subject pair. ISC values across the whole narrative were obtained by averaging the runwise ISC
values that were transformed into z-values using Fisher's transformation and weighted by the length of the runs. The ISC values were tested for statistical significance using a nonparametric voxel-wise resampling test that generated the null-distribution from ISC values calculated after circularly shifting each subject’s timeseries by a random amount such that the timeseries were not aligned in time across subjects. This allowed us to account for temporal autocorrelations of fMRI data. The null-distribution was approximated across 0.5 million realizations randomized across voxels and shifting points. The resulting p-values were adjusted for multiple comparisons using false discovery rate -correction. To see whether the family cultural background affects how similarly the subjects have processed the story, a nonparametric two-sample T-test with 5000 permutations was calculated between the z-transformed ISC values obtained for the subjects with Finnish and Russian backgrounds (Glerean et al., 2016). The resulting statistical maps were thresholded at $p<0.01$ and, thereafter, adjusted for cluster extent at 5%.

Figure 4. Schematic description of the inter-subject correlation analysis. First, the pairwise correlations of voxel time series are calculated separately for each voxel between every possible subject pair. Second, an average correlation value is calculated over the pairwise correlations for each voxel resulting an ISC map.

2.6 Measuring narrative interpretation using word association test

To see whether the cultural family background was reflected in the narrative interpretation in Study III, each of the subjects participated in an online association experiment where the subject was presented with the narrative in 101 sections. After each section, the subject typed a few words that described what had been on her or his mind at that point during the neuroimaging. The word2vec skip-gram model was applied to study the semantic relatedness between the word lists produced by the subjects (Gensim Python Library, https://pypi.org/project/gensim/). Word2vec is a two-layer neural net used to produce vector embeddings from a text corpus (Mikolov et al., 2013). The word2vec skip-gram method trains the neural net by using a given word of the corpus to predict its surrounding words such that the nearby words are weighted more than distant words. The weights of the hidden layer of the neural net are used as vectorized representations of the words. Finally, this training procedure results in a vector space where each word of the corpus is represented with a vector, and the distance between the word vectors is proportional their co-occurrences in the corpus (i.e. the vectors that correspond to contextually similar words are positioned close to each other).
The model was trained with the Finnish Internet Parsebank corpus (Haverinen et al., 2014) using the following parameters. Obvious spelling errors were corrected from each of the 101 word lists for each subject. Using the word2vec model, each word of the list was mapped to a corresponding vector representation that described the semantic content of the word. Thereafter, a vector sum was calculated over the vector representations of the words within each of the list. This resulted in 101 semantic vectors (one vector per narrative section) for each subject. The semantic similarities between the words reported by the subjects were estimated by calculating pairwise cosine distances between the vectorized word lists in each narrative section. The analysis was implemented using Gensim Python Library (https://pypi.org/project/gensim/).

To test whether the family cultural background was reflected in the semantic content of the word lists, a nonparametric t-test with 50,000 permutations was performed, comparing the Finnish- and Russian-background subjects in terms of the cosine similarities. The t-test was calculated separately in each of the 101 sections and across the whole narrative using an average cosine similarity over the sections after weighting the similarities in each section by the length of that section.

2.7 Implicit association test

In Study III, the implicit association test (IAT) was used to measure whether the subjects have an implicit preference for Finland over Russia or for Russia over Finland. The IAT measures the strength of associations between concepts and attributes that the subject may be unwilling to express or about which the subject is unconscious (Greenwald et al., 1998a, 1998b). In IAT, the subject categorizes two concepts (e.g. flower vs. insect names) with an attribute (e.g. pleasant vs. unpleasant) as fast as possible. The assumption is that the reaction times are faster when the strongly associated categories are assigned to the same response key (e.g. flower + pleasant) than when less associated categories are assigned to the same response key (e.g. insect + pleasant). IAT has been used to study bias for example in attitudes towards racial groups (McConnell and Leibold, 2001), religions (Rowatt, Wade and Franklin, Lewis, 2004), gender (White and White, 2006) and sexuality (Anselmi et al., 2013) as well as self-esteem (Greenwald and Farnham, 2000).

The IAT was implemented following the standard IAT procedure that consists of 7 consecutive tasks. In each task, the subject is instructed to categorize, as fast as possible, the words and/or pictures presented in the center of the computer screen either into the category presented in the left upper corner of the screen using the left hand or into the category presented in the right upper corner of the screen using the right hand. In Study III, the tasks were to

1) categorize images and words related to Finnish and Russian cultures into the categories “FINLAND” using the left-hand and “RUSSIA” using the right-hand key,
2) categorize positive and negative adjectives into the categories “GOOD” using the left-hand and “BAD” using the right-hand key,
3-4) categorize images and words related to Finnish and Russian cultures into the combined categories “FINLAND/GOOD” using the left-hand and “RUSSIA/BAD” using the right-hand key,
5) categorize the images related to Finnish and Russian cultures into the categories “RUSSIA” using the left-hand and “FINLAND” using the right-hand key (i.e. a repeat of the first task with the exception that the position of the categories on the screen and was reversed),
6-7) categorize images and words related to Finnish and Russian cultures into the combined categories “FINLAND/BAD” using the left-hand and “RUSSIA/GOOD” using the right-hand key (i.e. opposite pairings as in 3th and 4th tasks).

If the reaction times were shorter when the combined categories were “FINLAND/GOOD” and “RUSSIA/BAD”, the subject was interpreted to have more positive associations with Finland than with Russia. If the reaction times were shorter when the combined categories were “FINLAND/BAD” and “RUSSIA/GOOD”, the subject was interpreted to have more positive associations with Russia than with Finland.
3. Summaries of the studies

3.1 Study I: Previous exposure to intact speech increases intelligibility of its digitally degraded counterpart as a function of stimulus complexity

3.1.1 Aim of the study

This study investigated how the processing and intelligibility of acoustically distorted speech is affected by prior exposure to the speech signal in its undistorted form. Our study used a new experimental paradigm where the same set of speech stimuli was presented to the subject in a distorted, undistorted, and again in a distorted form. This appears to open up an interesting possibility to study speech intelligibility since any change between the responses to the distorted stimuli before and after the presentation of their undistorted counterparts cannot be elicited by acoustic variability but, rather, by cognitive processes involved with the successful match between acoustic information and memory representations.

3.1.2 Experimental setup

The study consisted of behavioral and MEG experiments. The behavioral experiment measured the subject’s ability to recognize the auditory stimuli, and MEG experiment was conducted to study how the comprehension of distorted speech was reflected in cortical activity.

In the behavioral experiment, the subject was presented with three D-I-D experiments, one with vowels, the second one with words, and the third one with sentences as stimuli (Fig. 5, for the description of the stimuli, see 2.2). In each D-I-D experiment, the stimulus set was presented first in a distorted form, then as intact speech, and finally again in the distorted form. The distorted stimuli were presented as a single block, as were the intact stimuli. Thus, an average delay between a specific stimulus being presented in its intact form and distorted form was 5 sec for vowels, 2 min for words, and 11 min for sentences.

The vowel set contained 12 repetitions of eight vowels (/a/, /e/, /i/, /o/, /u/, /y/, /ä/, /ö/). At each vowel presentation, eight alternative vowels were shown to the subject on the computer screen and the subject was instructed to indicate by a mouse click which of the eight alternative vowels she/he had heard or whether she/he could not recognize the vowel. Correctly identified vowels were classified as intelligible and incorrectly identified as unintelligible.
The word set and sentence set both included 192 stimuli which were divided into four subsets of 48 words or sentences. The 16 subjects who participated in the study were also divided into four groups and each group of the subjects was presented with one subset of words and one subset of sentences. Following the presentation of each word or sentence, the subject was instructed to type using a keypad what she or he had heard. Correctly identified words were classified as intelligible and incorrectly identified as unintelligible. Intelligibility scores for sentences were calculated by scoring the stems and suffixes of inflated words separately.

Similarly, as in the behavioral experiment, in the MEG experiment the subject was presented with three D-I-D experiments one with vowels, the other with words, and the third with sentences as stimuli (Fig. 5). The presentation order of the D-I-D experiments was counterbalanced across the subjects. The vowel set consisted of four vowels (/a/, /e/, /i/, /u/) that were repeated 120 times in a random order. The sentence set included 160 sentences selected from the 192 sentences presented in the behavioral experiment. The starting words of the sentences comprised the word set (i.e. 160 words). The stimuli were presented with offset-to-onset interstimulus interval (ISI) of 1 sec. During the experiment, the task of the subject was to watch a film without its soundtrack and to ignore the speech stimuli.

![Figure 5](image-url)  
Figure 5. Experimental design of Study I. The study consisted of behavioral and MEG experiments, each of which consisted of three D-I-D experiments, one with vowels, the second with words, and the third with sentences as stimuli. N= number of stimuli.

### 3.1.3 Results

The intact stimuli were 99% intelligible (intelligibility defined as a proportion of correct identifications). A single exposure to the intact speech stimuli improved the overall intelligibility of their acoustically distorted counterparts from 26% at their first presentation to 45% at their second presentation (Fig. 6a). Importantly, the intelligibility increased as a function of the complexity of the message: vowels remained unintelligible whereas the intelligibility of
words increased from 19% to 45%, and for sentences this increase was from 31% to 65%. In auditory cortex and surrounding areas, the distorted vowels elicited a stronger mean current at their second than at their first presentation at the N1m latencies of 130–160 ms (Fig. 6b). The corresponding effect for sentences was also found in the insula. The responses depended on stimulus complexity, with mean currents being a decreasing function of complexity at N1m latencies and an increasing function at P2m latencies of 200–270 ms. We found no differences between the mean currents of the sustained fields elicited by the first and second presentations of the distorted stimuli.

![Figure 6](image)

**Figure 6.** a) The mean intelligibility scores for the speech stimuli when they were presented in the intact and distorted forms. The intelligibility scores increased as a function of the complexity of the message. b) Brain areas where the N1m amplitude was higher at the second compared to the first presentation of the distorted vowels. Error bars indicate standard error of mean. Names for the ROIs: 1 = precentral sulcus, 2 = supramarginal gyrus, 3 = insula, 4 = transverse temporal gyrus and superior temporal gyrus, 5 = superior temporal sulcus.

### 3.1.4 Conclusions

The results of this study demonstrate that already a single prior exposure to an intact speech stimulus can remarkably increase the intelligibility of that stimulus if it is later presented in a distorted, noisy form, and that the complexity of the message conveyed by the stimulus enhances this effect. This prior exposure modulated the processing of distorted speech in the auditory cortex and surrounding areas already at the latencies of 130–160 ms from stimulus onset. This suggests that the memory representations of prior exposure to an intact speech stimulus affect speech processing already at the early stages. Interestingly, these effects occur even when the subject does not pay attention to the stimuli.
3.2 Study II: Predictive processing increases intelligibility of acoustically distorted speech: Behavioural and neural correlates

3.2.1 Aim of the study

Study II examined further the neural mechanisms underlying the instant disambiguation of degraded spoken sentences in situations where the subject has previously been exposed to the intact versions of the same sentences. Study I was able to identify the earliest signs after stimulus onset that prior exposure to the intact sentence was affecting the cortical processing of its distorted counterpart in the auditory cortex. Study II aimed to uncover whether also brain areas outside of the auditory cortex, such as frontal and subcortical areas, contribute to the rapid increase in intelligibility of the distorted sentences associated with prior access to the intact versions of these sentences.

3.2.2 Experimental setup

In the behavioral experiment, the subject was presented with 15 D-I-D experiments. In each D-I-D experiment, seven distorted sentences were first presented in a distorted form. Thereafter, five of the previous seven sentences were presented in their intact form. Finally, the seven distorted sentences were repeated. Two of the sentences were only presented in the distorted form to test whether repetition itself of the distorted sentence increases its intelligibility. The sentence set was always presented in the same order, notwithstanding sentence omissions in the presentation of the sentences in the intact form. After the presentation of each sentence, the subject typed using a keypad what she / he had heard.

The fMRI experiment was divided into two 19-min functional runs, in each of which 9 D-I-D experiments with six sentences as stimuli were presented to the subject (Fig. 7). Each three presentations (i.e. 1st distorted, intact, and 2nd distorted) of the stimulus set were followed by a period of 16 sec without auditory stimulation. After the second presentation of the sentence set in the distorted form, a question appeared on the computer screen instructing the subject to response by pressing a button (yes/no) whether the distorted sentences were easier to understand when they followed the presentation of their intact counterparts. The subjects were instructed to listen to the sentences attentively. A visual fixation cross was presented to the subject during the measurement. Separate groups of subjects participated in the behavioral and fMRI experiments to prevent carry-over effects.
Figure 7. Setup of the fMRI experiment in Study II. 18 D-I-D experiments were presented to the subject during fMRI measurement. In each D-I-D experiment, a block of six sentences was presented in the distorted, intact and, again, in the distorted form (unique set of sentences in each D-I-D experiment). The sentence blocks were separated by periods of 16 sec without auditory stimulation. Within a time-window of 6 sec after the ending of the D-I-D experiment, the subject indicated with a button press (yes/no) whether the distorted sentences were easier to understand after the presentation of their intact counterparts.

3.2.3 Results

The mean intelligibility score (defined as a proportion of correct identifications) of the distorted sentences at first presentation, averaged across subjects and stimulus sets, was 41% (Fig. 8). However, this score increased to the near-perfect level of 95% when the distorted sentences were presented after their intact counterparts. This dramatic increase in intelligibility was mainly caused by the exposure of the intact sentences since the intelligibility increased from 45% to a mere 53% if the distorted sentences were missing their counterpart in the set of intact sentences. The intelligibility of the first presentation of the distorted sentences increased with the ordinal position of the stimulus sets.

Figure 8. The mean intelligibility scores of the sentences at their presentations in the distorted, intact and distorted forms across the subjects for each of the 15 stimulus sets. Shaded error bars describe the standard error of the mean. The sentences presented only in the distorted form were excluded from the mean scores shown in this figure. In each stimulus set, the low intelligibility scores of the distorted sentences increased near to the perfect level when the distorted sentences were presented after their intact intelligible counterparts. The intelligibility of the distorted sentences increased at their first presentation with the ordinal position of the stimulus set.

We compared the BOLD responses elicited by the second presentation of the distorted sentences with those elicited by the first presentation. The second
presentation resulted in enhanced BOLD-responses in the bilateral lingual gyrus (LG) and anterior cingular cortex (CC), the right frontal operculum (FOC), and in an area extending from the left precentral gyrus (PCG) to the left middle frontal gyrus (MFG, Fig. 9). Two spherical areas also showed enhanced BOLD-responses bilaterally in the frontal pole (FP). Decreased responses were found in the right supramarginal gyrus (SMG), bilaterally in the area covering the posterior insula (pIns), in Heschl’s gyrus (HG), in planum temporale (PT), as well as bilaterally in the parahippocampal cortex (PHG). In the left hemisphere, the area with a decreased BOLD-response extended from the PHG to the occipitotemporal fusiform cortex (OTFC).

The intact sentences resulted in stronger BOLD-responses compared to the first presentation of the distorted sentences in the bilateral LG and MTG as well as the left anterior PHG, and weaker in the bilateral anterior and posterior CC as well as in the bilateral area covering HG and planum temporale (PT), in the left cerebellum and in the right FP and STG. The intact sentences elicited stronger BOLD-responses compared to the second presentation of the distorted sentences in the bilateral MTG and weaker in the bilateral posterior PCun, CC, the left CBC, orbitofrontal cortex (OFC) and FP as well as the right AG and FP.
Summaries of the studies

Figure 9. Group-level activation maps for the follow contrasts: 1) the second (high intelligibility) versus the first (low intelligibility) presentation of the distorted sentences, 2) the presentation of the intact versus the first presentation of the distorted sentences and 3) the presentation of the intact versus the second presentation of the distorted sentences. The significance of the results was tested with a cluster level nonparametric multiple comparison correction procedure using a cluster-forming threshold of \( p < 0.0001 \), 10 000 permutations and a cluster-level family-wise error correction at \( p < 0.05 \). Abbreviations: LG = lingual gyrus, CC = cingulate cortex, PCG = precentral gyrus, MFG = middle frontal gyrus, SMG = supramarginal gyrus, FOC = frontal operculum, PHG = parahippocampal gyrus, OTFC = occipitotemporal fusiform cortex, FP = frontal pole, STG = superior temporal gyrus, Ins = insula, HG = Heschl’s gyrus, PT = planum temporale, IFG = inferior frontal gyrus, MTG = middle temporal gyrus, PCun = precuneus, CBC = cerebellar cortex, OFC = orbitofrontal cortex, AG = angular gyrus, PT = planum temporale, a = anterior, p = posterior.

3.2.4 Conclusions

The results of this study further confirm the results of Study I demonstrating that a single prior presentation of an intact sentence can render its initially unintelligible distorted counterpart intelligible. The contrast between the BOLD responses elicited by the same distorted sentences when perceived as intelligible versus unintelligible revealed activations and deactivations in several extralinguistic brain areas. Previous studies have identified these areas as being associated with memory, attention, and performance monitoring during the comprehension of acoustically distorted speech (for a review, see Peelle, 2018). The BOLD modulations in the bilateral FP and in the bilateral auditory cortex might reflect activation of long-term memory since those areas were not associated with increased speech intelligibility when the disambiguating stimulus was presented immediately after the distorted
stimulus in previous studies (Clos et al., 2014; Hervais-Adelman et al., 2012; Zekveld et al., 2012). Based on previous studies on the memory retrieval mode (Lepage et al., 2000), response enhancement in the frontal and cingulate cortices found in this study could be an indication of distorted sentences being treated as retrieval cues that are compared to the information stored in memory. The deactivations occurring in auditory cortex might reflect a predictive coding mechanism (Friston, 2005; Huang and Rao, 2011; Murray et al., 2002) where responses generated in sensory areas reflect a prediction error that is the difference between the sensory information and the information stored in memory.

3.3 Study III: Processing of a spoken narrative in the human brain is shaped by family cultural background

3.3.1 Aims of the study

In studies I and II, the perception of the degraded speech stimuli was biased by prior exposure to the intact counterparts of the stimuli. Does prior experience more generally color speech perception so that this becomes specific to individual subjects or groups of subjects? People living in a shared cultural environment tend to share similar experiences, knowledge, and social habits, and this has been shown to be reflected in their perception (Ames and Fiske, 2010; Chiao et al., 2010; Gutches and Indeck, 2009). Study III, therefore, investigated whether a shared family cultural background can increase inter-individual similarity in how natural connected speech is interpreted and processed in the brain.

3.3.2 Experimental setup

Two groups of subjects, one with Finnish and the other with Russian family backgrounds (see 2.1), took part in the cultural background and personality questionnaires, an fMRI recording, simultaneous MEG and EEG recordings, as well as an IAT and a behavioral association test. In this thesis IAT, behavioral association test, T1-weighted images and the fMRI data collected during the presentation of the 71-min narrative (i.e. 10 runs, for the description of the stimulus, see 2.2) were analyzed. In the behavioral association experiment, the same narrative that was presented during fMRI was replayed in 101 segments. After each segment, the subject was given 20–30 sec to describe what had been on their minds while they had heard the segment during the neuroimaging. This description was attained by asking the subject to type few words.

3.3.3 Results

To estimate the between-subject similarities in the narrative interpretation, word2vec was first used to transform the word associations produced by the subjects into vectors that describe the semantic content of the words (see 2.6).
One vector was produced for each subject in each of the 101 narrative sections. Thereafter, cosine similarities between the vectors were calculated in each section. The cosine similarity described the semantic similarity of the words. Finally, the pairwise cosine similarities were compared between the two groups of subjects in each section. In 12 out of 101 sections, the cosine similarity between the vector representations of the word lists was larger among the subjects with a Finnish family background than among the subjects with a Russian background (Fig. 10a). In 44 sections, this similarity was higher for the subjects with a Russian background. In 45 sections, no between-group differences were found.

Compared to the subjects with a Russian background, there were a number of areas where the subjects with a Finnish background showed stronger ISC of the BOLD responses obtained during narrative listening. These areas were in the left posterior-middle Ins and anterior STG, in the right posterior STG and MTG, as well as in the bilateral HG, lateral occipital cortex (LOC, more extended in the right hemisphere), inferior occipitotemporal cortex (OTC), CBC, PCun, and superior parietal lobule (SPL, see Fig. 10b). A stronger ISC for the subjects with Russian family background was found in left-hemisphere areas extending from the HG to STG and MTG, as well as bilaterally in posterior-inferior parts of the PCun and in anterior parts of the cuneus.

In the IAT (see 2.7), subjects with a Finnish background had shorter response times when the task was to categorize words and pictures presented on the computer screen into the paired categories “FINLAND/GOOD” and “RUSSIA/BAD” compared to the task where the paired categories were “FINLAND/BAD” and “RUSSIA/GOOD. This suggests that these subjects had a more positive association with Finland than with Russia (Fig. 10c). For the subjects with a Russian family background, no differences were found in the response times between categorization tasks.
Figure 10: a) In the association experiment, each of the subjects listened to the narrative in 101 sections and, after each section, typed few words that were on her/his mind at that point during neuroimaging. In 12/101 sections, the subjects with Finnish family background reported more similar words than the subjects with the Russian family background. In 44/101 sections, the semantic similarity was higher between the words reported by the subjects with the Russian family background. No differences between groups were found in 45/101 sections. b) Brain areas where ISC was different between the subject groups. The statistical significance tested with 5000 elementwise permutations and corrected at the cluster level with $p<0.05$. c) Subjects’ preferences for Finland and Russia in IAT. Bar graphs indicate the number of subjects at each bias level. Abbreviations: Heschl’s gyrus = HG, superior temporal gyrus = STG, middle temporal gyrus = MTG, lateral occipital cortex = LOC, precuneus = PCun, SPL = superior parietal lobule, occipitotemporal cortex = OTC, LG = lingual gyrus, OP = occipital pole

3.3.4 Conclusions

The findings of this study provide both behavioral and neuroimaging evidence suggesting that the family background influences the way natural connected speech is processed. All subjects in our study were Finnish residents fluent in Finnish, most of them born and everyone raised in Finland. Thus, the difference between their cultural background was rather subtle via either one or both parents of half of the subjects being Russian. Our results suggest that, indeed, even such relatively subtle difference in the cultural backgrounds can result in higher similarity within one group compared to the other in how natural speech is interpreted and processed at multiple hierarchical levels in the human brain.
4. Discussion

4.1 Main findings

This thesis illuminated the effects of prior experience on speech perception using behavioural, MEG, and fMRI measurements. Study I revealed how initially unintelligible speech stimuli can be rendered intelligible by presenting them after their respective acoustically intact counterparts. Interestingly, this effect is stronger the more complex the stimulus is: Vowels remained difficult to understand (intelligibility 27%) even after the intact counterparts had been presented. In contrast, the intelligibility of the words increased from 19% to 45% and that of the sentences from 31% to 65% (Fig. 6a). In MEG measurements, the effect of presenting an intact speech stimulus was to enhance the response to a subsequently presented distorted stimulus. This enhancement occurred in the auditory cortex and surrounding areas at the latencies of 130–160 ms (Fig. 6b). Importantly, these differences cannot be attributed to acoustic variability but instead they likely reflect the prior exposure to the intact speech. Study II extended the results of Study I by suggesting that the dramatic increase in intelligibility was accompanied by increased BOLD responses in the in the bilateral frontal pole (FP), the anterior cingulate cortex (CC), paracingulate cortex (PCC), and the right frontal operculum; decreased BOLD responses were observed in the posterior insula (Ins), Heschl’s gyrus (HG), and the posterior superior temporal gyrus (STG, Fig. 9). Further, it seems that the primary auditory cortex is not only sensitive to acoustic-phonetic properties of the stimulus, but its activity also reflects prior experience.

Finally, Study III provided the evidence that prior experience accumulated through a lifetime can substantially modulate the perception of natural speech. The behavioural association experiment found evidence that the cultural family background of the listener can bias the interpretation of a narrative story (Fig. 10a). The fMRI results revealed that multiple hierarchical levels of language processing are shaped by cultural family background (Fig. 10b). Thus, it seems that even if we are using the same language, the meaning of the words may substantially vary depending on prior experience.
4.2 Rapid perceptual learning relies on word- and sentence-level long-term memory representations

Our results demonstrated that the intelligibility of distorted speech stimuli can increase dramatically if they are presented after their intact counterparts. For the first time, we show that this disambiguating influence provided by a prior exposure to the intact speech strengthens as a function of stimulus complexity. The distorted vowels were difficult to understand both before and after the presentation of their intact intelligible counterparts whereas the distorted words were considerably easier to understand when presented after the intact words. This effect was even stronger for the distorted sentences. This indicates that word- and sentence-level information is crucial in fast adaptive learning that allows comprehension of even severely distorted speech. Our results agree with previous studies which have found that words presented in semantically meaningful sentences are easier to recognize than words presented in isolation (Miller et al., 1951; Nittrouer and Boothroyd, 1990). Together, these results are in line with the psycholinguistic speech perception models which propose that the lexical access of a word in a sentence is increased by the preceding words which increase the activity of the memory representations of the words that are contextually consistent with the sentence (Gabriel, 2015).

Previous studies have shown that noise-vocoded speech stimuli can be rendered intelligible by presenting the disambiguating stimulus at the same time or immediately after the distorted stimulus (Clos et al., 2014; Obleser and Kotz, 2009; Sohoglu et al., 2012; Wild et al., 2012) or through long perceptual training (Davis et al., 2005; Giraud et al., 2004). The results of this thesis extend these results by showing that the memory representation activated by the disambiguating stimulus has a lifetime of at least several minutes and many representations of this kind can be maintained concurrently. This follows from our observation that improvements in comprehension of a set of several unique distorted speech stimuli in the D-I-D paradigm (48 sentences in Study I and 6 in Study I) last for at least six minutes (~2 min for words and ~6 min for sentences in Study I and 22–24 sec for sentences in Study II). Thus, it is unlikely that the intelligibility improvements observed in our studies rely primarily on echoic or short-term memory resources with limited lifetime and capacity, respectively. Instead, long-term memory storage would seem to have both the capacity and lifetime necessary to support the intelligibility enhancements.

Our results agree with Tiitinen et al. (2012) who showed that intelligibility of a set of 120 distorted sentences increased from 30% to 50% when compared before versus after the presentation of the set of intact sentences. However, the sentences in that experiment consisted of only seven starting words, three sentence stubs and four ending words, whereas in the studies of this thesis each sentence was unique and presented only once. Hannemann et al. (2007) also reported that from a set of 34 unintelligible digitally degraded nouns, intelligibility increased for the nouns that were previously presented in the intact form but not for the nouns that were only presented in the distorted form. However, this experiment included ten
repetitions of each noun. While repeating the same words and sentence stubs may have increased speech intelligibility in the above studies, our results show that even a single presentation of a speech stimulus leads to the formation of a memory representation that can be used in subsequent processing of distorted speech.

4.3 Brain responses already at the early stages of speech processing are modulated by experience

The D-I-D paradigm resulted increased mean current density to the second presentation of the distorted stimuli in the auditory cortex and surrounding areas already at the latencies of 130–160 ms, when comparisons were made to responses elicited by the first presentation. The corresponding contrast decreased BOLD activity bilaterally in the brain area extending from the HG to the middle STG and in the right hemisphere to the supramarginal gyrus (SMG, Study II). Additionally, ISC effects reflecting the influence of cultural background on speech perception were found already in HG, the earliest stage of the cortical speech processing hierarchy (Study III). These latencies and brain areas have typically been associated with acoustic and prelexical processing. Importantly, in this thesis these differences were found between cortical responses to acoustically identical stimuli and, therefore, they cannot be related to the acoustic differences of the stimuli. Instead, the differences probably reflect speech-disambiguating changes in brain function elicited by the prior presentation of the intact stimuli in studies I and II and cultural family background in Study III. Similarly, Wild et al. (2012) found that BOLD responses in the primary auditory areas reflect increased intelligibility of noise-vocoded speech obtained by presenting it with visual matching text. Also, Tiitinen et al. (2012) also found P2 response in auditory cortex to be modulated with speech intelligibility using the D-I-D paradigm. Together, these results suggest that already the phonetic-acoustic processing of speech at the early stages of the cortical speech processing hierarchy is modulated by experience.

4.4 Evidence for contribution of extralinguistic areas in the comprehension of acoustically distorted speech

Study II revealed more pronounced BOLD responses for the second compared to the first presentation of the distorted sentences in an extensive network of extralinguistic brain areas: the bilateral anterior CC/PCC and FP, the right frontal operculum as well as in the area from the left precentral gyrus (PCG) to the left middle frontal gyrus (MFG). These differences are probably related to the comprehension of degraded speech specifically rather than to speech comprehension in general since no differences were found in these areas in the contrast between the presentations of the intact stimuli and the first presentation of the distorted stimuli. Interestingly, the increased BOLD-response in the bilateral FP may reflect activation of long-term memory
representations since those areas were not associated with increased speech intelligibility when the disambiguating stimulus was presented immediately after the distorted stimulus in previous studies (Clos et al., 2014; Hervais-Adelman et al., 2012; Zekveld et al., 2012).

These results are in line with a recent review (Peelle, 2018) that proposed cingulo-opercular (consists of bilateral anterior CC and frontal operculum), premotor (consists of PCG) and frontoparietal networks (consists of bilateral MFG and inferior parietal cortex) to be involved in speech processing in adverse listening conditions. These networks have been associated with performance monitoring, verbal memory, and attention, respectively. However, the activation of the cingulo-opercular network was surprising in our study since it is usually activated when speech intelligibility is low but in our study the intelligibility of the distorted speech at their second presentation was at a near perfect level. Interestingly, bilateral activation of the anterior CC, FP, and frontal operculum have also been associated with the episodic memory retrieval mode where sensory information acts as “a retrieval cue” for information in episodic memory (Lepage et al., 2000). Similar processing might have increased intelligibility of degraded speech in our study. In this scenario, the distorted stimuli are treated as retrieval cues that are compared to the memory representations activated or created by a previous exposure to the intact stimuli. If the distorted speech contains enough information to trigger successful memory retrieval, the distorted sentence then becomes intelligible. This interpretation would also agree with the Easy of Language Understanding model which proposes that in acoustically suboptimal conditions, a mismatch between the distorted signal and the LTM trace triggers explicit memory retrieval during which the incomplete acoustic information is completed by the information stored in LTM (Rönnberg et al., 2013, 2010; Rudner and Rönnberg, 2008).

4.5 Similarity in family cultural background enhances BOLD synchrony between individuals at multiple levels of speech processing

We found behavioral and neuroimaging evidence that family cultural background can introduce divergence in the perceptions of a spoken narrative, even when the subjects have all grown up in the same country and are native speakers of the language of the narrative. Specifically, when the narrative was delivered in 101 segments, the behavioral word-listing association experiment (Saalasti et al., 2019) revealed group differences between cosine-similarity of word-list meanings in over half of the segments. The Finnish-background subjects showed significantly higher similarity in 12 segments compared to the subjects with a Russian background, and the reverse was true in 44 segments of the narrative. This suggests that those story segments have presumably contained elements that elicit the family cultural background related associations within either one group or the other, reflecting also higher similarity of perception within one group compared to the other. Interestingly, the between-group differences were not limited in the segments that contain
cultural elements, suggesting that cultural background can also modulate the perception of elements that are not obviously related to the culture.

Further, the results of Study III revealed that cultural family background modulated ISC of brain hemodynamic activity within the bilateral HG, STG, MTG, lateral occipital cortex (LOC), precuneus (PCun), superior parietal lobue (SPL), occipitotemporal cortex (OTC), lingual gyrus (LG), and occipital pole (OP). The STG, MTG, and PCun have been shown to be activated during speech comprehension and semantic processing in many neuroimaging studies (Regev et al., 2013, for a review, see Hickok and Poeppel, 2007). Lerner et al. (2011) demonstrated that isolated words and sentences activate mainly temporal cortices, whereas paragraphs of connected speech activate also the frontal cortices and PCun. This was proposed to reflect a distributed hierarchical network of brain areas where early sensory cortices accumulate information over short time scales of up to 100 ms and higher-order areas accumulate information over longer time scales up to minutes. Together, these previous studies and the results of the association word-listing experiment suggest that the between-group ISC differences in the STG, MTG, and PCun reflect differences in semantic processing at multiple timescales. Between-group ISC differences in the visual areas, the LOC, LG, and OP, might be explained by differences in visual imagery during narrative listening (Bartolomeo, 2002; Borst and Kosslyn, 2008).

The results of Study III can be seen as an advance in the recently emerged field of cultural neuroscience. Previous studies in this field investigating the influence of culture on cognition have presented subjects with pictures (Ames and Fiske, 2010) and gustatory stimuli (McClure et al., 2004). In our study, the narrative that described connected events and social interactions allowed us to study how cultural family background modulated both the interpretation of the narrative and the BOLD-responses elicited during listening. This was an ecologically highly valid setup, with natural connected speech being essential for smooth everyday communications. Further, most previous studies have examined cultural variation in psychological and neural processes using subjects from Western and Eastern countries (usually Americans vs. East Asians, for reviews, see Ames and Fiske, 2010; Chiao et al., 2010; Gutchess and Indeck, 2009). In contrast, all subjects in our study were Finnish residents fluent in Finnish, and all of them raised in Finland. Thus, the cultural differences between the subjects were quite subtle via half of the subjects having either one or two Russian parents, and the other half having Finnish parents. Our results suggest that even the relatively subtle factor of one’s parents cultural background influences how natural connected speech is processed and interpreted.

### 4.6 Limitations and future perspectives

Our results suggest that prior experience modulates responses already in the early stages of speech processing (i.e. in HG and STG). This could be explained with top-down models which propose that information stored in higher order
associative brain areas modulate speech processing in the sensory areas through feedback connections (Clos et al., 2014; Friston, 2005; Sohoglu et al., 2012; Tunnerhoff and Noppeney, 2016). However, our results also do not rule out the possibility that the predictive information is stored locally in the auditory cortices (Gagnepain et al., 2008). Further analyses, such as determination of causal relationships between the modulations of activity between the early auditory areas and the higher order associative brain areas could help to establish whether the results reflect bottom-up or top-down processing.

Another issue that deserves further examination is the influence of attention on the processing of degraded speech in the D-I-D paradigm. Currently, the evidence pertaining to the importance of attention in perceptual learning and processing of acoustically distorted speech is contradictory. Tiitinen et al. (2012) found that prior exposure to intact speech was reflected in the P2 response and the sustained field elicited by distorted sentences in the auditory cortex and areas anterior and posterior to that area when the subjects were not attending to the stimuli. In contrast, in this thesis, the prior exposure to intact speech was reflected in the N1 response, but not in the P2 response and sustained field when the stimuli were unattended. A previous fMRI study provided evidence that unattended noise-vocoded speech can be processed to some extent (Wild et al., 2012). However, behavioral evidence indicates that subjects who did not attend to noise-vocoded speech during the training session are not able to recognize noise-vocoded sentences better after training (Huyck and Johnsrude, 2012). Thus, attention seems to be necessary at least in adaptive learning that results in incremental improvements in perception over a timescale of minutes or longer. It would be interesting to study whether the intelligibility of the distorted speech increases in the D-I-D paradigm if the subject is not attending to the sentences when they are presented in the intact form. Further, neuroimaging experiment could be conducted to compare the brain responses to the attended and unattended degraded speech presented after attended intact speech.

Our results show that the intelligibility of the distorted sentences at their second presentation decreases when the number of unique sentences in the D-I-D set is increased. An average increase of intelligibility over 18 D-I-D experiments with six sentences was stronger (54 percentage points) than the increase of intelligibility in one D-I-D set with 48 sentences (34 percentage points). Together, these results suggest that there is a capacity limitation of the memory system probed in the D-I-D paradigm. The capacity limitation of the memory store could be tested by investigating how the number of sentences in the sentence set affects the intelligibility of the distorted sentences at their second presentation (e.g. three D-I-D sets with 5, 20 and 80 sentences). If the capacity of the memory store is limited, it can be expected that the intelligibility decreases when the number of sentences increases. Further, duration of the memory store could be tested by varying the length of the time gap between the intact sentences and the second presentation of the distorted
sentences. If the duration of the memory store is limited, it can be expected that the intelligibility decreases when the time gap increases.

In Study III, we used a nonparametric t-test to show that the cultural family background modulates ISC values of BOLD-responses to the narrative. The behavioral association experiment together with the previous neuroimaging studies about semantic processing (Binder et al., 2009; Lerner et al., 2011; Specht, 2014) suggest that at least some of the between-group differences in the ISC values reflect differences in story interpretation. We also compared the behavioral results with the BOLD responses directly to see which brain areas reflect differences in the interpretation of the story, but this yielded no statistically significant results. A possible reason for this could be that word associations from more than currently used 101 time instants of the narrative would be needed to track changes in the narrative interpretation reflected in BOLD responses measured with a sampling rate of 100 ms.

An important question for further studies is also which issues in the cultural family background specifically (e.g. the fluency in Russian language, frequency of visits to Russia, how Russian or Finnish does the subject feel) are reflected in the perception of the narrative. This could be tested, for example, by applying regression methods to find brain regions where the scores of the background questionnaire and IAT could be associated with the BOLD responses. Moreover, the cultural family background is only one factor that affects what kind of experience and knowledge we accumulate through our lifetime. Other factors are, for example, age, and socioeconomical or political backgrounds. It would be interesting to conduct a related experiment as in Study III to see whether these other factors affect speech perception and processing. An interesting question also is whether the results obtained in Study III can be generalized to differences between any cultures.
5. Conclusions

Our behavioural and neuroimaging results suggest that prior experience is critical in speech comprehension. Only a single prior presentation of an intact speech stimulus renders the corresponding initially unintelligible distorted stimulus intelligible. The intelligibility of sentences increased more than the intelligibility of words, whereas vowels remained unintelligible. This result presumably reflects activation of word- and sentence-level long-term memory representations. The prior exposure to intact speech modulated processing of its acoustically distorted counterpart in the primary auditory areas and early latencies indicating that already early stages of speech processing are modulated by prior experience. The final study of this thesis provided evidence that experience accumulated throughout our lifetime modifies how we interpret and process narrative. The results of this thesis increase our understanding of the neural basis of speech perception. As such, they may help to refine models of speech perception in noisy real-life environments and in hearing impaired individuals. Moreover, the results might be used to promote mutual understanding between individuals from different cultural backgrounds.


Cutler, A., Oahan, D., Van Donselaar, W., 1997. Prosody in the Comprehension of Spoken Language: A Literature Review*, LANGUAGE AND SPEECH.


Füllgrabe, C., Rosen, S., 2016. Investigating the Role of Working Memory in


References


References


doi:10.1002/mrm.20997
Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 435, 869. doi:10.1038/nature06976


