Unspoken Speech

Speech Recognition Based On Electroencephalography

Lehrstuhl Prof. Waibel
Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA
Institut für Theoretische Informatik
Universität Karlsruhe (TH), Karlsruhe, Germany

Diplomarbeit

Marek Wester
Advisor: Dr. Tanja Schultz

31.07.2006

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 31.7.2006

Marek Wester

Abstract

Communication in quiet settings or for locked-in patients is not easy without disturbing others or even impossible. A device enabling to communicate without the production of sound or controlled muscle movements would be the solution and the goal of this research.

A feasibility study on the possibility of the recognition of speech in five different modalities based on EEG brain waves was done in this work. This modalities were: normal speech, whispered speech, silent speech, mumbled speech and unspoken speech.

Unspoken speech in our understanding is speech that is uttered just in the mind without any muscle movement. The focus of this recognition task was on the recognition of unspoken speech. Furthermore we wanted to investigate which regions of the brain are most important for the recognition of unspoken speech.

The results of the experiments conducted for this work show that speech recognition based on EEG brain waves is possible with a word accuracy which is in average 4 to 5 times higher than chance with vocabularies of up to ten words for most of the recorded sessions. The regions which are important for unspoken speech recognition were identified as the homunculus, the Broca's area and the Wernicke's area.

Acknowledgments

I would like to thank Tanja Schultz for being a great advisor, providing me feedback and help whenever I needed it and providing me everything that I needed to get my thesis done and have a good stay at CMU. I would also like to thank Prof. Alex Waibel who made the InterAct exchange program and through this my stay at CMU possible. Great thanks to Szu-Chen Stan Jou for helping me to get to know Janus. I also want to thank Jan Calliess, Jan Niehues, Kay Rottmann, Matthias Paulik, Patrycja Holzapfel and Svenja Albrecht for participating in my recording sessions. I want to thank my parents, my girlfriend and my friends for their support during my stay in the USA. Special thanks also to Svenja Albrecht for proof reading this thesis.

This research was partly funded by the Baden-Württemberg-Stipendium.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic speech recognition is supposed to provide a solution in human-machine communication. It enables the communication with computers in a natural form. In the beginning of the research in speech recognition computing power was a problem in order to do reliable speech recognition in real time. Since the fast increase of computing power this problems vanished but other conceptual problems remained. The recognition of speech in noisy environments is still an unsolved problem. Speech impaired people having problems to utter speech correctly are also a difficult task for a speech recognizer. Sometimes it would be even desirable to communicate while uttering speech is not possible like in different environments e.g. under water or in very quiet environments. In the described situations communication through unspoken speech would be ideal because it would be the only solution for the described problems.

In this work we define unspoken speech as follows: it is speech which is thought as if it would be spoken. To learn the production of unspoken speech a person would have to start with uttering a word in normal speech. The next step would be to think of nothing while uttering the word besides the thoughts needed to produce this speech. The final step would be to do the same as in the step before without any muscle movement. This is what we understand as unspoken speech.

1.1 Goal of this Research

In this work we want to investigate if the recognition of unspoken speech is feasible. To show this we employed electroencephalography (EEG) measurement of the human brain at the scalp. The underlying idea is that every muscle movement is preceded by an activation of neurons in the brain. This activation involves electrical signals which are measured with electrodes attached to the scalp. The research in this field shows that there is a connection between the recorded EEG-data and speech production. We want to investigate if this is also true for unspoken speech. To achieve this goal we divided this work in three subtasks.

The first subgoal is to find out if the recognition of normally spoken speech using EEG-data is possible. This step should show that there are patterns in the EEG-data while speech is produced in a normal speech modality which could be recognized with the methods of automatic speech recognition.

In the second subgoal we want to investigate how well this recognizer performs for different modalities of speech production, namely: whispering, silent speech, silent mumbling and finally unspoken speech. This modalities can also be seen as a degeneration of normal speech production to unspoken speech.

In the final subgoal that is described in this work we investigated if data that is collected around the region of the brain that is considered as being responsible for muscle movement (homunculus) and the regions that are considered to be responsible for speech (Broca's area and Wernicke's area) are sufficient to recognize unspoken speech.

The main goal of this work is to investigate if it is possible to recognize naturally thought arbitrary unspoken speech with adjusted methods of standard automatic speech recognition applied on EEG data.

1.2 Motivation

Language recognition without the need to speak out loudly or speak at all is useful for many applications.

Sometimes it would be very convenient to have an EEG based speech recognizer. An example is a very quiet setting like an opera performance or a library. No sounds should be produced there. It is for example not possible to answer a phone call. Communication in this situations would be possible with a recognizer for unspoken speech. For example the person in the opera performance could use unspoken speech to answer the phone and just listen to what the caller has to say and answer with a limited set of unspoken words which are then synthesized into audible speech for the caller.

While solving a convenience problem is a nice to have feature, there are areas where no general purpose solution exists today to enable people to communicate with others. One area where our research can help are people like locked-in patients whose only chance to communicate with their environment is currently through rough speech or eye blinking. This people could use an EEG system to control a computer with their thoughts. Even a small vocabulary of about ten words would be sufficient to control basic commands on the computer. Using a T9 [10] spelling system like it is used in most cell phones these days they could even write letters or chat with other people. The enrichment of their lives might even be worth the hassle with an EEG cap on their head and the gel filled hair after using it.

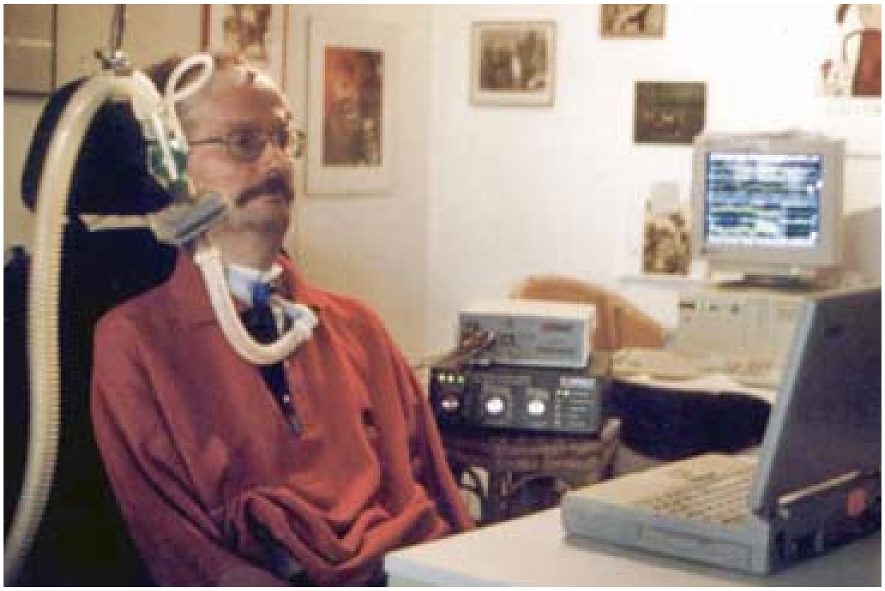


Figure 1.1: Locked-In patient using the Thought Translation Device[1] to control a computer

Another group of people who would benefit from the system would be people who are in situations where usual speech recognition or even simple communication is not possible. This are for example fire fighters while wearing a thermal protecting fire suit with an oxygen mask when fighting a fire. While exposed to extreme temperatures the firefighters are already in bad physical conditions and through the noise produced by the fire it is hard for them to produce speech that can be understood trough the radio communication system. It would be less stressful to call for reinforcements while just thinking it than to shout it. Again a small set of commands is sufficient in such situations. Another group are scuba divers. Since most rebreahers are put into the mouth it is not possible for them to utter anything. A set of thinkable commands would help them to get any communication. For both of the described groups an additional EEG cap would not add to the burden of the equipment they are already wearing to accomplish their tasks.

1.3 Ethical Considerations

The recording and recognition of human thoughts is an invasion of the privacy of the recorded subject. The recorded data alone includes personal information about the subject.

The recorded data can e.g. include information about mental disease of the subject as

Koles describes in[11]. The subjects from whom the data was collected, were apprised of this fact before they decided if they wanted to take part in the recording. Since our group has no intent to investigate mental disease in EEG-data, data would not be used for examinations other than research topic of speech recognition in EEG-data, as the data was collected just for this purpose.

Future improvements may make it possible to not just recognize trained data but also random thoughts. This possibility might be used for interrogations and lie detection. Interrogation methods which involve mind reading may be considered as being illegal e.g. in the USA through the Fifth Amendment of the United States Constitution: ”...nor shall be compelled in any criminal case to be a witness against himself...”[12]. Though this kind of technology may be misused by criminal people for interrogations. The purpose of our research is not mind reading but the recognition of unspoken speech and we refrain from misusing this technology against the will of people.

The only purpose of the research done for this work is to support people in the fulfilling of their tasks and not to spy on them or to intrude their privacy.

1.4 Structure of the Thesis

In chapter 2 the theoretical background is described that is necessary to understand the following chapters. Information about the speech recognition system Janus, about the feature extraction methods used, electroencephalography, the brain and the recording technology can be found there.

In chapter 3 the related work in the field of unspoken speech recognition in EEG data are discussed and it is shown which new contributions come from this thesis.

An overview over the recording system, the process of recording, the training and the recognition is given in chapter 4.

The collected data is described in chapter 5. The different corpora and modalities are introduced there.

The results of the conducted experiments and therefore the main part of this thesis are explained in chapter 6. Chapter 7 describes the demo system that was built to test our recognition methods online. In Chapter 8 a summary with conclusions and an outlook to future work is given.

The appendix describes the technical background and a documentation of the software created for this thesis. A list of all recordings can also be found in the appendix.

Chapter 2

Background

2.1 Janus

The Janus Recognition Toolkit is a framework developed for speech recognition of normal speech developed by Interactive System Labs at University of Karlsruhe, Germany and Carnegie Mellon University, Pittsburgh, USA[13]. The recognition system developed for unspoken speech recognition is based on the Janus framework. A technical overview can be found in the Appendix A.1. A theoretical overview will be provided in this section.

To initiate the recognition system a state of the art recognizer for normal speech was chosen and iteratively adapted to a recognizer for unspoken speech.

The first step in the training of the recognizer is the segmentation of the speech. The recordings are always starting with silence followed by a word and then again silence. The detection of silence in EEG- data is an easy task if muscle movement is involved, since the movement results in large amplitudes of the brain waves which make the distinction of speech and silence easy.

A problem arises when no muscle movement is involved. Brain waves of speech vs silence are hard to discriminate. Because of that, speech had to be marked in a procedure controlled by the subject. This was done by one eye blink before uttering the unspoken word and one eye blink after the uttering. The high amplitudes produced by the eye blinking which were easy to detect served as a marker for the speech part. Because the recordings did concern single isolated words rather than continuous sentences, a more sophisticated segmentation was not needed.

Features were computed as described in the next section. This computation resulted in a high dimensional feature vector of 192 dimensions. This feature space was reduced to 35 dimensions with the linear discriminant analysis.

A left-to-right Hidden Markov Model[14] with five states and 25 gaussians per state was trained for every word in the vocabulary. The shape of the gaussians is represented by a diagonal matrix. The Expectation Maximization algorithm with four iterations was used for the training.

Finally the recognition was conducted by the computation of a Viterbi path for every word of the vocabulary which was recorded. The word with the best Viterbi score was selected as the hypothesis.

2.2 Feature Extraction

The features in speech recognition are different from the features which were used for the recognition of unspoken speech. Usually acoustic speech recognition relies on frequency based features, extracted from the speech signal. There is a huge difference in the data density of the recorded waves. While in audible speech data is recorded through one channel with 16 kHz, brain waves were recorded through sixteen channels with 300 Hz each. An example for brain waves in contrast to audible sound wave can be found in figure 4.4

The following features were used in the unspoken speech recognizer:

- windowed Short Time Fourier (STF)[15] coefficients: the STF coefficients were used with a window size of 26.6 ms and a window shift of 4 ms. This parameters were chosen because of experimental results.

- delta coefficients: the delta coefficients were used and also the delta coefficients of the delta coefficients (delta delta coefficient) were used. A delta coefficient is the first deviation, while the delta delta coefficient is the second deviation.

- delta mean coefficients: the delta mean is a delta coefficient of a windowed mean

The resulting features were concatenated to form a single feature vector. The dimensionality of the resulting feature vector was reduced with the linear discriminant analysis[16].

2.3 Electroencephalography

The recording of electrical activity of the human brain, known as electroencephalography, was first done by Hans Berger in 1929[17]. The electroencephalography (EEG) is a method to record the electrical potentials produced by the brain close to its surface. For this purpose electrodes are positioned either on the scalp or directly on the cortex. In the case of this thesis we used electrodes positioned on the scalp.

The electric potentials that can be measured on the surface of the skull are due to the information transfer which happens in the brain between the neurons which the brain consists of. This process is described in more details in section 2.4.1.

The EEG is considered to have a high temporal resolution of up to 80Hz. We used a higher sampling rate for our recordings as it would be required to avoid aliasing. The slope in the bandpass filter of our amplifier is very small so that we are using a sampling rate of 300 Hz. This makes it ideal for speech recognition of thoughts. On the other hand it records a three dimensional compound using electrodes at the surface while reducing it to a two dimensional space. And even the spatial resolution in this two dimensionalities is not high as Paul Nunez states in [18]. He says that one scalp electrode records electrical currents generated in cortical tissue containing approximately 30-500 million neurons. While technologies like e.g. computer tomography, positron emission tomography or magnetic resonance image have a high spatial resolution, EEG has the highest temporal resolution. This is important for the recognition of unspoken speech that requires the observation of rapid changes over time. Another advantage of EEG is that it is relatively inexpensive and easy to transport because the recording device fits in every pocket, while this is not true for the recording devices of the other structural brain imaging methods.

EEG is also the only method which measures the electrical potentials produced by the neurons in the brain directly. Other methods rely on the blood flow or metabolisms which are not coupled with the electric potentials produced by the neurons.

The EEG recording system consists of electrodes, amplifiers and a recorder. The electrodes are attached to a cap which is placed on the subjects head to keep them in position. The cap is covered in section 2.5. The most commonly used way to distribute the electrodes over the scalp is an uniform distribution using the International 10-20 System introduced by the International EEG Federation in 1958 [19]. Figure 2.1 shows an example for the 10-20 distribution.

To reduce impedance, a conductive gel is often applied between the scalp and the electrodes. The gel also helps to get the electrodes connected to the scalp through hair so there is no need for shaving the head of the subject.

The electrodes are connected to an amplifier and filter combination and the resulting signal is recorded. The recorded signals are called brain waves. The amplitude which can be measured on the scalp is about $200\mu V$[18].

There are three ways to measure the potential. Average reference derivation is the name of the first way. All signals are averaged and the resulting signal is used as a common reference for the amplifier. The second way is the common reference derivation. The reference electrodes are placed e.g. at the earlobes. All electrodes are measured then relative to this

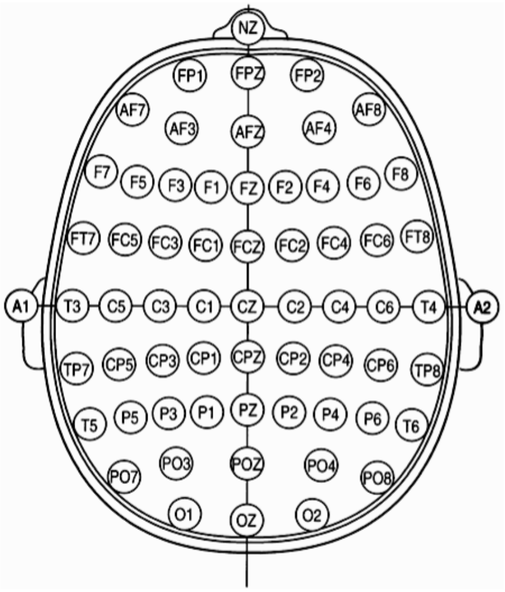


Figure 2.1: The international 10-20 system for distributing electrodes on human scalp for EEG recordings[2]

reference. The last way is the bipolar derivation. The electrodes are connected in a way that potential differences between adjacent scalp electrodes are measured, e.g. an amplifier measures the difference between electrode 1 and electrode 2. The second amplifier measures the difference between electrode 2 and 3 and so on.

EEG recordings are very vulnerable to artifacts. These artifacts can be produced by the environment. A source might be the VGA[1]-outlet of a computer which produces electromagnetic interferences. Another source for artifacts might be the recording hardware. The artifacts can also come from the recorded subject. Every body movement causes large artifacts. Automatic artifact removal works as Nunez states in [18] only for the largest artifact because the artifact band and the important band, which contains the EEG information that should be extracted, overlap.

2.4 Brain

This section we will explain the basic unit of the brain, the neuron, and how it works and how through its work electrical potentials are produced which can be measured afterwards. After that follows a introduction of the different language areas in the brain. This is followed by a section describing the process of speech production. The last section explains the idea

[1] Video Graphics Array (VGA)

behind this work.

2.4.1 Information transfer

The major class of cells which are responsible for message transfer in the brain are called neurons. They are also the foundation of the nervous system. A typical neuron as shown in figure 2.2 consists of the cell body (soma) filled by cytoplasm that is containing a nucleus. There are two extension of the soma which are dendrites which collect electrical potentials from other neurons and the axon transports electrical potentials to other neurons or muscle cells.

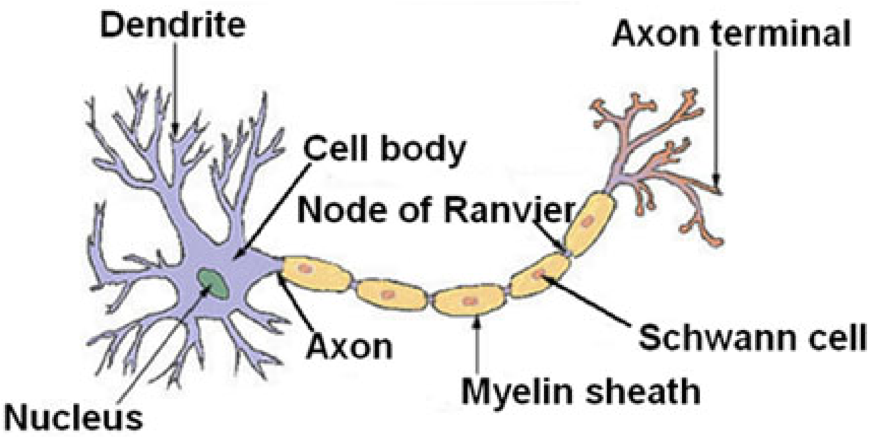


Figure 2.2: Model of a neuron[3]

For a communication between neurons to occur they have to be connected to each other one the one side with the dendrite and on the other side with the axon terminal. This connection is called synapse. Through this junctions the cells exchange electrical potentials through chemical processes. There are two kinds of synapses: exhibitory and inhibitory. Exhibitory synapses increase the potential in the connected neuron and inhibitory synapses decrease this potential. If and only if enough exhibitory potentials are generated to exceed a certain threshold a so called action potential is evoked. This potential is then transported through the axon of the neuron to other neurons or muscle cells.

The potential inside a neuron is about -70mV. This is measured relative to extracellular fluid. In order to have such a negative level the cell has to keep charged ions inside the soma. So it has a cell membrane that does not let the ions inside the cell or let them get outside the cell.

To keep this negative level the cell has two strategies. The first passive one is that the soma has proteins that can be opened and closed for K^+ and Na^+ ions. The ion concentration of a neuron cell of a mammal is shown in table 2.1. If this protein is opened, K^+ flow out of the cell because of diffusion until the electric potential which changes with this flow stops the diffusion. The second strategy is active and is an ion pump which actively pumps two K^+ ions in and three Na^+ ions out. This results in a more negative soma.

	intracellular fluid	**extracellular fluid**
K^+	155	4
Na^+	12	145
Cl^-	4	120
Ca^{++}	10^{-8}-10^{-7}	2

Table 2.1: Ion concentration in a muscle cell of a mammal[8]

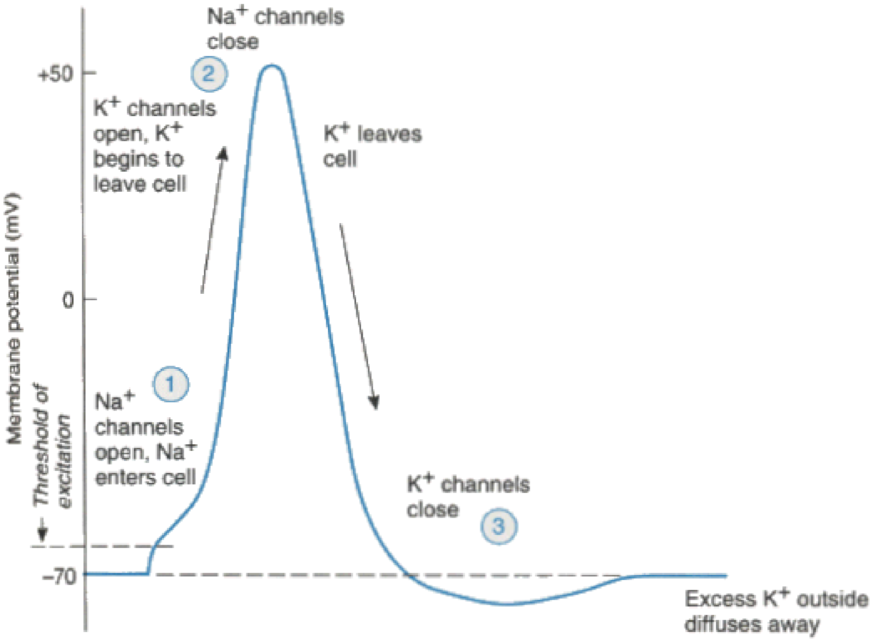


Figure 2.3: The flow of ions during an action potential[4]

When an action potential hits a synapse it causes a flow out of neurotransmitters which opens the proteins to let Na^+ flow in (caused by diffusion) and the cell membran gets more positive as shown in figure 2.3 (1). At a certain threshold the K^+ proteins open and K^+ flows out (2). After reaching the highest point the Na^+ proteins close and because there is still more K^+ inside the cell than outside, the K^+ are still leaving the cell. Finally the K^+ protein closes (3) and the ion pump does the rest of the work to get a concentration as in table 2.1.

It is believed that a summation of this action potentials in cortical cells is what can be measured with EEG. While Meyer-Waarden[20] also explains this theory he states that there were no experimental results proofing this. He explains another theory that the signals can also come from the brains surface where mostly dendrites and synapses are located.

In order to active a muscle to e.g. produce speech the action potential finally has to reach a muscle fiber and make it contract. The connection between an axon and a muscle is called neuromuscular junction and is also a synapse. For the action potential to pass the neuromuscular junction it activates the spilling of the neurotransmitter acetylcholine in the neuromuscular junction. This transmitter binds to receptors at the motor end plate located at the muscle which causes the motor endplate to be depolarized which causes a depolarization of the muscle fiber and results in a muscle contraction.

2.4.2 Brain and Language

While in normal speech recognition the vocal tract as the part of speech production is the point of interest in this work the brain as the source of unspoken speech is the subject of investigation. Ramachandran [18] gives a detailed explanation of the brain and its functions. This section will focus on the parts of the brain which we believe to be most important for the production of unspoken speech.

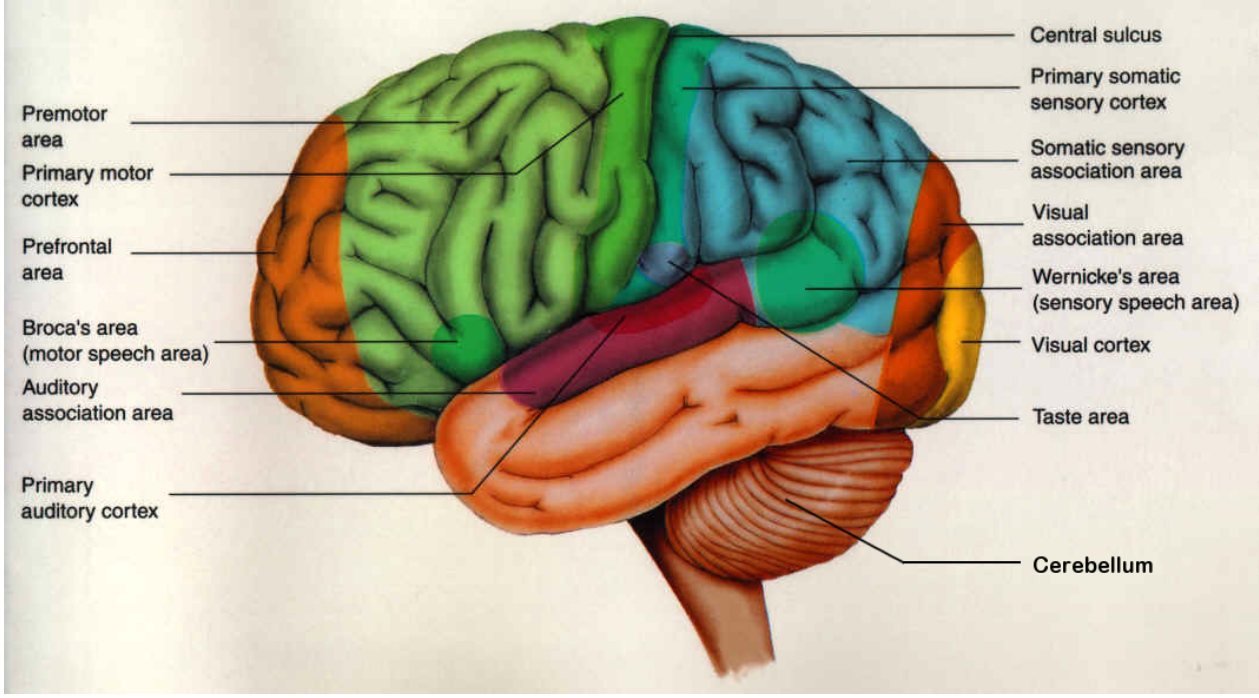

Central sulcus
Primary somatic sensory cortex
Premotor area
Somatic sensory association area
Primary motor cortex
Visual association area
Prefrontal area
Wernicke's area (sensory speech area)
Broca's area (motor speech area)
Visual cortex
Auditory association area
Taste area
Primary auditory cortex
Cerebellum

Figure 2.4: Left side of the brain showing the important regions of the brain for speech production like primary motor cortex, Broca's area and Wernicke's area (modified from [5])

A model of the human brain is depicted in figure 2.4. The model shows the left side of the brain with the front of the brain on the left side of the figure. Three parts of the brain are interesting for this work and for speech production: Broca's area, Wernicke's Area and the primary motor cortex.

The Broca's area was discovered by Paul Broca[21] in 1861 by autopsy. Broca found out that this area was injured in the brains of persons having difficulties to articulate words. Sometimes they could just utter a hand full of words. This area of the brain is thought to be responsible for the articulation of words. Broca's area is located on the left side of the brain.

The Wernicke's area is also located at the left side of the brain as shown in figure 2.4. It was discovered by Carl Wernicke in the 19th century. Wernicke found that a lesion in this area leads to speech without language. This means that people can speak fluently but the spoken output makes no sense. They are just able to utter meaningless words and sentences sounding correctly.

The primary motor cortex, also known as "homunculus", is depicted in figure 2.5. This part of the brain is responsible for the movements of most parts of the human body and more specifically for the vocal speech tract. The figure shows which parts of the motor cortex are responsible for which part of the body. The size of the body on the map do not correspond to the actual size, but to the actual brain portion part to control this particular part of the body. So there is as much brain mass to control the face as to control the legs but the face is much smaller. The consequence is that there is a lot information to be gathered from the homunculus concerning the movement of the face and therefore speech production.

Before the primary motor cortex lies the premotor area which supports the primary motor cortex in the planing of movements. The Broca's area is located in the premotor area though generates the movement patterns for the production of speech. It works together with the cerebellum. The cerebellum is a connection point of sensory feedback and the muscle movement. It coordinates the movement depending on the sensory feedback like e.g. how hard to push a button.

2.4.3 Speech Production in the Human Brain

The production of speech in the human brain is a field of ongoing research. In this section the Wernicke-Geschind-Model[22] is going to be introduced which is a well know classic theory about the production of speech after hearing a word. More recent research shows that this model is oversimplified [23]. Nevertheless the Wernicke-Geschwind-Model is the basis for more sophisticated models. The model also gives a theoretical fundament for the findings in this work.

Figure 2.6 shows the path that the neural signal follows according to the Wernicke-Geschwind-Model when a person hears a word and then repeats the word. First the word

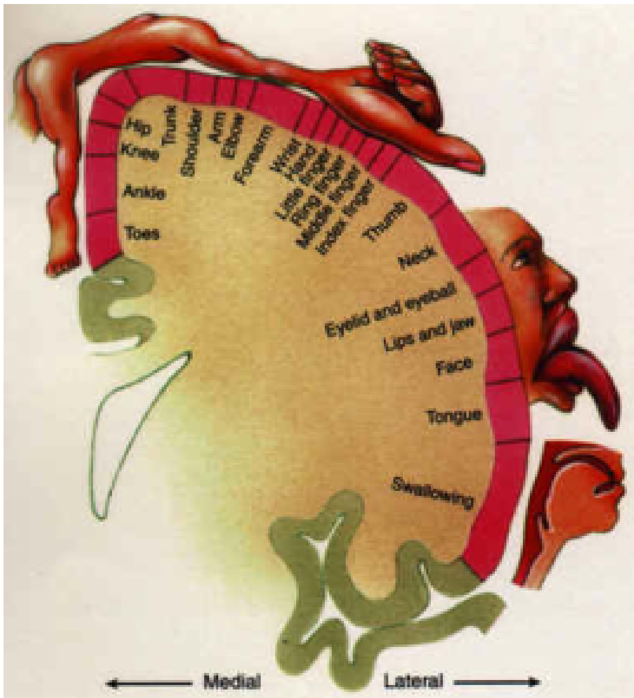


Figure 2.5: Homunculus area, also know as primary motor cortex. This part of the brain controls most movements of the human body[5]

is processed in the primary auditory area. The semantics are extracted and also added in the Wernicke's area. As Mamoli [8] states a lesion of the Wernicke's area can lead to wrong naming of words in speech production therefore semantics are also added to the word which is going to be uttered. The signal advances through the arcuate fasciculus which is the connection between the Broca's area and the Wernicke's area to the Broca's area. A plan for the motor cortex is formed in the Broca's area. The plan is implemented then in the motor cortex with the manipulation in the vocal tract.

2.4.4 Idea behind this Work

Normal speech involves the innervation of muscles. To innervate muscles action potentials are needed which can be measured with the EEG. Brain waves result from action potentials which finally lead to the innervation of muscles and through this to speech production. This brain waves affect different areas in the left part of the brain according to the Wernicke-Geschwind-Model and to further work in this area. Following the Wernicke-Geschwind-Model it can be said that this process is involved in every speech production. The idea behind this work is that it should be possible to recognize patterns from the data collected through the EEG while speech is produced.

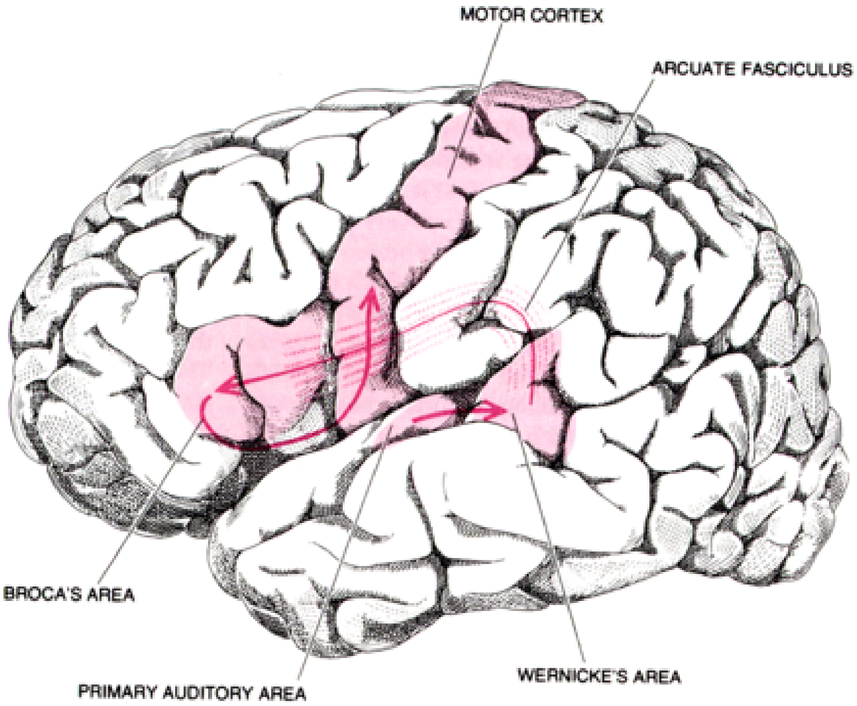


Figure 2.6: A graphical representation of the Wernicke-Geschwind-Model[6]

During the different modalities the muscle movement decreases more with every modality until in the unspoken modality no muscle movement is involved at all. Through this process the involvement of the primary motor cortex gets lower. But as we believe the involvement of the other regions involved in speech production stays at a level that pattern recognition is still possible because speech is still produced. The Wernicke-Geschwind-Model stays valid because unspoken speech as defined by us is speech without muscle movement. But still movement patterns should be produced in the Broca's area which then should be recognized. No mind reading should be done, just patterns should be recognized in the process of speech production as described in the Wernicke-Geschwind-Model.

2.5 Cap

The cap that was used for the recordings was supplied by Electro-Cap International, Inc[2]. It is equipped with 20 electrodes using the International 10-20 method [19]. It is made of an elastic spandex-type fabric. The electrodes are made of Ag/AgCL and are recessed and attached to the fabric. Because they do not touch the skin of the subject directly they have to be filled with a conductive gel as shown in picture 2.7. The process of filling the electrodes

[2] http://www.electro-cap.com/

also lowers the impedance of the skin because during this process skin is abraded.

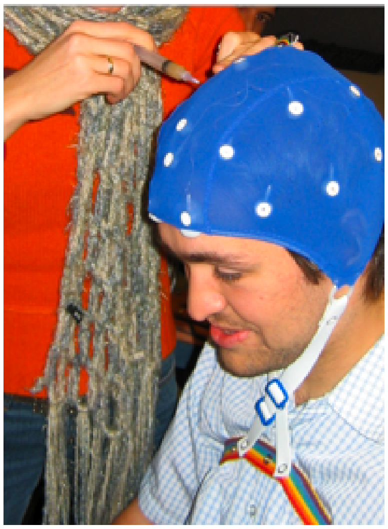


Figure 2.7: Electro-Cap being filled with a conductive gel

The cap is attached to the subject with straps which presses the electrodes closer to the scalp. The straps are connected to a band which is attached around the upper part of the body under the axles. This tension is important so that the gel can not run out of the electrodes. On the other hand this pressure inflicts pain to the subject over time because the electrode fittings are made out of hard plastic. This pain may lead to artifacts in the recordings.

Chapter 3

Related Work

This chapter describes the related work. However since this study is to the best of our knowledge the first that addresses the recognition of unspoken speech with EEG therefore no literature was found that describes approaches to the given problem. Instead this chapter introduces the main topics in the EEG brain wave recognition community which are related to this work and show how the recognition of human thoughts was approached.

3.1 Early work

The first work that describes speech in EEG is from 1971. McAdam [24] conducted experiments measuring brain waves while the subject was speaking. His results showed that the recordings of the inferior frontal sites of the left hemisphere (presumably Broca's area) showed larger negative potential than the recordings from the right hemisphere. This was the first evidence for a crude localization of speech production with EEG.

3.2 Brain computer interface

Brain computer interfaces (BCI) should make the control of computers with just the usage of the mind possible. Work in this area is successfully showing that binary decisions are possible to be done with thoughts. The subject have to learn and train particular thinking patterns. The burden is on the side of the subjects rather then on the side of recognizer to discriminate real life thoughts.

There is a distinction between dependent and independent BCIs. A dependent BCI relies on the presentation of a stimulus that activates a brain region. This activation is then detected. An example are the visual evoked potentials. This systems use the visual evoked potential (VEP) recorded from the visual cortex to recognize the direction of an eye gaze.

Middendorf[25] built a device where several buttons on a screen were presented. This buttons were blinking at a different rate. The user selected a button by focusing on it. The device could recognize the choice by measuring the frequency of the photic driven response over the visual cortex. If it matched the frequency of the flashing button then the device selected this button as a hypothesis.

A independent BCI is one which the user can use without an external presentation of a stimulus. An example is the P300 evoked potential.

The BCIs can be divided into four groups based on the electrophysiological signal they use (Figure 3.1 visualizes three of the signal types):

3.2.1 Slow cortical potentials

The slow cortical potentials (SCP) are the slowest which can be recorded by EEG. The potentials are lasting between 300ms and several seconds. There are negative and positive SCPs. People can learn to control the production of them. Birbaumer [1] built a device for locked-in patients where this persons had to learn to control the slow cortical potentials of their electroencephalogram. This enabled the locked-in patients to transmit binary decisions to the computer.

3.2.2 P300 evoked potentials

A subject is presented a large number of frequent events. When one infrequent event occurs then a positive peak can be measured in the brain waves at about 300ms after this event. Farwell and Donchin [26] built a device showing a matrix of letters. Every row and column was flashing in a random order one at a time. The subjects were counting the number of times the desired letter was flashing. The counting of the flashing of the row or column containing the desired letter generated an infrequent event which evoked the P300 potential. The flashing of rows or columns not containing the character on the other hand was the frequent event. The detection works without a long training of the subject.

3.2.3 Mu rhythm

The mu rhythm is the 8-12Hz activity which can be measured at the central sensory motor cortex. Mu rhythms are present when the subject is relaxed. They disappear in the left hemisphere of the brain when body parts on the right side are used and vice versa. It is possible to learn after some weeks of training to control the amplitude of the mu rhythm just by thoughts. Wolpaw and McFarland[27] introduced a system which can recognize the amplitudes of the mu rhythms on both sides of the brain and by that to control the movement of a computer mouse.

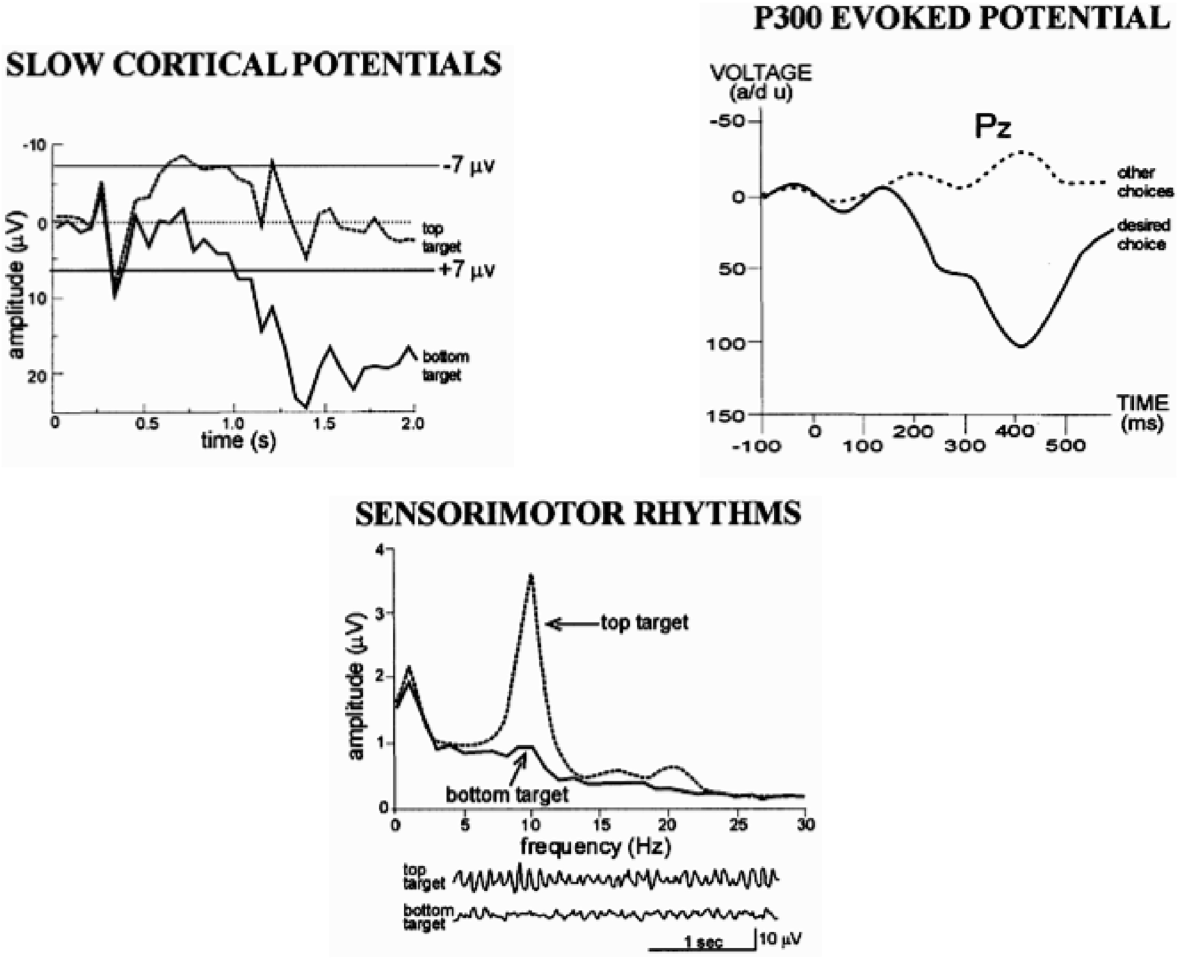


Figure 3.1: (Modified from [7]) (Top left): User learns to move a cursor to the top or the bottom of a target. (Top right) The P300 potential can be seen for the desired choice. (Bottom) The user learns to control the amplitude of the mu rhythm and by that can control if the cursors moves to the top or bottom target. All the signal changes are easy to be discriminated by a computer.

3.2.4 Movement related EEG potentials

Studies show that particular EEG signals can be derived while a subject imagines to move a body part. An example for this approach comes from Dornhege[28] who presents the subjects the letters 'L' and 'R'. The subject images to perform a movement of a finger of the corresponding hand (L=left, R=right). The evaluation shows that the signal for left and right can be discriminated. Also Wentrup[29] uses this approach. The Berlin Brain-

Computer Interface group used an approach where the subject imagines the movement of the whole left or right hand[30].

3.2.5 Discussion

The drawback of this is that the subject needs to train made up thoughts to control the computer. Furthermore mostly binary decisions are possible. Therefore these approaches are more suitable for a command receiving system than for a system which enables people to communicate with a computer via unspoken speech.

3.3 Recognizing presented Stimuli

One group of work investigates the possibility of recognizing stimuli. The task consists of the presentation of a visual or auditory stimulus. While doing so EEG-data is recorded. Later a recognition of what was shown in the EEG-data is tried[31] [32]. This differs from visually evoked potentials because here the stimulus is detected, not the eye gaze.

Suppes et al. [33] presented a system capable to detect from brain waves audible or visual stimuli followed by nothing, spoken or silent speech.

This methods are also used to build functional maps of the brain or to develop theories how the parts of the brain work together.

This approaches help us to understand the brain but are not useful for our communication task.

3.4 State Detection

Singh [34] built a system that recognizes certain mental states such as if eyes are closed or open and if the person has Alzheimer or not. Another work in the mental state detection was done by Honal[9] where six different user states such as reading, listening or resting could be discriminated in brain waves.

3.5 Contribution

This work differs from the described work because it investigates the possibility to recognize unspoken speech out of brain waves. This means that the subject does not have to imagine unnatural things to communicate its commands like moving the left finger. Commands can be uttered in a natural way as they are usually spoken. We do this with an adapted state of the art speech recognizer which is also different from the approaches that the work presented here used for solving this task.

The idea as it is described in section 2.4.4 is a different approach than ones that the presented work took because we try to extract speech out of the moving patterns that the Broca's area generates in all modalities including the unspoken modality. As the experimental results show the primary motor cortex together with the Broca's area and Wernicke's area produce enough collectable information to make an unspoken speech recognition possible that performs as good as if it would be using also the information of all the other non movement related areas where we placed electrodes.

Also the number of detectable states is different. It is increased to 10 different recognizable states for all modalities. This gives the person more options and makes the system more flexible.

Chapter 4

System Overview

This chapter will describe how the data collection was done and how the training of the model and the recognition process were performed.

4.1 Setup

The goal of this work is to show that the recognition of unspoken speech using EEG data is possible. To keep focus on this specific task we had to get rid of as many influences on the recorded subject as possible. Any distractions on the subject such as movements of any body part, pain, additional thoughts or environmental influences could cause artifacts in the EEG signal which would make recognition harder. We tried hard to keep as many artifacts as possible out of the signal. The dispositions we used to reach this goal will be described in the next sections.

4.1.1 Overview of the recording setup

Our recordings were done in quiet rooms during day and night times. The recording setup is shown in the picture 4.1. The picture shows a room at the interAct-labs at Carnegie Mellon University in Pittsburgh, Pennsylvania, USA in which most of the recordings were done. Other locations with a very similar setup were also tried.

The subject was sitting in the chair in the front and the advisor was sitting in the chair at the opposite side of the table. The subject was facing the CRT display and looking at it. The investigator was controlling the recordings on a laptop which was attached to the CRT display.

The subject was told that it can quit the experiment without any consequences at any time. The subject was also allowed to ask for as many breaks as it wanted. During this

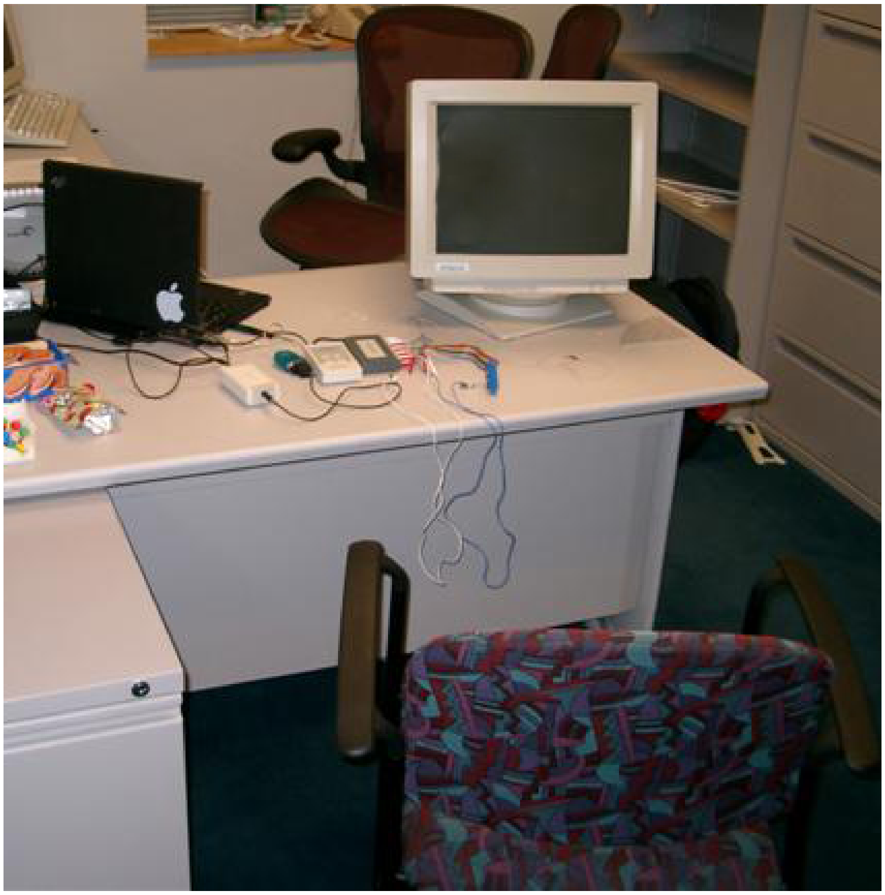


Figure 4.1: recording setup

breaks candies and beverages were provided for the subject. The sessions were continued when all eating and drinking was finished and the subject had enough rest.

The screen showed instructions which the subject had to follow. If the subject did any mistakes then it was asked by the adviser to repeat the current recording. The recording was then deleted and repeated and the subject could also ask the investigator for a repetition of the recording if the subject noticed a mistake.

The recordings were done on the investigator controlled laptop[1] with the "UKA {EEG—EMG} Studio 2.10mwR"[35] software. The software and the modification done for this recording setup are described in Section A.2.

4.1.2 Recording Procedure

EEG recordings differ a lot from other kinds of recordings because of the high impact of artifacts on the recognition. Because of that the subject was not allowed to do any uncontrolled

[1] IBM T40p 1.6GHz, 1GB RAM

motions during the recordings. To avoid this motions we enforced a special procedure during the recording process.

The succeeding steps were followed for the recording of every utterance:

1. The subject sat quietly and without any movement in front of a white screen
2. The instructor started the recording process by pressing a button.
3. The screen showed the words which should be uttered in black letters. In brackets it showed the modality of the utterance.
4. After 1 second the screen showed the words: "inhale and exhale".
5. After 1 second the screen turned black.
6. After 2 seconds the screen turned white.
7. The subject was instructed to wait for about 1 second
8. The subject utters the word which was shown on the screen in step 3.
9. The instructor stopped the recording with the pressing of a button as soon as the subject uttered the words from step 3.

The sequence of screens (Figure 4.2) through the steps 3 to step 7 was chosen to force the subject in a certain rhythm of recording. The result of this rhythm was that the initial situation for every recording of every utterance was always the same. The subject saw the word to utter, inhaled and then exhaled so that it could start the utterance in an exhaled state. In this way we could produce comparable recordings. The appearance of the black screen in step 5 was chosen to not later recognize a picture which the subject might have in its head from the words in step 3.

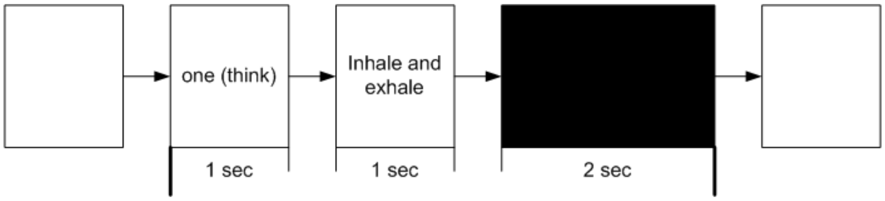


Figure 4.2: The screens showed to the subject before it uttered the word

Obviously the condition for stopping the recording in step 9 was hard to determine by the instructor for the thinking modality. So another procedure had to be used to determine the end of a recording. Several possibilities of showing the end of the thinking are possible. The subject could show the end of its thinking by either stopping the recording by itself or giving the instructor a sign to stop it.

However every voluntary movement would involve the production of additional artifacts. If the subject stops the recording process by itself then it produces artifacts before pressing a button to stop this process. To get a proper alignment there artifacts should be easy to recognize. This is a non trivial task for artifacts which are produced by movements of the lower part of the body e.g. the hands. So a good alignment is harder to find than with an alternative approach which was used in our system.

Eye blinking produces a significant increase of the amplitude in the recorded EEG signal at the Fp1 and Fp2 electrodes. This can be recognized very reliable and an alignment can be found easily. So when the subject finished thinking it did one eye blink. After this blinking the instructor stopped the recording.

4.1.3 Subject

The first source for artifacts is the subject itself. Every movement of the subjects body produces artifacts. We found that the closer the muscle of the movement is located to an electrode the higher the amplitude of the measured artifact on this electrode is.

The subject was therefore told not to move any part of the body during the production of the utterance. The eyes should be open all the time and focusing a point in front of the subject on the screen. Any eye blinking was not allowed during the modalities which involved facial movement and during the thinking process.

During the recording process before the actual recording as described in the steps 1 to 7 in section 4.1.2 the subject was asked to stay in a neutral position as described in [24]. This neutral position means that the subject should stay in a relaxed but immobile posture, the eyes should fix the screen, the lips should stay together and the tongue should rest on the floor of the the subjects mouth. The subject should not move the eyes, swallow or do movements of the head, the limbs or the trunk. The word production should be as fast and accurate as possible. This rules were just strictly enforced during and after step 5, since this was the phase which could impact the recording.

The subjects were all from Germany and none of them was a native English speaker. All were graduate students. All the subjects were capable of completing the recording task and did not use any medication. Table 4.1 summarizes the subjects data which was relevant for the experiments.

Speaker ID	age	sex	recorded utterances	minutes
S1	25	male	5345	772.76
S2	24	male	250	25.78
S3	24	male	250	27.36
S4	25	female	250	27.85
S5	27	male	250	22.73
S6	23	female	1256	167.9

Table 4.1: subjects (a more detailed view of the statistical data is given in appendix B)

4.1.4 Hardware Setup

To capture the signal from the scalp we used two caps (figure 4.4 from Electro-Cap International, Inc). They differed in size only. For electrode positioning on the cap the 10-20 system was used. The caps are equipped with 20 Ag/AgCL electrodes. Because we had an amplifier with 16 channels we used 16 electrodes of the cap simultaneously. These are Fp1, Fp2, F2, F3, F4, F7, C3, C4, Cz, T3, T4, T5, T6, P3, P4, Pz as shown in figure 4.3.

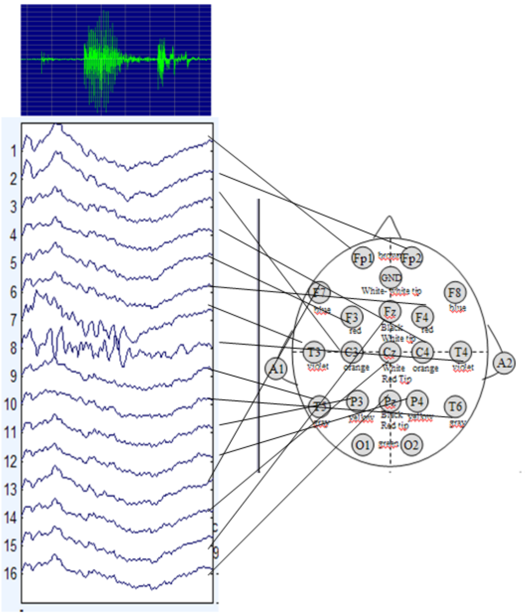


Figure 4.3: This figure shows a sample recording of a subject uttering "eight" in the speech modality. The signal at the top is the waveform of the audio recording simultaneously. The head on the right shows which channels are connected to which electrodes. A1 and A2 are the reference electrodes.

We left out the electrodes O1 and O2 which cover the optical cortical regions since we do not focus on visual information. We also had to leave out one more and decided for F8 because speech is considered to take place mostly on the left side of the brain [36] and the

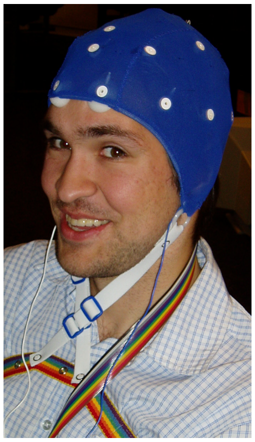


Figure 4.4: subject with Electro-Cap cap

front of the frontal cortex is also not considered to have anything to do with speech.

We used the VarioPort^{TM} [Becker 2003] (figure 4.5) as the amplifier and recorder to amplify and digitalize the captured signal. The specifications of the amplifier are collected in table 4.2. All recordings were done with a sampling rate of 300 Hz.

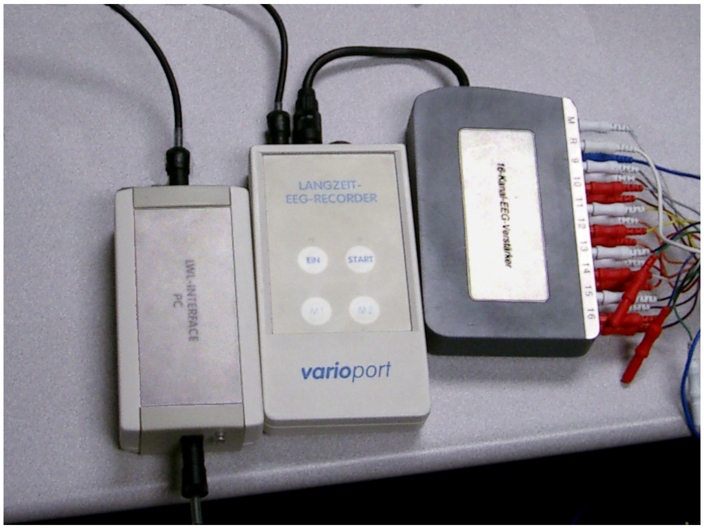


Figure 4.5: From left to right: optical waveguide, computer interface, amplifier

The amplifier was connected to the computer through an interface and an optical waveguide which was connected to a RS232 port which itself was connected through an USB-adapter to a computer. The computer was equipped with an Pentium M 1.6 GHz processor and with 1GByte RAM. All recordings were done under Windows XP.

For the non thinking modalities we also recorded sound files with an sampling rate of 16 KHz. For that we used a close throat microphone (ISOMAX E6 directional microphone).

Amplification factor	2775
Input Range	$\pm 450\mu V$
A/D conversion	12 Bit (4096 steps)
Resolution	0,22 V / Bit
Frequency Range	0,9 ... 60 Hz

Table 4.2: Technical specification of the amplifier used for the recordings [9]

This was optimal because it could fit under the cap behind the left ear and did not apply any additional physical pressure on the subjects scalp.

As monitor we used a 17" CRT with 1024x768 resolution and a horizontal frequency of 75 Hz.

4.2 Training

After the recordings were conducted a training on the data had to be done. The speech recognition system Janus (see section 2.1 for details) was used for this. Janus was run on the condor-cluster at the Carnegie Mellon University InterAct-Labs. A framework was developed based on a state of the art ASR to train a model based on the recorded data. It was also used to evaluate the trained model.

A detailed description of the usage is given in A.1. A brief description of the theoretical background is given in 2.1.

4.3 Recognition

The recognition could be done offline for the purpose of testing our recorded data and our recognition system. But it could also be done online, in realtime to do a demo recording as described in section 7.

4.3.1 Offline

The recognition offline is done when an evaluation of the recognizer is needed. A set of recordings is selected for the recognition using the leave one out cross validation approach. The system is always trained for one session in one modality. For this the data is divided into two groups. For the evaluation set one utterance of every word of the vocabulary is selected. The remaining utterances are used for the training.

The evaluation of every utterance of the evaluation set is done by the computation of the Viterby score for every utterance with the selection of the word with the best score as the hypothesis. The word accuracy is computed from this results. This process is repeated until every utterance was once in the evaluation group. The resulting average of the word accuracies is taken as the resulting word accuracies presented in this work.

4.3.2 Online

The online recognition is needed for the demo system. First a set of training data is recorded. Then the recognizer is trained based on these data. In the online recognition the evaluation set comes straight from the online recordings. After segmentation the hypothesis is built as in the offline system.

Chapter 5

Data Collection

Data was collected in different modalities and with diverse vocabularies. In this chapter the corpora used for this work and the modalities are described. All data were recorded in English.

5.1 Corpora

We used different vocabularies in our data collection. A set of this words which we used in a recording session is going to be called corpus. We used several corpora to show that the speaker is not adapting to a particular corpus. The vocabularies of the corpora are shown in table 5.1. Since we used a full word model for our recognizer the sentences of the *lecture* corpus can be seen as one word.

Name	**Vocabulary**
digit	one, two, three, four, five, six, seven, eight, nine, zero
digit5	one, two, three, four, five
lecture	good afternoon ladies and gentlemen, welcome to the interact center my name is marek wester, let me introduce our new prototype, thank you for your attention, any questions
alpha	alpha, bravo, charlie, delta, echo
identifier gre	brittle, cordial, diffidence, regicide, profundity, presage, nonplused, insipid, fluster, tepid
phone	yes, no, accept, deny, wait
player	start, back, next, louder, turn down

Table 5.1: Corpora used during the data collection. The table shows the name which is used as an identification to refer to the corpus

5.1.1 Digit and Digit5 corpora

The *digit* corpus represents the English numbers from zero to nine. We used this corpus because numbers are universally composable and the size is limited to ten and so the training of a system does not take a long time. The longer a training takes the less comfortable it is for the subject. The situation can get even painful since the cap is very tight. More information about the problems we had with the cap can be found in section 2.5.

The *digit5* corpus consists of the numbers for one to five. It was used for the training of our demo system. To train a model we needed training data. Since session independence could not be shown (as described in section 6.4) we had to do a data collection preceding every demo recording. To save some time and to assure the well being of the subject we used the digit5 corpus.

The digit corpus makes our results easier to compare with other work in this field because it is frequently used in the EEG-community.

5.1.2 Lecture Corpus

The lecture corpus was used to see how good our recognition system can recognize sentences with the full word model. The corpus consists of sentences used during the demonstration of an EMG system at several press conferences of the interAct labs. Using the same sentences would allow for comparison.

5.1.3 Alpha Corpus

The *alpha* corpus consists of the words alpha, bravo, charlie, delta, echo. These words are used by the International Civil Aviation Organization (ICAO) as spelling alphabet. The words are chosen to be easy to distinguish. We wanted to have an easily distinguishable vocabulary to check if the methods we developed improve when making the recognition task easier. The alphabet can also be used universally and allows the comparison with results from other groups. The number of words is constricted to five for fast turn-around time in our experiments.

5.1.4 Gre Corpus

In order to avoid that the subject get used to the vocabulary of our corpora and make sure that the subject does not picture the words in from of imaginary images and to eliminate the resulting artifacts we introduced the GRE corpus that contained words that are rarely used in English language. They were selected from the Graduate Record Examination (GRE) which is a standardized test that many colleges in the USA require their students to do. The GRE includes a test of vocabulary knowledge which was our source to randomly pick ten words (brittle, cordial, diffidence, regicide, profundity, presage, nonplused, insipid, fluster, tepid). These word were shown to the subject some minutes before the data collection and just the pronunciation was explained.

The GRE corpus makes the adaption to words unlikely and since the semantics of these words was not revealed the subject could not picture them.

5.1.5 Phone Corpus

The *phone* corpus consists of the words yes, no, accept, deny, wait. It can be used to answer or reject phone calls. Since we recorded just two sessions with the *phone* corpus it can be seen as a further proof of concept that our recognizer is able to recognize a variety of words.

5.1.6 Player

The *player* corpus consists of commands (start, back, next, louder, turn down) to control an mp3 player. It was designed to be used during a demo which was not further developed due to a problem that came from the fact that the music that was played to the subject was recognized instead of the commands that the subject thought.

5.2 Modalities

We did most of our recordings in 5 different modalities. These are normal speech, whispered speech, silent speech, mumbled speech and unspoken speech. With recording this modalities we could test our recognition system under different circumstances of speech production in this modalities. These modalities create a sort of a continuous degeneration of speech. This degeneration works on two levels.

The first level is the acoustic level. While normal speech can be seen as the optimal way to utter words so that they are most easy to recognize, it gets harder with whispered speech and impossible with silent, mumbled and unspoken speech.

The second level is the movement level. With the normal speech modality the movement of the speech related muscles is very easy. Because of the vanishing feedback in whispered speech and the missing feedback in silent speech it gets harder, as the subjects reported, to move the muscles necessary for correct speech production. Mumbled speech was considered as the hardest to utter by the subjects because the lips were closed in this form of speech.

During all recordings of all modalities the subjects were asked to think the words they utter very clearly. And not to think of anything else. So that they could develop a feeling for how to think a word that they uttered. This was a preparation for the unspoken speech modality where they had to only think the word.

5.2.1 Normal Speech

The subject was asked to pronounce the word as naturally and clearly as possible in an audible fashion. Later we asked the subject to utter words in the *digit* corpus using phonetic knowledge.

5.2.2 Whispered Speech

To utter words in the whisper modality the subject was asked to whisper the words naturally and clearly so that they were barely audible. No special restrictions were made in this modality.

5.2.3 Silent Speech

The modality of silent speech was defined as natural speech production without the production of any sound. The silent speech modality was felt hard to utter by the subjects.

5.2.4 Mumbled Speech

The mumble modality was defined as natural speech production without opening the lips and producing any sound. This was the most degenerated kind of speech. It was felt as the hardest to utter by the subjects.

5.2.5 Unspoken Speech

In the unspoken speech modality the subjects were asked to think the word loud and clearly as if they were uttering the word in the normal speech modality. To think a word ”loud”, focused and clearly means that they should not think of anything else. They should think the word in the same way as they did in the normal speech, whispered speech, silent speech and mumbled speech modality. They were also asked to think nothing before the thinking and after the thinking of the word.

Chapter 6

Experiments

This chapter presents the results of our experiments and the way we developed our recognition system.

For all experiments the evaluation method as explained in 4.3.1 is used. The results of the crossvalidation are presented as word accuracy.

To refer to the different sessions the following notation is used:

> subject-session-modality/repetitions → [0-9][0-9]-[0-9][0-9]-[nwsmu]+/[0-9]+
>
> e.g. 02-05-wu/20 refers to a session recorded with subject 02 in the subjects session 05 with 20 repetitions in the whispered speech modality per word and 20 repetitions in the unspoken speech modality per word.

Every time it is referred to significance in the description of the experiments then the t-student-test was used to determine this. A result of this test is considered to be significant if the error probability is $\leq 5\%$.

In the first part of this chapter a description of how we found the parameters for the feature extraction and normalization is shown. In the next section the results for the first subgoal of this, as defined in the introduction, is shown. In the next two sections the problems with speaker and session dependency are discussed. The results for the second subgoal of this work are discussed in section 6.5. The following two sections discuss the recognition of sentences and unknown words. The final section in this chapter presents the results for the third and last subgoal of this work.

6.1 Feature Extraction and Normalization

This section contains a description of how the parameters of the recognizer influence the recognition results. The result of this investigation should be no optimal system since an optimization done on the compared to speech recognition small amount of data would be just an optimization of the system on these specific data and not an optimization of the task of recognizing speech in brain waves. So the result of this experiments should be to get knowledge about which parameter can be a good lever for getting better recognition results and which parameters influence the system most in which modality and what values work best for this data.

A baseline system was used for this investigation. Within this system one parameter was varied at a time so that the influence of the parameter can be seen. The following parameters were investigated (in brackets are the values of the baseline system which were gathered empirically in the course of developing the recognizer):

- The first parameters which were investigated concern the data processing which is the process of transforming the brain waves info feature vectors
  - Window Size of the STFT (26.6ms)
  - Windows Shift of the STFT (4ms)
  - the extracted feature (stft and delta delta)
  - use LDA or not (use LDA)
  - number of dimensions of the feature vector used for the training (35 dimensions)
- The second group of parameters concerned the HMM which is the classifier we used for the recognizer
  - number of gaussians for every state of the HMM (25 gaussians)
  - number of states of the HMM (5 states)

The experiments were conducted with the sessions 01-02/5, 01-04/5, 01-05/5, 01-06/5. The following figures in this section are going to show the mean of the word accuracy of four different sessions on which the experiments were conducted in the five modalities for a better overview. A detailed listing of the results can be found in appendix C.

First the window size was investigated. As figure 6.1 shows, the window size of 106.6ms performs worst. The sizes 26.6ms and 53.3ms show no significant difference for the modalities involving muscle movement. Just the unspoken speech modality shows a large improvement of 10.5 % points when using a window size of 53.3ms.

The next investigation concerns the window shift for the STFT. Here the results in figure 6.2 show very clearly that a window shift of 4ms has the best results through all modalities.

Now that we have a feeling of how big the windows have to be and in what shift they have to move we can investigat the influence of the different features next. The investigation

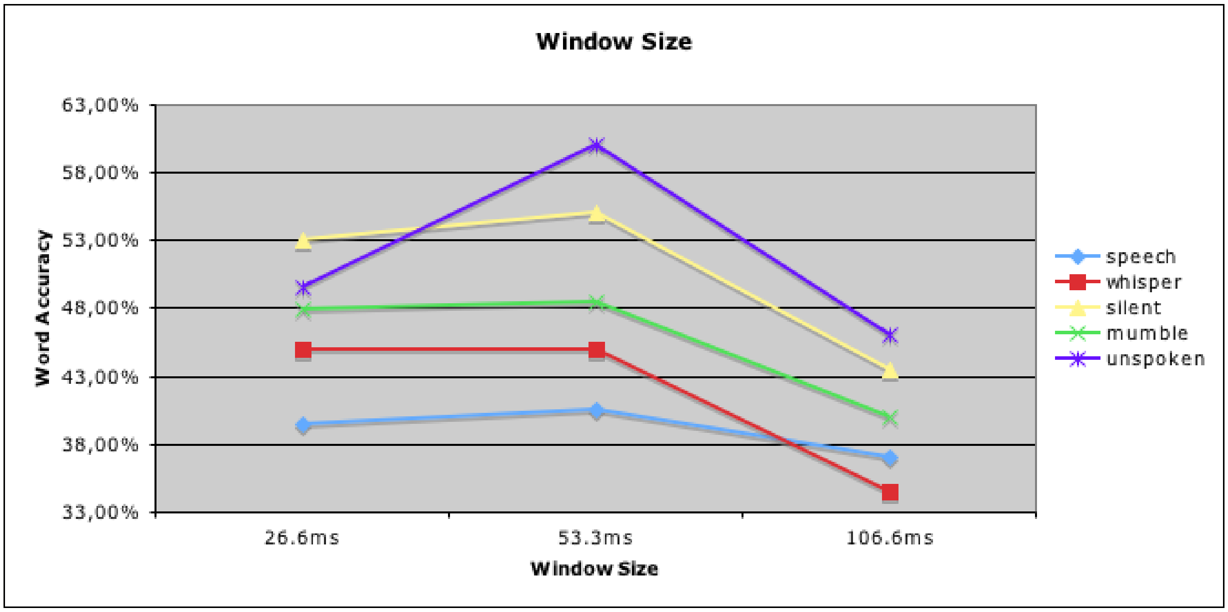


Figure 6.1: The window size of 53.3ms is better for unspoken speech

started with the STFT which was also included in all of the following experiments concerning the features. Then the delta, delta delta and delta delta delta was varied. The last experiment was done with STFT, delta delta and delta mean.

The results in figure 6.3 show that the features are dependent on the modality. The speech and whisper modality which involve the production of audible speech perform best when just using the STFT. This also means that delta features do not help to discriminate speech in these modalities.

For the silent modality delta works best. But a double delta and just the STFT are also not significantly worse. The mumble and the unspoken speech gain a lot from the delta features. Both perform best with the delta mean feature. But there is also a significant peak for the delta delta feature. The gain from the features for mumbled speech is 10% points and for unspoken speech 15% points. This shows that this parameter is a good lever for the mumbled and unspoken speech modality.

Now that features are selected we have to investigate if there is any gain if we cut off dimensions which make the training task hard due to the lack of more data. This can be done with the LDA whose basic idea is that the dimensionality of a feature is reduced while discriminative information is preserved as good as possible. The results in figure 6.4 show that the usage of the LDA has an significant improvement of 31.40 % points. The mode where no LDA was used seems to be undertrained which is normal in this situation because we have a high dimensionality of 192 dimension but not much data. So using the LDA is a

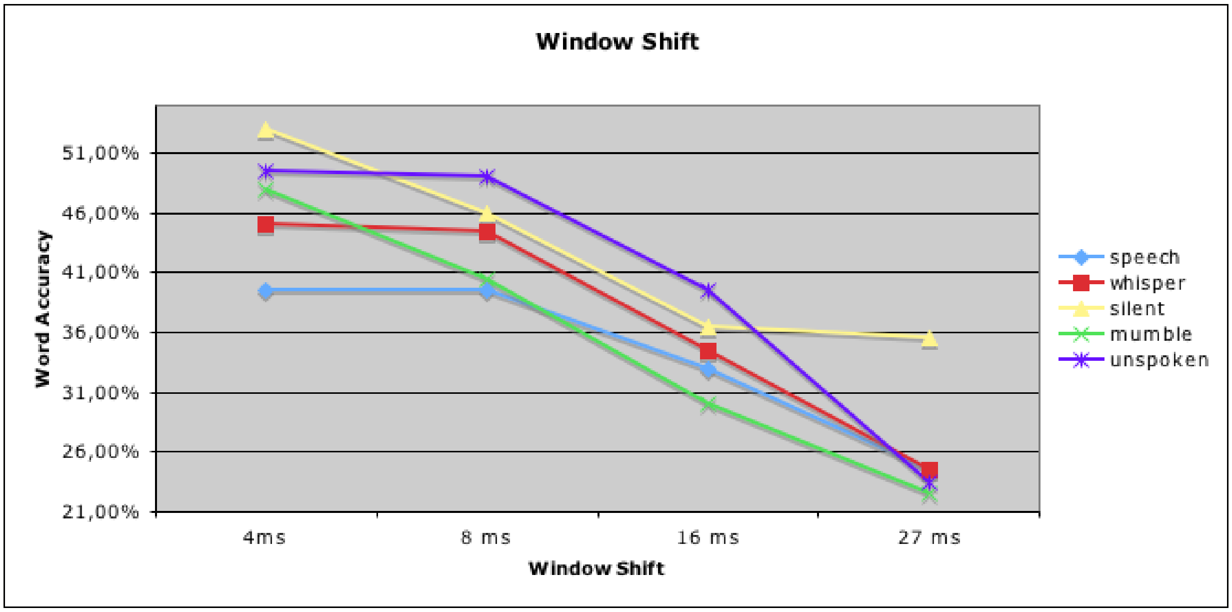


Figure 6.2: A window shift of 4ms is ideal

good lever to improve the recognition system.

The next question to answer is how many dimensions should be kept after the LDA. In this investigation the number of dimensions of the feature vector after performing the dimensionality reduction was varied.

A dimensionality of 16 dimensions shows the best results for all modalities besides the mumble modality (see figure 6.5). For the mumble modality 8 dimensions are optimal. The whisper modality is also not much worse with 8 dimensions. Since the range can be 14 % points this parameter is also important in building the recognizer.

The next investigation was conducted to see the influence of the number of gaussian mixture models (gaussians) used for every state in the HMM.

For the speech and silent speech modality best results can be gained by using 4 or 16 gaussians (figure 6.6). Best results for the whisper and unspoken speech modality can be gained with 4 and 32 gaussians.Ffor the mumbled speech modality there is no value for the parameter which provides the significantly best result. The numerically best results can be gained with 25 gaussians.

The last investigation concerned the number of states in the HMM. As figure 6.7 shows this parameter has the maximum distribution of optimal values over the parameter value. Just the normal speech and silent speech share 6 as the optimal number of states. For the unspoken modality 3 states work best while 5 states are best for the mumbled speech modality and 7 states for the whispered modality. The gain with this parameter is up to 8%

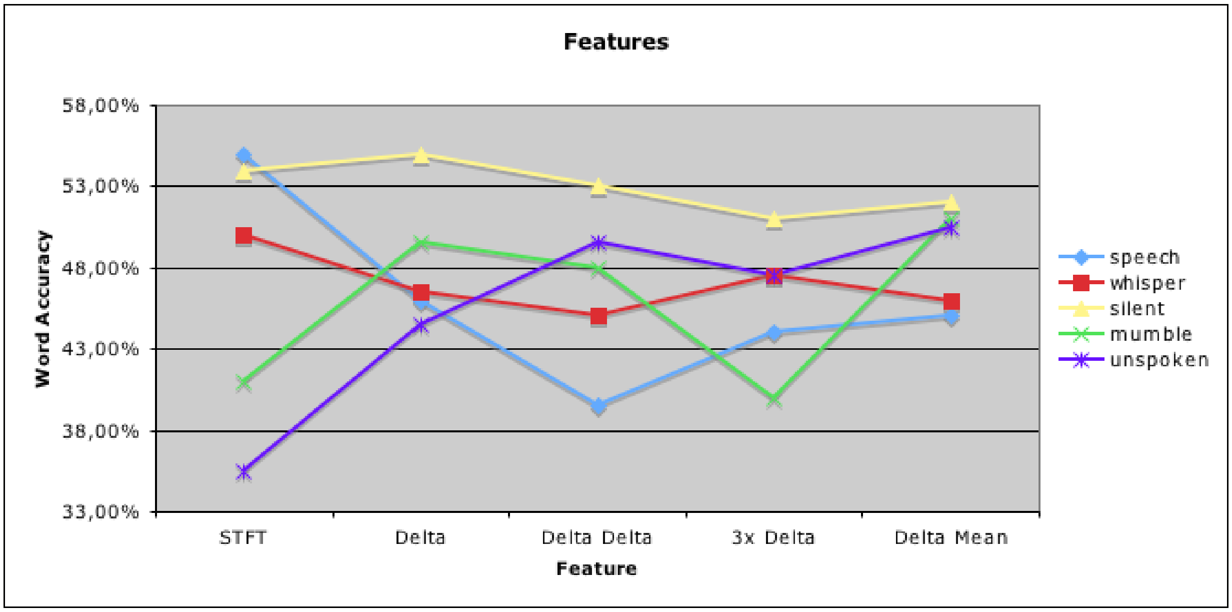


Figure 6.3: delta features increase the recognition of unspoken speech

points for the mumbled speech modality so this parameter can be seen as a lever to optimize a recognizer. But it is hard to find an optimal value for a large number of modalities.

As a conclusion we can state that using the LDA is the best lever for all modalities. It is also clear that a window shift of 4ms performs best and also a window size of 53.3ms seems to be optimal for all modalities.

It gets harder to state which features are best. The mumbled and the unspoken modality clearly gain from delta features, while this is not true for the normal speech, whispered speech and silent speech modality. The number of dimensions after the LDA also falls in the group of parameters with two best parameter values. The mumbled speech modality performs best with 8 dimensions while the other modalities perform best with 16 dimensions. There is a large gain in this parameter of up to 14.5 % points.

The parameters concerning the HMM are mostly modality dependent. A general conclusion can only be, that finding the optimal parameter in the HMMs can be hard but the gain can be high.

6.2 Recognition of Normal speech

The first subgoal was to see if it is possible to recognize normal speech in EEG-data. For this data was collected as described in section 4.1.2. The recognizer was trained with the recorded data and the evaluation of the recognizer showed results as shown in the confusion

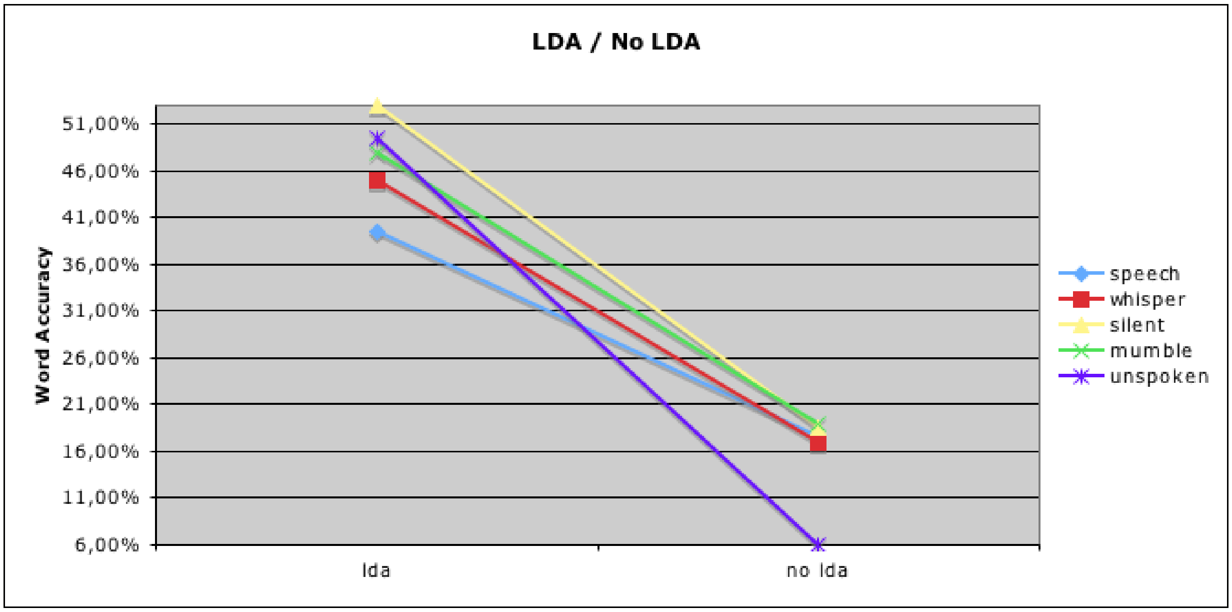


Figure 6.4: lda is very important for the current recognizer

matrix 6.1.

						hypothesis						
		one	two	three	four	five	six	seven	eight	nine	zero	word accuracy
reference	one	17		1		2	3		2			68%
	two		20	1		1					3	80%
	three	2		17	1					2	3	68%
	four	4	5	5	5	1	1	1	2		1	20%
	five	1	1			13	1	1		4	4	52%
	six					1	21	3				84%
	seven		2				2	17		4		68%
	eight			4		1	2	5	11	2		44%
	nine		1			1	1	2	1	18	1	72%
	zero		5		2						18	72%
												62.8%

Table 6.1: confusion matrix for results of session 01-07-n/25

The worst result can be seen for the word ”four”. The production of this word involves not much facial movement. Therefore not much EEG-data is produced in the homunculus area which can be the reason for the worse recognition. Another reason may be bad recordings for this word.

Chart 6.8 shows the results for the recognition of speech in different sessions with the digit corpus. The results do not significantly differ besides session 6 and session 7. The bad result

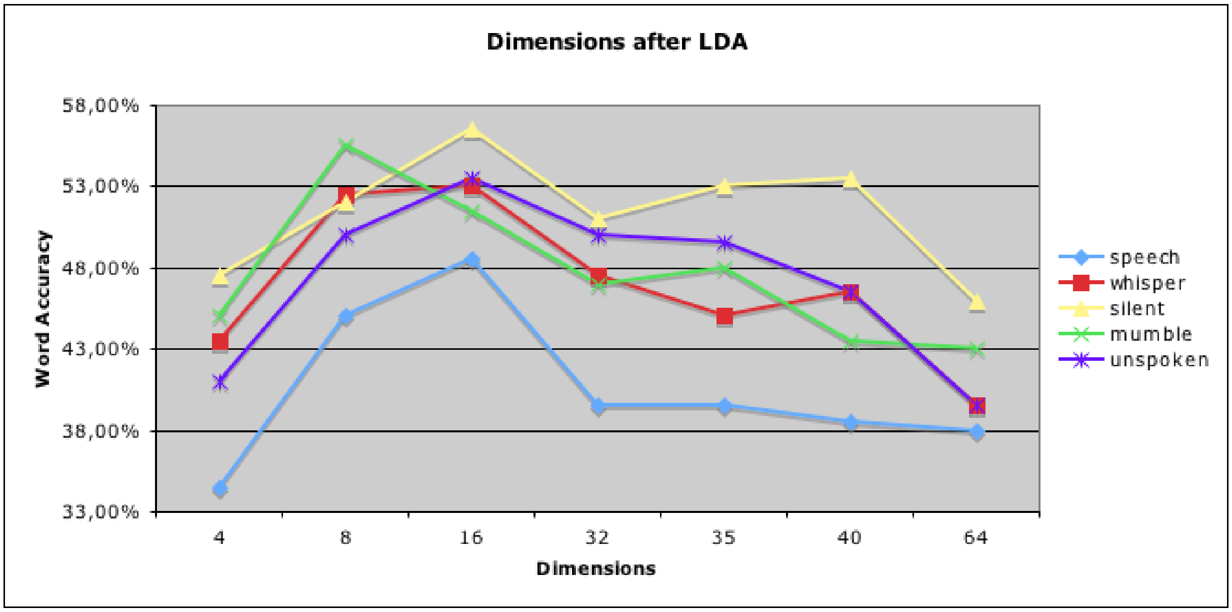


Figure 6.5: up to 35 coefficients are best for the recognizer after the dimensionality reduction was done

in session 6 results from not well articulated words. The speaker was not focused enough during the recording of this session. This shows how important well done recordings are.

The result in table 6.1 and the results in chart 6.8 show that recognition of speech in EEG-data is possible. The achieved results are about 5 times higher than chance so we can say with more likelihood that goal 1 is reached.

6.3 Variation between Speakers and Speaker Dependancy

Unspoken speech is a kind of thinking. Every person speaks different in the persons mind. There is no notation of a phonetic alphabet for unspoken speech. No subunit of a thought word is known which is constant between different people. In this section an investigation of the speaker dependency of the recognition system was conducted.

To test if a system is speaker dependent we trained the recognition system with the data of one speaker and recognized a session of another speaker which was recorded with the same corpus and the same number of repetitions under the same conditions. The results across speakers are significantly worse than within speakers. Table 6.2 shows the results of an experiment where a session in the modalities normal speech, silent speech and unspoken speech using the digit corpus was trained with subject 1 and then evaluated on comparable

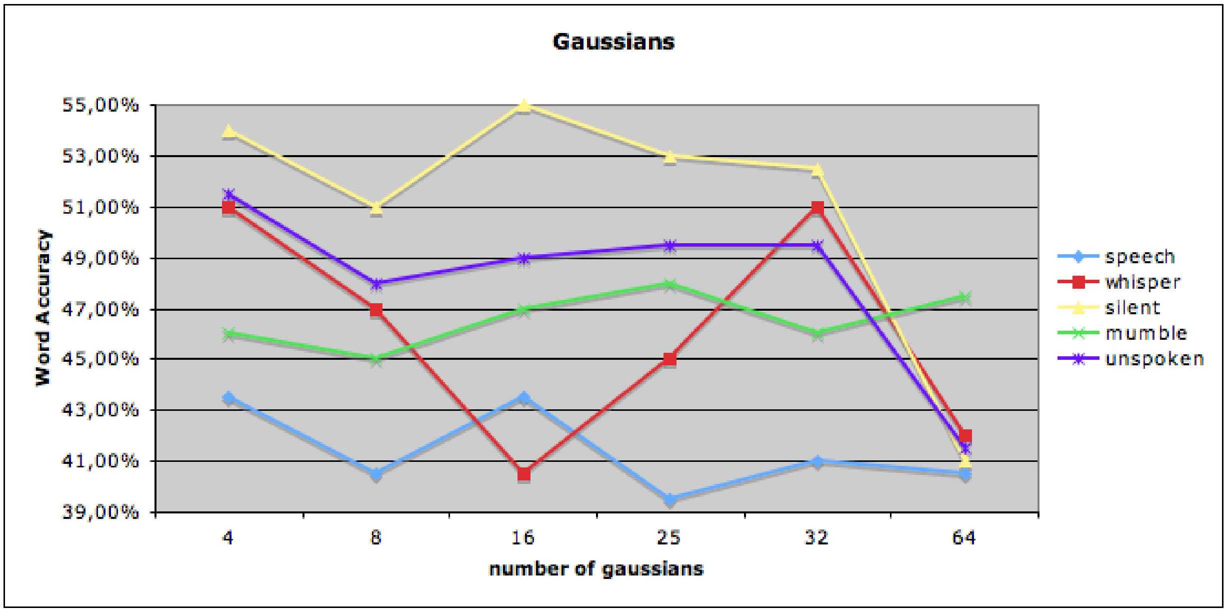


Figure 6.6: No significant difference can be seen for up to 32 gaussians. 64 gaussians are too much.

data of subject 6 and vice versa. The results show that the recognition rate is not significantly different from chance. This showed that the system is very speaker dependent.

evaluation session

model session	06-06-n/10	01-11-n/10
06-06-n/10	92%	11%
01-11-n/10	9%	99%
	06-06-s/10	01-11-s/10
06-06-s/10	100%	11%
01-11-s/10	7%	91%
	06-06-u/10	01-11-u/10
06-06-u/10	98%	10%
01-11-u/10	10%	96%

Table 6.2: Results of the experiment with the digit corpus show high speaker depedency

This is due to the fact that the brain waves that can be measured while speech is produced seem to be very different between every person. The first problem is that to get stable results in the recognition some training in producing constantly the same speech is needed. The subjects need to be instructed very carefully. During the experiment the subject has to be very focused on the task of clear production of speech. Interruptions because of technical

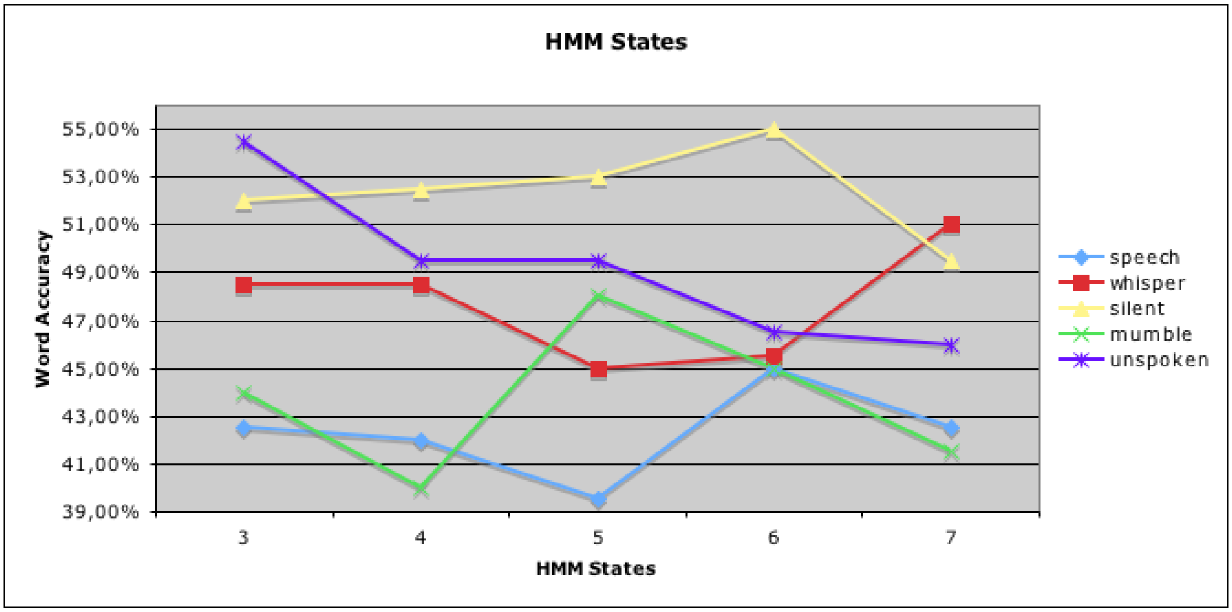


Figure 6.7: no significant difference in the overall performance but unspoken speech seems to do best with 3 states

recording issues or through a noisy environment are borne different by every subject. This results in smaller or higher artifacts. Figure 6.9 shows the result word accuracy for five different subjects. Since for subject 6 no comparable session was recorded the results of this subject are not in the chart. The digit corpus was used in this sessions[1]. Large variations can be seen in word accuracy between the subjects and within the same subject and the different modalities.

For most of the other experiments subject 1 was used to get results which are better comparable. This subject also turned out to produce recordings which could be better recognized.

A larger amount of data was also collected with subject 6. The results through different sessions show that the results are worse compared to subject 1 who had more training. Table 6.3 shows that the results in the different comparable sessions are sometimes significantly worse and sometimes comparable. In numbers the results of subject 1 are always better. Due to this all other results presented here are from recordings of subject 1.

[1] 02-01-nwsmu/5, 03-01-nwsmu/5, 04-01-nwsmu/5, 05-01-nwsmu/5

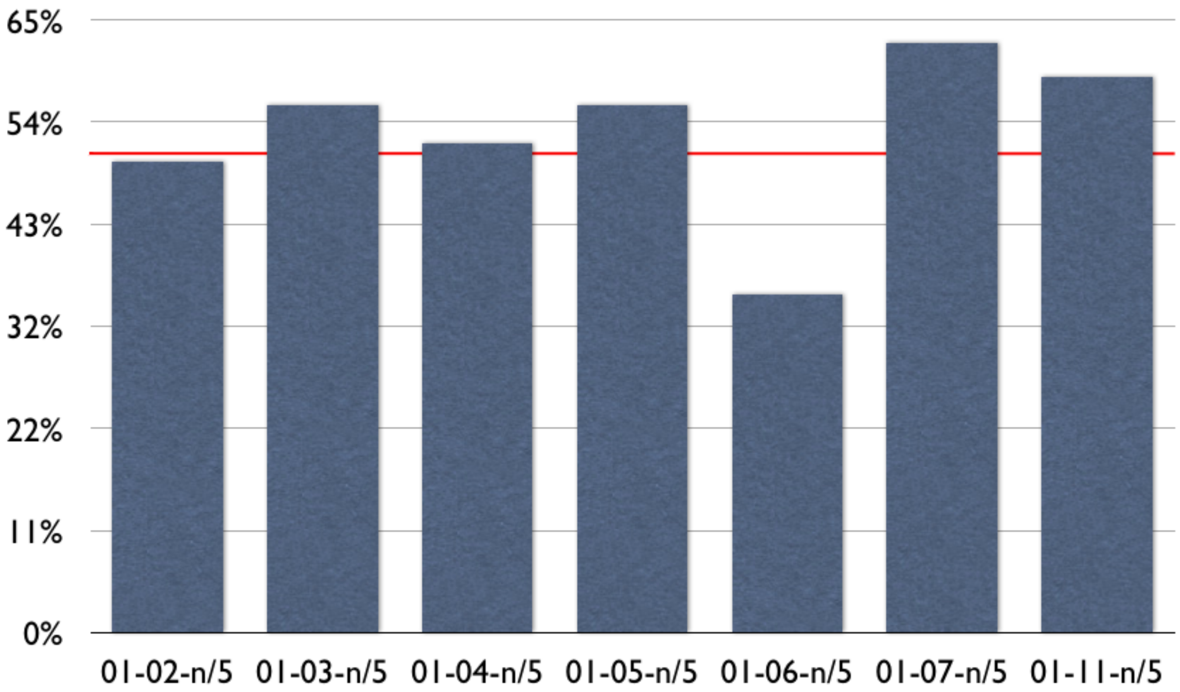


Figure 6.8: Word accuracy for the digit corpus in different sessions with normal speech modality. The red line shows the average.

6.4 Variation between Sessions and Session Dependancy

In normal speech recognition a recognizer can be trained with the recordings of one session and can then recognize other untrained recordings. To test if this is also possible for the recognition of unspoken speech we tested this with training the recognizer with the recordings of one session. Then we tried to recognize recordings of another session with this recognizer. The results in word accuracy were worse than chance. Even feature adaption such as MLLR did not give significant results.

Variations between sessions are due to the different recordings conditions and more importantly different mind states of the recorded subject.

6.5 Modalities

The results for the investigation of the second subgoal are presented in this section. It should be investigated how well the developed recognizer works for different modalities: normally spoken speech, whispered speech, silent speech, mumbled speech and unspoken speech.

The results for the different modalities are shown in chart 6.10. In every session of this chart all 5 modalities were recorded with the digit corpus. Five examples were recorded for

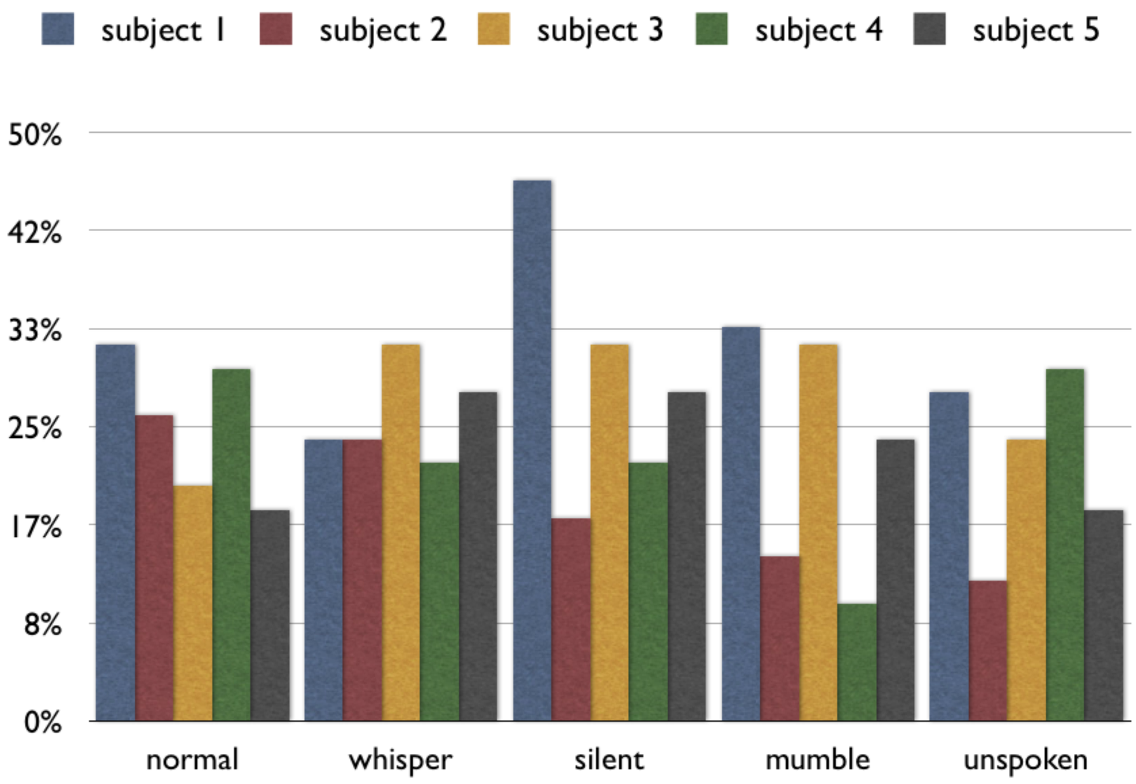


Figure 6.9: word accuracy for different subjects

every word in every modality. This results in 250 recordings per session. All five session were recorded with the same speaker in the same recording setting. This five sessions were chosen as examples for the other sessions recorded and tested which performed comparably.

The speech modality has an average word accuracy for the five session of 50 %. This is five times higher than chance which is 10 %. The other modalities which involve muscle movement are in average not significantly different[2]. Some sessions like e.g. session 2 in mumble modality show worse results. This can be explained with bad recordings. The subject may not have uttered the words correctly, the environment produced noise or problems with the cap led to worse data.

For the unspoken speech which involved no muscle movement the results were slightly worse. But a significant difference could not be shown. In average this results are comparable with the results from the other modalities.

The second subgoal is therefore reached. In average we get a word accuracy rate that is four to five times higher than chance.

[2] The t-student test was performed

		word accuracy		
domain	session id	normal	silent	unspoken
digit	1-11-nsu/10	59,0%	63,8%	35,0%
	6-06-nsu/10	42,0%	51,0%	31,0%
phone	1-14-u/20			42,0%
	6-05-u/20			38,0%
digit	1-09-u/25			45,0%
	6-01-u/25			29,6%
	6-03-u/25			33,7%

Table 6.3: comparison of the word accuracy for subject 1 and subject 6 for different sessions with different modalities and different corpora.

6.6 Recognition of sentences

We investigated how well sentences are recognized with the recognizer. For the investigation every sentence was modeled as word in the recognition framework. The test was done in three sessions[3] with the lecture corpus with two different subjects.

session	repetitions	modality	word accuracy	$\frac{\textit{word accuracy}}{\textit{chance}}$
06-04	20	unspoken	42.5 %	2.1
01-10	25	unspoken	56 %	2.8
01-12	15	normal	67.7 %	3.3
01-12	15	silent	84 %	4.2
01-12	15	unspoken	67.7 %	3.3

Table 6.4: Results for the recognition of sentences

The results from table 6.4 show a high word accuracy compared to the word accuracy for the single word recognition. The reason is that the number of sentences is five and not ten like in the digit corpus. The sentences are also longer and therefore less confusable. Therefore the probability to choose a word by chance is 16.66% compared to 10%. The last column still shows a word accuracy of four to five times higher than chance for the session 01-12.

6.7 Meaningless Words

It would be possible that not the uttered word is recognized but an image of the word that is produced in the mind. Therefore we investigated if the recognition also works for words which have no meaning to the subject. Ten words which were meaningless to subject 1 were randomly chosen by the recording assistant out of the Graduate Record Examinations which

[3] The sessions are 06-04-t/20, 01-10-t/25 and 01-12-nsu/15

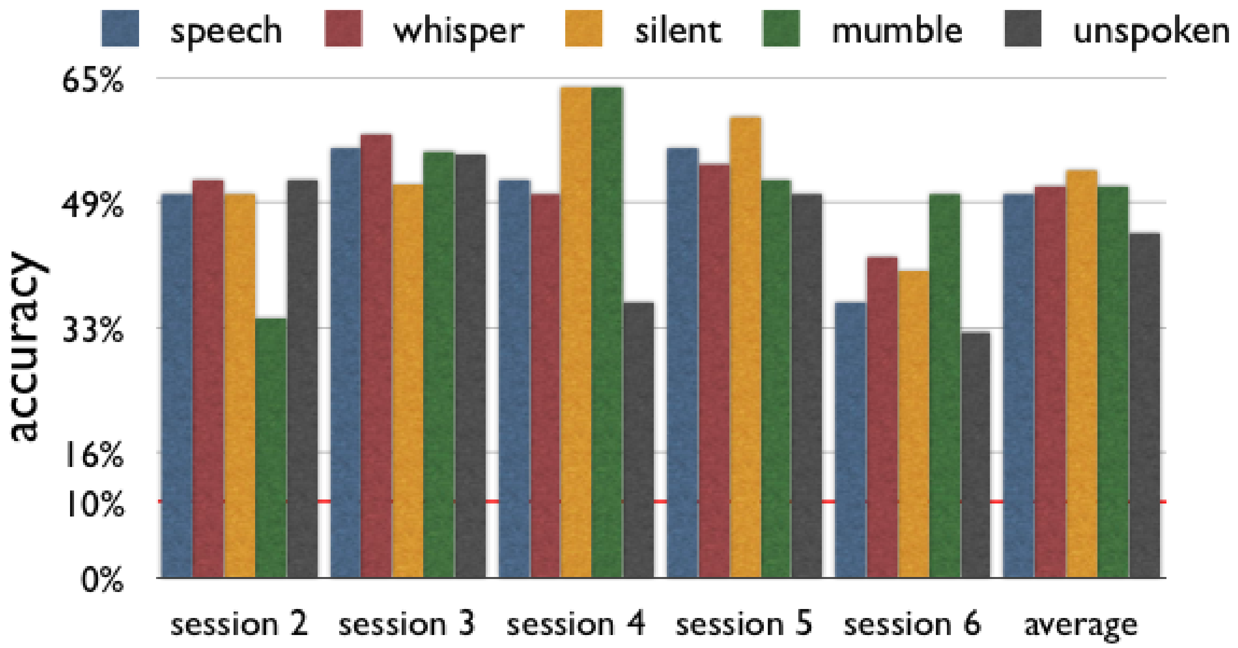


Figure 6.10: Results of the different modalities

resulted in the gre corpus as described in section 5.1.4. Meaningless means here that all subjects are non native English speakers. Therefore this rarely used words are not know to the subject and because of that they have no meaning to the subject.

Because the words had no meaning to the subject only one session was recorded and evaluated. The words of this corpus were not know by the subject and the subject could not derive them from known words.

Table 6.5 shows a confusion matrix of the evaluation result. The word accuracy was 38.50% which is approximately four times higher than chance. This result could be seen as an indication that the detection is based on the brain waves resulting from producing the speech rather than imaging a picture.

6.8 Electrode Positioning

The third subgoal was to investigate which electrode positions are most important for the recognition of unspoken speech. Session 01-24-u/30 in the unspoken speech modality using the digit corpus was chosen to investigate this. The result for the evaluation with all electrodes is a word accuracy of 47.24%.

Training and evaluation experiments were performed, in which we left out particular electrodes in order to see how important the corresponding channel information is. The

		hypothesis									
		brittle	cordial	diffide.	fluster	insip.	nonp.	pres.	profu.	reg.	tepid
reference	brittle	5	3	3		1	5	1	1		1
	cordial	4	6	1				8		1	
	diffidence	7	2	7			2	2			
	fluster	3	2	5	1	1	4	1	2		1
	insipid	1				16	3				
	nonplused					2	8	6			4
	presage		1				2	16	1		
	profundity					1		7	8	4	
	regicide		2					7	6	5	
	tepid	1			1		8	4	1		5

Table 6.5: Confusion matrix for the recognition of unknown words shows a word accuracy of 38.50%. The rows are the expected words while the columns are the predicted words.

evaluation criteria is word accuracy measured on unspoken speech of session 01-24-u/30. The left out electrodes were chosen to be left out because the region around the electrodes T3, C3, Cz, C4, T4 seems to be most promising to detect unspoken speech because the homunculus is located there. Also the electrode F7 where the Broca's area is located and electrode T5 where the Wernicke's area is located seem to be interesting.

The electrodes in the back P3, Pz, P4 and T6 were left out first. Then we left out the electrodes in the front: Fp1, Fp2, F3, Fz, F4. The result for the word accuracy in figure 6.11 shows no significant difference to the result with all electrodes. This indicates that the electrodes in the front and in the back do not to provide information that help in recognition of unspoken speech.

In the next step we left out the electrodes in the front and the back, namely P3, Pz, P4, T6, Fp1, Fp2, F3, Fz, F4. The result regarding the word accuracy is shown in figure 6.11 in the bottom right. The word accuracy does not differ significantly from the word accuracy with all electrodes. This indicates that the remaining electrodes are sufficient to recognize unspoken speech. This supports the theory as described in chapter 2 that this areas of the brain are not much involved in speech production and therefore are also not much involved in the production of unspoken speech.

The next point of the investigation is to see if the Broca's and Wernicke's area are as important as it seems or if the area around the homunculus is sufficient for the recognition of unspoken speech. The result for this question with an error probability of 0.018% is significantly worse compared to the result with all electrodes.

The first result in the top row of figure 6.12 shows also a significantly worse (error probability 0.001 %) word accuracy compared to the word accuracy with all electrodes. Here we investigated if possibly the inverse of the best result of figure 6.12 shows better results. But again the electrodes around the homunculus together with the Broca's area and Wernicke's area are showing the best result.

We investigated the influence of the Broca's area and Wernicke's area. Since Broca's area is responsible for fluent pronunciation and Wernicke's area is responsible for semantic processing then Wernicke's area should not provide a lot of additional information for unspoken speech on single words as used in this experiment. So we used in one experiment the electrodes on the homunculus and only the Broca's area and in the second experiment the homunculus and only the Wernicke's area. As figure 6.13 shows the information of the Wernicke's area are such important that both results in this experiment are nearly the same and significantly worse than the best result. This supports the Geschwind-Wernicke-Model that says that the Wernicke's area is also an important part of speech production and that Broca's area and Wernicke's area work together to produce speech.

The last question was to see if just the Broca's area, the Wernicke's area and the area between them would provide a high word accuracy. As figure 6.12 shows in the bottom left this result is in between the best and the worst result. Compared to the result with all electrodes this result is significantly worse (error probability 2.83 %).

In conclusion we can say that the best result is achieved with all electrodes (16) but that no significant difference exists when focusing on the homunculus and Broca's area and Wernicke's area (7 electrodes) and that this leads to the best result among all other settings.

Subgoal three is reached. It can be shown that the region around the homunculus and the Broca's area and Wernike's area are sufficient for the recognition of unspoken speech.

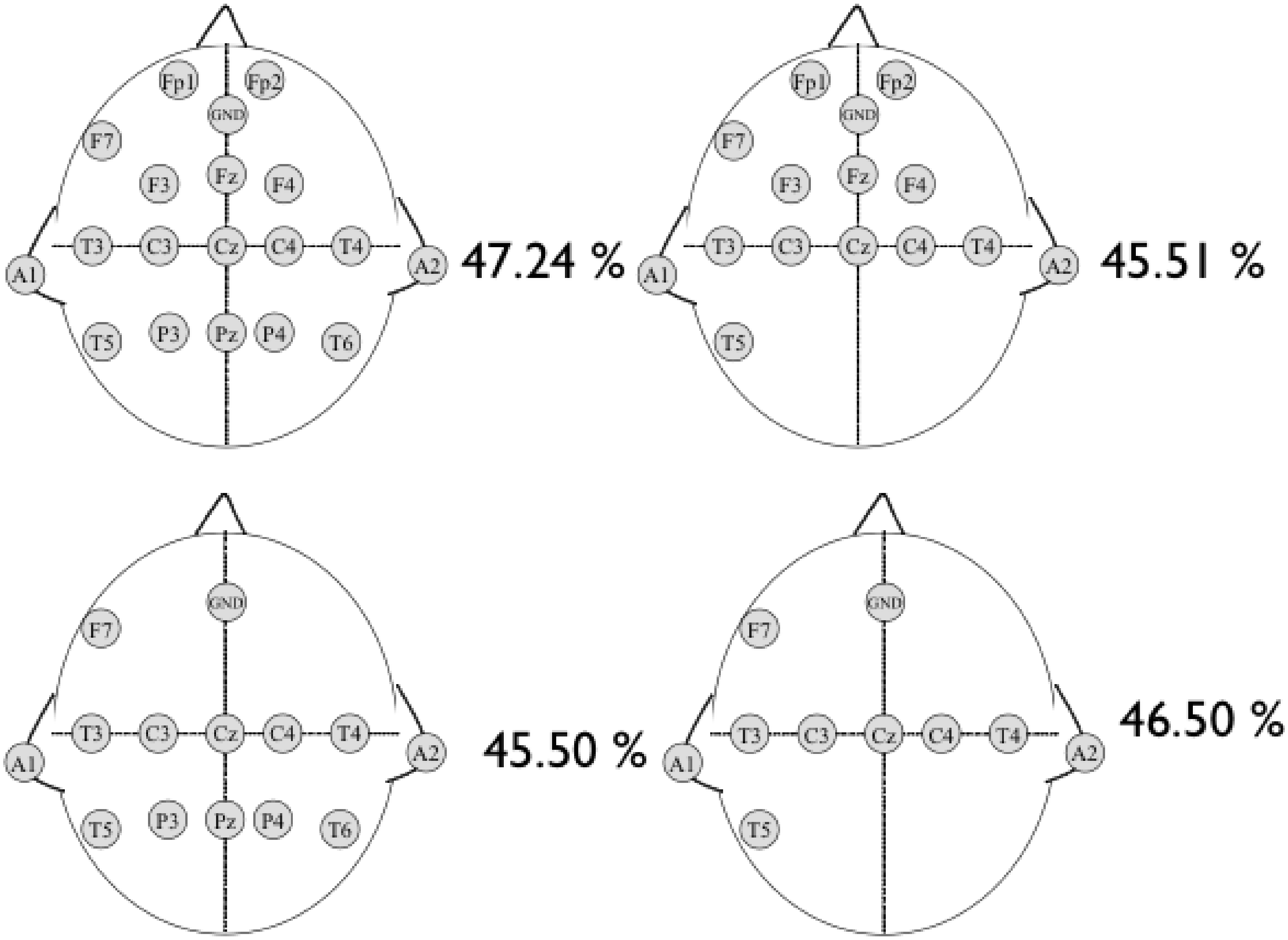


Figure 6.11: Electrode Layout with the word accuracy gained using just the shown electrodes in training and evaluation. The electrodes A1 and A2 are the reference electrodes while the electrode GND is the ground electrode.

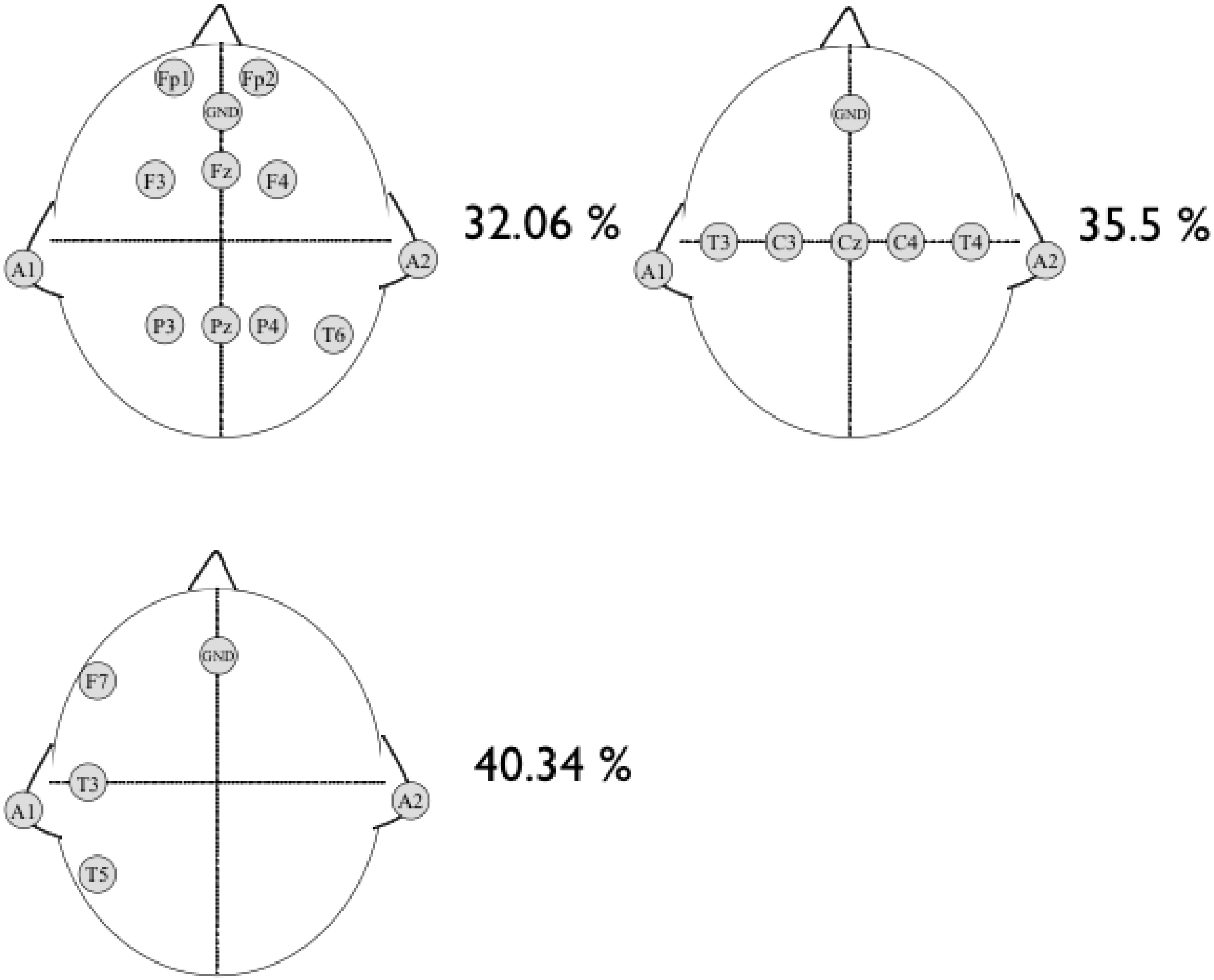


Figure 6.12: The results as word accuracy for the experiments with different electrode positions

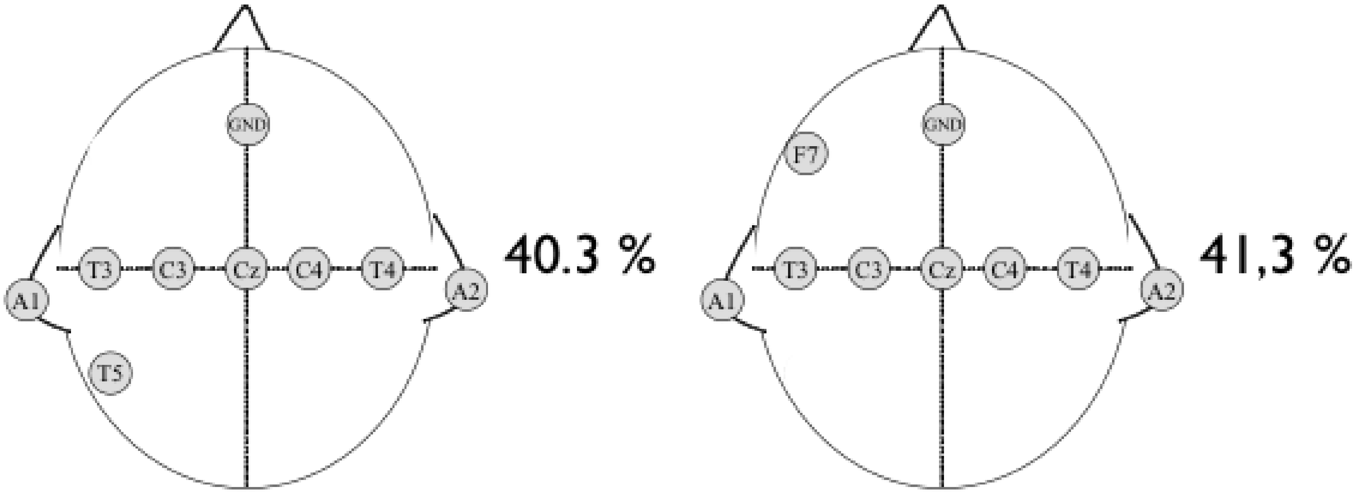


Figure 6.13: Broca's area and Wernicke's area alone do not perform as good as they do together

Chapter 7

Demo System

To test the online recognition capabilities of the unspoken speech recognizer a demo system was built. The results of the offline recognition were very promising so online recognition should be possible.

The setup of the recording room was the same for the demo setup as for the normal recording as presented in section 4.1.1. Only the software needed to be exchanged. The task in the demo was to produce letters with the unspoken speech modality.

The procedure was the following:

1. the subject makes one eye blink
2. the subject utters a word with the unspoken speech modality
3. the subject makes one eye blink
4. the recognizer tries to recognize the word and outputs the hypothesis to the screen as shown in picture 7.1

The vocabulary for the demo was the alpha corpus as introduced in section 5.1.3. To save space just the first letter of the words was output.

The subject was looking on a white screen all the time to have the same conditions as during the recording of the training data.

Before the demo could be started training data had to be recorded due to the session dependency of the recognizer. Then the recognizer needed to be trained and finally the demo system could be started. The subject was given the task to utter five times ”alpha” then five time ”bravo”... and then five times ”echo”. The subject was not interrupted during or in between the process of uttering the 25 words. For later analysis everything was recorded with a video camera.

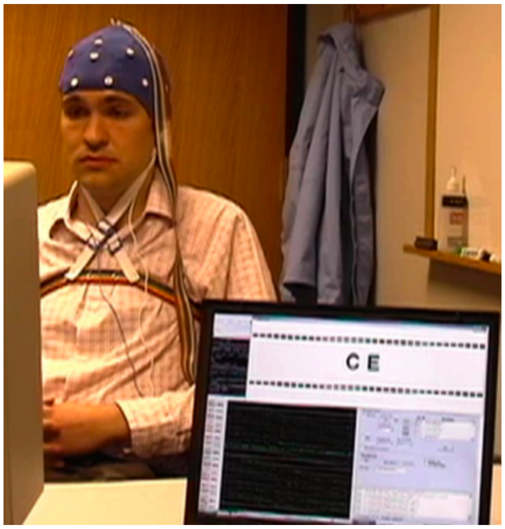


Figure 7.1: The demo setting. The laptop screen shows the hypothesis of the last 2 recognized words, which are "C" and "E"

Nine sessions of which each included the recording of training data and the online recognition were done. Six sessions were done with the alpha corpus, two sessions were done with the digit5 corpus and one session was done with the digit corpus.

None of the sessions produced results measured as word accuracy which were significantly different from chance.

The reason for the bad results may be due to problems with the cap. The collection of training data takes about two hours. Then the recognizer needs to be trained and the demo system needs to be set up which can take also about one hour. During this time the subject needs to wear the cap because it is not possible to get exactly the same electrode positioning as before.

The cap needs to be very tight because the electrodes need a good connection with the scalp. The electrode mountings are made out of hard plastic which is pressed against the scalp. This inflicts pain after about 90 minutes as the subject reported.

There are two consequences because of the pain. The first is that the subject cannot be as focused with pain on the scalp as without pain. Because of that the unspoken speech during the online recognition is not uttered in the same way as during the recording of the training data. Therefore the learned patterns from the training data differ from the patterns during the online recognition.

The second consequence is that brain waves changes with pain. Baltas [37] even built a pain detection system based on EEG data. Therefore the learned patterns from the training data also differ from the patterns produced during the online recognition task.

A solution for this problem would be to use a cap which is more comfortable to wear and does not inflict pain. Another solution might be to try to get rid of the pain artifacts with a better approach in the preprocessing.

Chapter 8

Conclusions and Future Work

8.1 Summary and Conclusion

In this work we showed a setup for recording EEG-data during the production of speech in five different modalities: normal speech, whispered speech, silent speech, mumbled speech and unspoken speech. Furthermore we introduced a system to recognize speech in this five modalities which uses methods of speech recognition for spoken audible speech. The main focus was on the recognition of unspoken speech which is uttered without any muscle movement. Finally an investigation was done to identify the regions of the brain which produce the most interesting brain waves for unspoken speech recognition.

The results of the experiments which were conducted showed that speech recognition on EEG brain waves is possible with a word accuracy four to five times higher than chance for vocabularies of up to ten words. The same results were found for the other modalities. Unspoken speech was slightly but not significantly worse than the other modalities. The results also showed that the important regions for unspoken speech recognition seem to be the homunculus, the Broca's area and Wernicke's area.

Still there are defiances to be solved. Speaker and session dependency makes the usage of the system difficult. For every recognition task training data has to be collected beforehand. The largest problem to solve is the inability of online recognition due to pain inflicted by wearing the cap for longer than 90 minutes.

This results show that there is a potential for breaking barriers in interaction with computers and through this with other humans. For physically challenged people unspoken speech is sometimes the only efficient way to communicate with their environment.

8.2 Outlook

This work is to be seen as a feasibility study. It does not claim completeness. There are still areas which need to be improved. Improvements are needed in the preprocessing. Methods for feature extraction like wavelets or independent component analysis could improve the recognition and make artifact detection easier.

Improvements are also needed for the cap with respect to the number and location of electrodes and comfort. A higher density of electrodes might provide more information for the recognizer. This would make the system also more reliable for interferences which could come from single electrodes. A cap which is more comfortable to wear would decrease the infliction of pain and therefore online recognition might be feasible.

Appendix A

Software Documentation

This chapter should give a brief overview of the technical details of the recognition system. It is a starting point to get an understanding of how to use it.

A.1 Janus

For all recognition tasks the Janus Recognition Toolkit was used. This is a framework to build speech recognition systems. It is written in C[1] and provides a TCL[2] wrapper to control it. It is available for various platforms. For the recognition task in this work the Linux operating system was used.

The recognition system was adapted from a state of the art speech recognizer. To make the exchange of parameters for the experiments easier all important parameters were made available in two files. This files are *desc/baseDesc.tcl* and *desc/featDesc_eeg.tcl*. This first file contains parameters concerning the recognition system like the number of states for the HMM, the number of gaussians, the corpus and so forth. The second file contains the description of the feature extraction.

Because of the high number of different modalities a lot of recognizers needed to be trained and evaluated. The recognizer was developed at Carnegie Mellon University where a Condor-Cluster[38] is available to compute high numbers of parallel tasks. The recognition task of different modalities was parallelable such that every recognition system can run independent from each other.

To build a parallel system three task had to be solved. First the system needs to be changed to work in a parallel manner. Second the system needs to be started from a central spot. Third the system needs to deliver the recognition results from the parallel tasks.

[1] http://en.wikipedia.org/wiki/C_programming_language

[2] http://en.wikipedia.org/wiki/Tcl

The recognizer that was used as the base for the resulting system was already partly able to solve the training and evaluation task in parallel. This just worked for a lot of data but not as in the task of this work for a small amount of data but different recognition tasks. It was also not possible to start the recognizer with different parameters at the same time as needed for our task.

To make the system work in parallel there are two ways. The first way would be to develop a complex system residing in one directory capable of doing different recognition tasks with different parameters at one time. The Condor-Cluster tries to dispatch different jobs like the recognition task to computers with free computing capacities. Therefore the recognizer would have had to cope with the problem that different computers try to write data to the same directories and files at the same time.

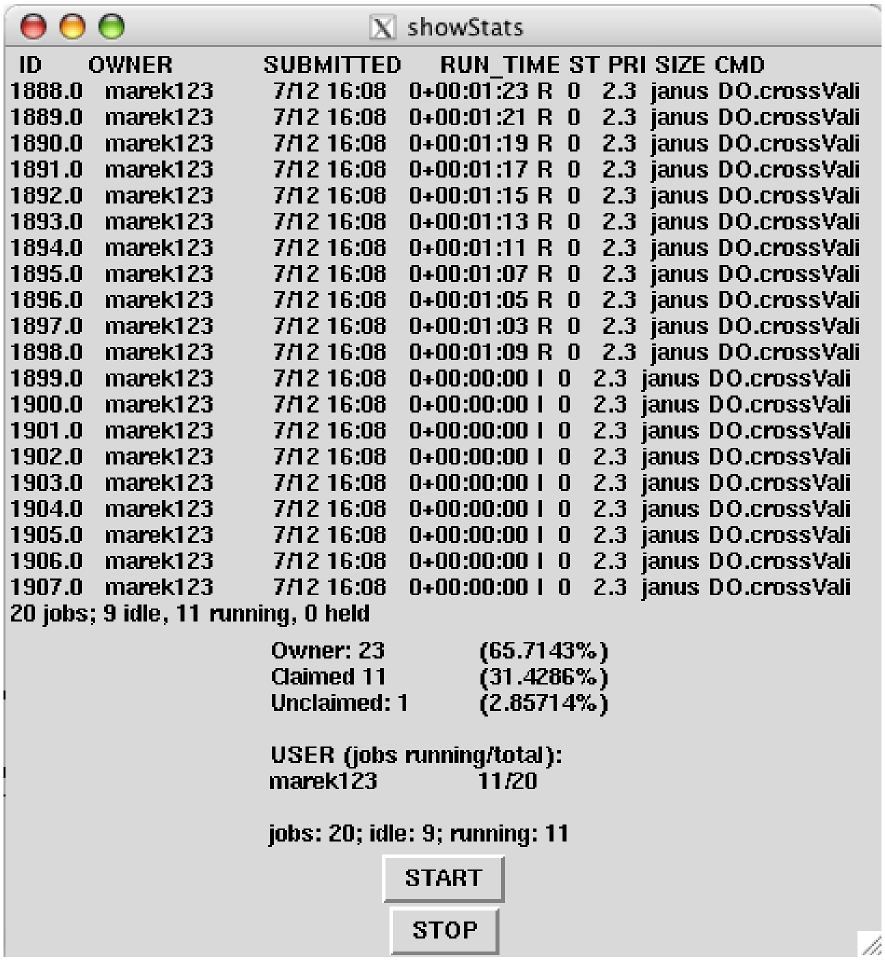
The second way of solving this problem is to take the already working recognizer and to duplicate it and start the systems isolated from each other. This has the advantage that the developed and working recognizer can be used and no additional efforts need to be invested in solving problems coming from parallelisation. The disadvantage is that tools need to be developed to control this set of duplicates.

Because the seconds approach seemed to provide fast and more reliable results it was selected for this work. The spot of controlling the system is the master.tcl script. It can be used to solve all the three tasks mentioned earlier. The first task is to enable the system to work in parallel. For this it has to be duplicated and initialized with initial parameters. This is done with the *build* parameter in the master.tcl script.

The syntax is: *janus master.tcl build <speaker-id>_<session-id>_<modality> -domain <corpus>*. E.g. *janus master.tcl build 02_03_whisper -domain digit* builds a recognizer for the subject 02 in session 03 speaking in modality whispered speech with the digit corpus.

There has to be a possibility to update parameters in the different recognizers. This is done by the *updateFile* parameter. It would be desirable for this task to address more than one recognizer at a time. The first possibility to do this is to simply concatenate the different recognizer names e.g. *janus master.tcl updateFile "01_02_whisper 01_02_mumble 02_03_whisper" eeg_recognizer/desc/baseDesc.tcl*. In this example the file eeg_recognizer/desc/baseDesc.tcl is copied to the recognizers "01_02_whisper 01_02_mumble 02_03_whisper". This can also be expressed in a shorter way by using the % symbol which works like the usual Kleene-star *. The example could also be written like this: *janus master.tcl updateFile "01_02_% 02_03_whisper" eeg_recognizer/desc/baseDesc.tcl*.

The next step to solve for the task of parallelisation is to control the recognizers. The recognizers have to be started, monitored and possibly terminated. For that the following parameters can be attached to the master.tcl: *start, showStat, kill.* To start the training and evaluation for the session 02 of subject 04 in all modalities the command is: *janus master.tcl start 04_02_%*. To start the monitoring of the started jobs this command is needed: *janus showStat _*. This opens a TK[3] window as shown in figure A.1 showing the status of all jobs and the Condor cluster. To possibly kill the jobs from job number 1023 to 1059 this command will work: *janus master.tcl kill 1023-1059*.



```
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
1888.0   marek123         7/12 16:08   0+00:01:23 R  0   2.3  janus DO.crossVali
1889.0   marek123         7/12 16:08   0+00:01:21 R  0   2.3  janus DO.crossVali
1890.0   marek123         7/12 16:08   0+00:01:19 R  0   2.3  janus DO.crossVali
1891.0   marek123         7/12 16:08   0+00:01:17 R  0   2.3  janus DO.crossVali
1892.0   marek123         7/12 16:08   0+00:01:15 R  0   2.3  janus DO.crossVali
1893.0   marek123         7/12 16:08   0+00:01:13 R  0   2.3  janus DO.crossVali
1894.0   marek123         7/12 16:08   0+00:01:11 R  0   2.3  janus DO.crossVali
1895.0   marek123         7/12 16:08   0+00:01:07 R  0   2.3  janus DO.crossVali
1896.0   marek123         7/12 16:08   0+00:01:05 R  0   2.3  janus DO.crossVali
1897.0   marek123         7/12 16:08   0+00:01:03 R  0   2.3  janus DO.crossVali
1898.0   marek123         7/12 16:08   0+00:01:09 R  0   2.3  janus DO.crossVali
1899.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1900.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1901.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1902.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1903.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1904.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1905.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1906.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
1907.0   marek123         7/12 16:08   0+00:00:00 I  0   2.3  janus DO.crossVali
20 jobs; 9 idle, 11 running, 0 held

                         Owner: 23          (65.7143%)
                         Claimed 11         (31.4286%)
                         Unclaimed: 1       (2.85714%)

                         USER (jobs running/total):
                         marek123           11/20

                         jobs: 20; idle: 9; running: 11
```

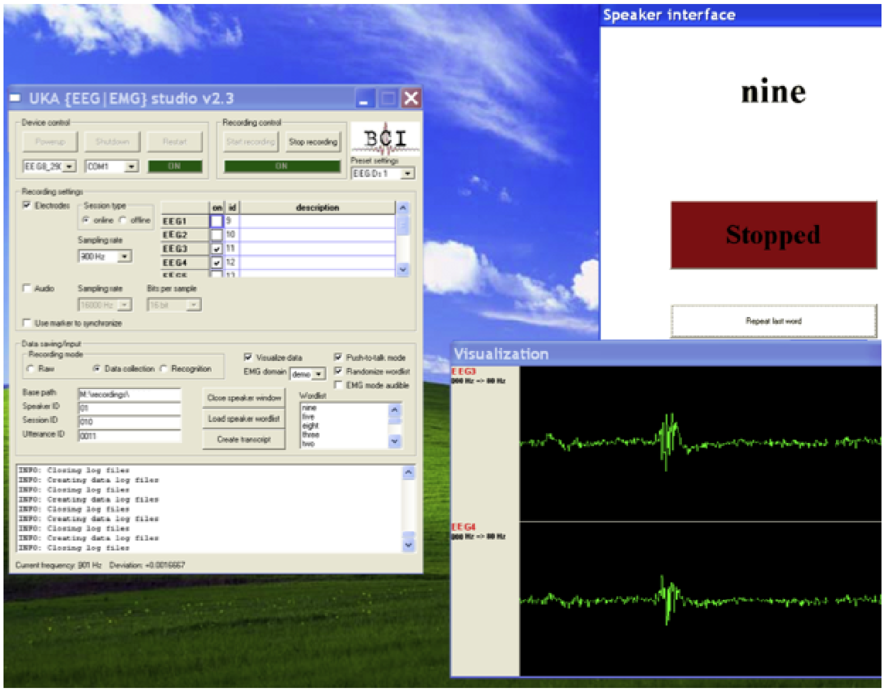
Figure A.1: TK window showing the status of the jobs and the cluster

The remaining task is to get the results as word accuracy and as a confusion matrix from the jobs that ran. The following command will present the results for session 02 of subject 04 in all recorded modalities: *janus master.tcl results 04_02_%*.

[3] http://en.wikipedia.org/wiki/Tk

A.2 Recording Software

The recording software "UKA {EEG—EMG} Studio 2.10mwR" (a screenshot can be seen on image A.2) was developed at University of Karlsruhe in Germany at the ITI Waibel labs and modified for this work at CMU. This software is developed in C++[4] and runs in the Microsoft Windows operation system only.



EEG8_29(
COM1
ON
ON
Preset settings: EEG D: 1

☑ Electrodes
Session type: ● online ○ offline
Sampling rate: 300 Hz

	on	id	description
EEG1	□	9	
EEG2	□	10	
EEG3	☑	11	
EEG4	☑	12	
EEG5	□	13	

□ Audio Sampling rate: 16000 Hz Bits per sample: 16 bit
□ Use marker to synchronize

Recording mode: ○ Raw ● Data collection ○ Recognition
☑ Visualize data
EMG domain: demo
☑ Push-to-talk mode
☑ Randomize wordlist
□ EMG mode audible

Base path: M:\recordings\
Speaker ID: 01
Session ID: 010
Utterance ID: 0011

Wordlist:
nine
five
eight
three
two

```
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
```

Current frequency: 301 Hz Deviation: +0.0016667

nine

Stopped

Figure A.2: The software used for the recordings of brain waves

The recording software was developed for the recording to be done on one screen. For our recording task a system is needed which has a different screen for the recording assistant and the subject that is recorded. For this the control window that shows the push to talk button and the word that has to be uttered needed to be modified. The window was enlarged so that it would span over one and a half screen showing on the one screen the controls for the recording assistant and on the other screen the word that has to be uttered for the subject.

Another requirement was to implement the successively changing words on the subjects screen as described in section 4.1.2. This requirement was implemented so that after the recording assistant pressed the recording button the sequence of words started to show on the subjects screen.

The recording software also needed to be changed for the demo system. In the case of the

[4] http://en.wikipedia.org/wiki/C_Plus_Plus

demo system the recording software had to detect eye blinks. The procedure for the demo system was the following:

1. start writing recorded data to a to a file called *recording-<number>.adc* where <number> is a number starting with "1" increased by one after every recording

2. detect the first eye blink

3. detect the second eye blink

4. close the file and start over

The janus recognizer was waiting for the file with the name *recording-1.adc*. After this file appeared janus had to wait for the appearance of *recording-2.adc* because *recording-1.adc* was still recorded. When *recording-2.adc* appeared it did the recognition of the uttered word in file *recording-1.adc* and showed the hypothesis in a TK window and waited for *recording-3.adc* to appear and then did the recognition on *recording-2.adc* and so on.

Appendix B

Recorded Data

id	speak	whisper	silent	mumble	think	Σ	minutes	domain
02-01	5	5	5	5	5	250	25.78	digit
03-01	5	5	5	5	5	250	27.36	digit
04-01	5	5	5	5	5	250	22.85	digit
05-01	5	5	5	5	5	250	22.73	digit
06-01					25	250	31.33	digit
06-03					25	250	30.5	digit
06-04					20	100	17.33	lecture
06-05					20	100	11.71	phone
06-06	10		10		10	300	37.4	digit
06-07					25	125	18.33	digit5
06-08					3	3	2.46	1234554321
06-09					25	125	16.46	alpha
06-10					3	3	2.35	abcdeedcba
01-01	5	5	5	5	5	250	24.11	digit
01-02	5	5	5	5	5	250	24.4	digit
01-03	5	5	5	5	5	250	22.55	digit
01-04	5	5	5	5	5	250	23.93	digit
01-05	5	5	5	5	5	250	23.71	digit
01-06	5	5	5	5	5	250	22.28	digit
01-07	25					250	32.1	digit
01-08			25			250	37.48	digit
01-09					25	250	43.13	digit
01-10					10	50	12.2	lecture

01-11	10		10		10	300	42.4	digit
01-12	15		15		15	225	45.48	lecture
01-13					20	200	28.65	gre
01-14					20	100	11.88	phone
01-15					15	75	9.5	player
01-16					15	75	9,01	player
01-17					10	10	16	player_long
01-18					15	75	17.23	player
01-19					10	10	18.2	player_long
01-20					30	150	22.86	digit5
01-21					30	150	24.61	alpha
01-22					20	100	14.83	digit5
01-23					20	100	14.6	digit5
01-24					30	300	46.66	digit
01-25					30	150	23.06	alpha
01-26					15	75	13.05	alpha
01-27					60	300	50	alpha
01-28					30	150	23.26	alpha
01-29	20		20		20	300	46.76	alpha
01-30			20		20	200	28.66	alpha

Table B.1: Overview of how many utterances were recorded in every session

Appendix C

Results of the experiments from section 6.1

This is a detailed report of the experimental results concerning the parameters of the recognizer. The maximum values per modality per parameter are marked bold.

window size	**session**				**Accuracy**
26.6ms	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	**45.0%**
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	48.0%
unspoken	54.0%	50.0%	58.0%	36.0%	49.5%
					47.0%
53.3ms					
speech	40.0%	44.0%	42.0%	36.0%	**40.5%**
whisper	52.0%	48.0%	50.0%	30.0%	**45.0%**
silent	60.0%	62.0%	54.0%	44.0%	**55.0%**
mumble	46.0%	52.0%	44.0%	52.0%	**48.5%**
unspoken	60.0%	56.0%	72.0%	52.0%	**60.0%**
					49.8%
106.6ms					
speech	40.0%	30.0%	38.0%	40.0%	37.0%
whisper	52.0%	26.0%	32.0%	28.0%	34.5%
silent	46.0%	42.0%	46.0%	40.0%	43.5%
mumble	34.0%	42.0%	42.0%	42.0%	40.0%
unspoken	38.0%	52.0%	56.0%	38.0%	46.0%
					40.2%

Table C.1: The window size of 53.3ms is better for unspoken speech.

window shift	**session**				**Accuracy**
4ms	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	36.0%	38.0%	52.0%	32.0%	**39.5%**
whisper	52.0%	54.0%	46.0%	28.0%	**45.0%**
silent	52.0%	56.0%	58.0%	46.0%	**53.0%**
mumble	38.0%	52.0%	58.0%	44.0%	**48.0%**
unspoken	54.0%	50.0%	58.0%	36.0%	**49.5%**
					47.0%
8ms					
speech	30.0%	46.0%	40.0%	42.0%	**39.5%**
whisper	44.0%	46.0%	50.0%	38.0%	44.5%
silent	34.0%	50.0%	58.0%	42.0%	46.0%
mumble	34.0%	48.0%	52.0%	28.0%	40.5%
unspoken	54.0%	52.0%	58.0%	32.0%	49.0%
					43.9%
16ms					
speech	30.0%	42.0%	30.0%	30.0%	33.0%
whisper	46.0%	26.0%	40.0%	26.0%	34.5%
silent	18.0%	40.0%	42.0%	46.0%	36.5%
mumble	24.0%	40.0%	30.0%	26.0%	30.0%
unspoken	46.0%	34.0%	52.0%	26.0%	39.5%
					34.7%
27 ms					
speech	24.0%	18.0%	36.0%	20.0%	24.5%
whisper	36.0%	24.0%	18.0%	20.0%	24.5%
silent	34.0%	36.0%	44.0%	28.0%	35.5%
mumble	28.0%	28.0%	14.0%	20.0%	22.5%
unspoken	32.0%	20.0%	28.0%	14.0%	23.5%
					26.1%

Table C.2: A window shift of 4ms is ideal.

gaussians	**session**				**Accuracy**
4 gaussians	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	42.0%	44.0%	52.0%	36.0%	**43.5%**
whisper	66.0%	46.0%	58.0%	34.0%	**51.0%**
silent	58.0%	54.0%	60.0%	44.0%	54.0%
mumble	42.0%	48.0%	58.0%	36.0%	46.0%
unspoken	5.0%	46.0%	64.0%	46.0%	**51.5%**
					49.2%
8 gaussians					
speech	40.0%	44.0%	46.0%	32.0%	40.5%
whisper	46.0%	5.0%	58.0%	34.0%	47.0%
silent	52.0%	48.0%	58.0%	46.0%	51.0%
mumble	38.0%	48.0%	54.0%	40.0%	45.0%
unspoken	44.0%	44.0%	64.0%	40.0%	48.0%
					46.3%
16 gaussians					
speech	40.0%	44.0%	48.0%	42.0%	**43.5%**
whisper	5.0%	48.0%	42.0%	22.0%	40.5%
silent	62.0%	54.0%	58.0%	46.0%	**55.0%**
mumble	44.0%	5.0%	56.0%	38.0%	47.0%
unspoken	46.0%	48.0%	60.0%	42.0%	49.0%
					47.0%
25 gaussians					
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	45.0%
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	**48.0%**
unspoken	54.0%	5.0%	58.0%	36.0%	49.5%
					47.0%
32 gaussians					
speech	42.0%	44.0%	48.0%	30.0%	41.0%
whisper	54.0%	58.0%	54.0%	38.0%	**51.0%**
silent	52.0%	62.0%	54.0%	42.0%	52.5%
mumble	38.0%	48.0%	54.0%	44.0%	46.0%
unspoken	54.0%	46.0%	62.0%	36.0%	49.5%
					48.0%
64 gaussians					
speech	40.0%	46.0%	42.0%	34.0%	40.5%
whisper	52.0%	52.0%	38.0%	26.0%	42.0%
silent	36.0%	5.0%	34.0%	44.0%	41.0%
mumble	40.0%	56.0%	54.0%	40.0%	47.5%
unspoken	42.0%	44.0%	48.0%	32.0%	41.5%
					42.5%

Table C.3: No significant difference can be seen for up to 32 gaussians. 64 gaussians are too much.

states	**session**				**Accuracy**
3 states	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	36.0%	42.0%	52.0%	40.0%	42.5%
whisper	58.0%	54.0%	46.0%	36.0%	48.5%
silent	48.0%	54.0%	62.0%	44.0%	52.0%
mumble	40.0%	48.0%	44.0%	44.0%	44.0%
unspoken	64.0%	56.0%	56.0%	42.0%	**54.5%**
					48.3%
4 states					
speech	44.0%	36.0%	52.0%	36.0%	42.0%
whisper	52.0%	48.0%	54.0%	40.0%	48.5%
silent	46.0%	64.0%	62.0%	38.0%	52.5%
mumble	36.0%	44.0%	36.0%	44.0%	40.0%
unspoken	52.0%	52.0%	56.0%	38.0%	49.5%
					46.5%
5 states					
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	45.0%
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	**48.0%**
unspoken	54.0%	50.0%	58.0%	36.0%	49.5%
					47.0%
6 states					
speech	46.0%	48.0%	42.0%	44.0%	**45.0%**
whisper	52.0%	50.0%	48.0%	32.0%	45.5%
silent	50.0%	60.0%	60.0%	50.0%	**55.0%**
mumble	42.0%	48.0%	52.0%	38.0%	45.0%
unspoken	46.0%	48.0%	50.0%	42.0%	46.5%
					47.4%
7 states					
speech	40.0%	46.0%	48.0%	36.0%	42.5%
whisper	58.0%	50.0%	54.0%	42.0%	**51.0%**
silent	54.0%	48.0%	52.0%	44.0%	49.5%
mumble	34.0%	54.0%	40.0%	38.0%	41.5%
unspoken	48.0%	54.0%	50.0%	32.0%	46.0%
					46.1%

Table C.4: no significant difference in the overall performance but unspoken speech seems to do best with 3 states

coeff. after LDA	**session**				**Accuracy**
4	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	26.0%	38.0%	38.0%	36.0%	34.5%
whisper	50.0%	48.0%	42.0%	34.0%	43.5%
silent	42.0%	48.0%	50.0%	50.0%	47.5%
mumble	40.0%	54.0%	54.0%	32.0%	45.0%
unspoken	36.0%	40.0%	48.0%	40.0%	41.0%
					42.3%
8					
speech	42.0%	50.0%	52.0%	36.0%	45.0%
whisper	62.0%	58.0%	52.0%	38.0%	52.5%
silent	42.0%	50.0%	64.0%	52.0%	52.0%
mumble	50.0%	66.0%	60.0%	46.0%	**55.5%**
unspoken	54.0%	42.0%	58.0%	46.0%	50.0%
					51.0%
16					
speech	50.0%	50.0%	54.0%	40.0%	**48.5%**
whisper	58.0%	56.0%	62.0%	36.0%	**53.0%**
silent	50.0%	56.0%	64.0%	56.0%	**56.5%**
mumble	52.0%	56.0%	54.0%	44.0%	51.5%
unspoken	54.0%	44.0%	74.0%	42.0%	**53.5%**
					52.6%
35					
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	45.0%
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	48.0%
unspoken	54.0%	50.0%	58.0%	36.0%	49.5%
					47.0%
40					
speech	40.0%	34.0%	50.0%	30.0%	38.5%
whisper	48.0%	56.0%	54.0%	28.0%	46.5%
silent	46.0%	68.0%	60.0%	40.0%	53.5%
mumble	40.0%	44.0%	58.0%	32.0%	43.5%
unspoken	52.0%	42.0%	56.0%	36.0%	46.5%
					45.7%
64					
speech	36.0%	38.0%	48.0%	30.0%	38.0%
whisper	42.0%	56.0%	30.0%	30.0%	39.5%
silent	44.0%	62.0%	48.0%	30.0%	46.0%
mumble	30.0%	46.0%	52.0%	44.0%	43.0%
unspoken	54.0%	44.0%	42.0%	18.0%	39.5%
					41.2%

Table C.5: up to 35 coefficients are best for the recognizer after the dimensionality reduction was done

features	**session**				**Accuracy**
stft	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	58.0%	52.0%	60.0%	50.0%	**55.0%**
whisper	58.0%	44.0%	58.0%	40.0%	**50.0%**
silent	40.0%	58.0%	76.0%	42.0%	54.0%
mumble	28.0%	60.0%	36.0%	40.0%	41.0%
unspoken	46.0%	46.0%	24.0%	26.0%	35.5%
					47.1%
delta					
speech	48.0%	42.0%	46.0%	48.0%	46.0%
whisper	58.0%	46.0%	52.0%	30.0%	46.5%
silent	44.0%	68.0%	56.0%	52.0%	**55.0%**
mumble	42.0%	54.0%	54.0%	48.0%	**49.5%**
unspoken	56.0%	50.0%	48.0%	24.0%	44.5%
					48.3%
delta delta					
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	45.0%
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	48.0%
unspoken	54.0%	50.0%	58.0%	36.0%	49.5%
					47.0%
delta delta delta					
speech	38.0%	48.0%	52.0%	38.0%	44.0%
whisper	48.0%	52.0%	56.0%	34.0%	47.5%
silent	48.0%	60.0%	54.0%	42.0%	51.0%
mumble	38.0%	48.0%	42.0%	32.0%	40.0%
unspoken	54.0%	42.0%	56.0%	38.0%	47.5%
					46.0%
delta mean[1]					
speech	44.0%	48.0%	46.0%	42.0%	45.0%
whisper	50.0%	50.0%	46.0%	38.0%	46.0%
silent	48.0%	58.0%	56.0%	46.0%	52.0%
mumble	52.0%	56.0%	56.0%	40.0%	51.0%
unspoken	56.0%	48.0%	68.0%	30.0%	**50.5%**
					48.9%

Table C.6: delta features increase the recognition of unspoken speech

LDA/no LDA	**session**				**Accuracy**
with lda	**01-02/5**	**01-04/5**	**01-05/5**	**01-06/5**	**Average**
speech	36.0%	38.0%	52.0%	32.0%	39.5%
whisper	52.0%	54.0%	46.0%	28.0%	45.0%
silent	52.0%	56.0%	58.0%	46.0%	53.0%
mumble	38.0%	52.0%	58.0%	44.0%	48.0%
unspoken	54.0%	50.0%	58.0%	36.0%	49.5%
					47.0%
without lda					
speech	14.0%	24.0%	20.0%	12.0%	17.5%
whisper	16.0%	14.0%	14.0%	24.0%	17.0%
silent	12.0%	18.0%	20.0%	24.0%	18.5%
mumble	18.0%	24.0%	14.0%	20.0%	19.0%
unspoken	8.0%	6.0%	4.0%	6.0%	6.0%
					15.6%

Table C.7: lda is very important for the current recognizer

Bibliography

[1] Niels Birbaumer. The Thought Translation Device (TTD) for Completely Paralyzed Patients. *IEEE*, 2000.

[2] Marc R. Nuwer et al. IFCN standards for digital recording of clinical EEG. *Electroencephalography and clinical Neurophysiology*, (106):259 – 261, 1998.

[3] U.S. National Cancer Institute's Surveillance. Benign Brain Tumor Reporting `http://training.seer.cancer.gov/ss_module00_bbt/unit02_sec04_b_cells.html`.

[4] Carlson, Niel. A. *Foundations of Physiological Psychology.* Needham Heights, Massachusetts: Simon & Schuster., 1992.

[5] Neurology for Physiology Students `http://www.science.uwc.ac.za/physiology/neurology/neuro.cont.htm`.

[6] Greg Hickok. The Neuroscience of Language. Lecutre Notes.

[7] Jonathan R. Wolpawa, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, Theresa M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 2002.

[8] *Das menschliche Gehirn.* Christian Brandsätter - Wien-München, 1999.

[9] Matthias Honal, Tanja Schultz. Identifying User State using Electroencephalographic Data. *Proceedings of the International Conference on Multimodal Input (ICMI)*, 2005.

[10] T9 predictive text `http://www.tegic.com/`.

[11] Koles ZJ, Lind JC, Flor-Henry P. Spatial patterns in the background EEG underlying mental disease in man. *Electroencephalogr Clin Neurophysiol.*, 1994.

[12] Congress of the United States. Bill of Rights, March 1789.

[13] M. Finke and P. Geutner and H. Hild and T. Kemp and K. Ries and M. Westphal. The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proc. ICASSP '97*, pages 83–86, Munich, Germany, 1997.

[14] Lawrence R. Rabinaer. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, 1989.

[15] E. O. Brigham. *The fast Fourier transform and its applications.* Prentice Hall Signal Processing Series, Englewood Cliffs, NJ, 1988.

[16] Haeb-Umbach, R. and Ney, H. Linear Disriminant Analysis for Improved Large Vocabulary Continous Speech Recognition. In *Proceedings of the ICASSP*, 1992.

[17] Hans Berger. Über das Elektroencephalogramm des Menschen (On the human electroencephalogram). *Archiv f.Psychiatrie u.Nervenkrankheiten*, 1929.

[18] V. S. Ramachandran. *Encyclopedia of the Human Brain*, volume 2. Academic Press, 2002.

[19] H.H. Jasper. The Ten-Twenty Electrode System of the International Federation. *Electroencephalography and Clinical Neurophysiology. EEG Journal*, (10):371–375, 1958.

[20] Prof. Dr.-Ing. K. Meyer-Waarden. *Bioelektrische Signale und ihre Ableitverfahren.* Schattauer (Stuttgart - New York), 1985.

[21] P. P. Broca. Perte de la parole; ramolissement chronique et destruction partielle du lobe antérieur gauche de cerveau. *Bulletins de la Société d'anthropologie de Paris*, 2:235–238, 1861.

[22] Geschwind, Norman. Specializations of the Human Brain. *Scientific American*, 241(3):180–99, September 1979.

[23] Jeffrey R. Binder et al. Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, 1997.

[24] McAdam, D. W. & Whitaker, H. A. Language production: electroencephalographic localization in the normal human brain. *Science*, 172(982):499–502, 1971.

[25] Matthew Middendorf. Brain–Computer Interfaces Based on the Steady-State Visual-Evoked Response. *IEEE Transactions on rehabilitation engineering*, 2000.

[26] L.A. Farwell and E. Donchin. Talking off the top of youe head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 1988.

[27] J. R. Wolpaw and D. J. McFarland. Multichannel EEG-based brain-computer communication. *Electroenceph. Clin. Neurophysiol.*, 1994.

[28] Guido Dornhege et al. Combining Features for BCI. 2003.

[29] Moritz Grosse Wentrup. EEG Source Localization for Brain-Computer-Interface. *Proceedings othe the 2nd International IEEE EMBS*, 2005.

[30] Roman Krepki, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller. The Berlin Brain-Computer Interface (BBCI). 2003.

[31] Patrick Suppes et al. Invariance between subjects of brain waves representations of language. 1999.

[32] Patrick Suppes et al. Invariance of brain-wave representations of simple visual images and their names. *PNAS*, 1999.

[33] Patrick Suppes, Zhong-Lin Lu and Bing Han. Brain wave recognition of words. *Proc Natl Acad Sci U S A*, 1997.

[34] Sameer Singh. EEG Data Classification with Localised Structural Information. *IEEE*, 2000.

[35] Mayer C. UKA EMG/EEG Studio v2.0.

[36] William Orr Dingwall , Harry A. Whiteaker. Neurolinguistics. *Annual Review of Anthropology*, 1974.

[37] E. Baltas et al. An LVQ Classifier of EEG Coherence Patterns for Pain Detection. 2002.

[38] Condor Cluster http://www.cs.wisc.edu/condor/.