Further Investigations on Unspoken Speech

- Findings in an attempt of developing EEG-based word recognition

Studienarbeit

Jan-Peter Calliess

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

November 15, 2006

Advisors: Dr. T. Schultz, Prof.Dr.rer.nat. A. Waibel

Hiermit versichere ich, die Arbeit selbständig erstellt und keine anderen, als die angegebenen Hilfsmittel benutzt zu haben.

Vorname, Name Pittsburgh, den Datum

Abstract

Classification of signals originating from brain activity has become an active area of research in the past years. While the trend has evolved in the direction of using modern imaging techniques such as fMRI, we consider brain-activity based speech recognition more suitable to be attempted using techniques with high temporal resolution such as electroencephalography. Past achievements demonstrate that EEG-based approaches are promising [39].

This thesis is a continuation of a previous thesis ([43]), written and implemented by Marek Wester that had been modifying a speech recognizer to be applied on brain waves in order to attempt the recognition of words from a small vocabulary domain, uttered in different modalities. The modalities included normal *speaking*, *mumbling*, yet the primary focus had been on the recognition of words uttered in a modality called *unspoken speech*. In the latter modality the test subject was asked to think of the word in question without making any audible sound or involvement of facial muscle movements at all.

Our main task was to aggregate electrodes on the scalp above the orofacial motor cortex and compare the recognition results to the ones achieved using the standard layout already used in [43] for the modalities *speaking* and *unspoken speech*. For this purpose,we acquired a new cap for EEG - recordings that was, due to a higher electrode density, better suited for realizing such arrangements of electrodes, adapt the implementation of the work mentioned above and executed experiments. We found no significant differences between the estimated recognition rates on data produced with the aggregated layout of the new cap and the data acquired with the standard layout of the old low-density cap.

However, we executed investigations making it doubtful whether the estimated recognition rates obtained in [43] and at the beginning of this work are meaningful. Instead, we believe to have demonstrated that, with the current setting, the structure of the recorded training examples actually caused the speech recognizer to learn a wrong, temporal concept that misleadingly was correlated to the labeling of the first training data sets.

On the training data presumably unaffected by such problems, the estimated recognition results were inconclusive. For *unspoken speech* utterances, some were significantly higher than random guessing, some were not. We had only one unaffected training data set with instances of the *speaking* modality. It was significantly higher than random guessing. However, further investigations producing more unaffected data will have to be made.

In the course of the work, additional improvements and ideas were developed. These include the derivation and implementation of a cross-correlation based pattern detection algorithm that was then used for blink detection, as well as methods and investigations for the purpose of being able to estimate the test subject's brain tissue conductivity properties, which translate to an estimation of the degree of ability to assign meaningfulness to some aspects of the experiments.

Acknowledgements

I want to thank my advisor, Dr.Tanja Schultz for guidance and providing me with the means to work in this interesting field. I also want to express my gratitude to Marek Wester and Matthias Honal who repeatedly provided helpful advice regarding some of the technical details of this work. Szu-Chen Jou was helping a lot with ordering textbooks whenever I needed it and was always willing to help with a hint as far as his demanding schedule would allow it. This work would not have been possible without the aid of many people, including Michael Wand and especially Friedrich Faubel and Mangala Srinivas, who spent several hours volunteering to participate in the recording sessions of this work and who have also been very supportive in several discussions. Special thanks to them!
This research was partly funded by the Baden-Württemberg-Stipendium (InterACT) granted by the Landesstiftung Baden-Württemberg.

Contents

Chapter 1

Foundations

1.1 Preliminaries

The idea that underlies the efforts this thesis is based on, is that the mental or verbal production of words is reflected by physiological processes in the brain that can be recorded by suitable measurement devices and which follow on average different patterns that in principle would allow a discrimination of the distinct words. Many discoveries in the past decades have been encouraging , even though to our knowledge it has not yet been accomplished to discriminate words of the same semantic category and hence the first attempt to do so, with seemingly positive results, was done in [43] which this thesis is meant to continue. Since the data the classification mechanism is applied on eventually determines the nature of the patterns one is able to detect, the selection of the recording technique is crucial, if the necessary information is to be contained that allows a successful discrimination of the words. Today, there are various ways of monitoring activity in the living brain. Each of them has its advantages and drawbacks[1]. Since we are concerned with the recognition of words, where the duration of the articulation of phonemes that make up words happens in splits of seconds, only two brain monitoring techniques are possible candidates due to their sufficient temporal resolution : magnetoencephalogram (MEG) and electroencephalogram (EEG). With both techniques events with time scales in the order of milliseconds can be resolved ([16]) as opposed to imaging methods like FMRI that smear their measurements over time intervals in the range of 2-10 seconds

[1]For a tabular overview, refer to [40] or [16].

[40]. Even though MEG has better spatial resolution, it has some serious drawbacks that also preclude it from being used in our case. First, it cannot be used on all subjects since no metal on the patient is allowed (e.g. the patient must not have gold fillings, has to take off ear rings etc...). Second, it is a huge apparatus that needs to be operated in a specially shielded room. Third, MEG devices are very expensive to buy as opposed to relatively small costs for EEG recording devices. Energy consumption is also a point, which is on the one hand a financial issue, on the other could lead to restrictions on conceivable areas of application.

For these reasons, the decision fell on the use of electroencephalography, which is the oldest and most widely used method and a term that is unseparably connected to the German psychiatrist Hans Berger. He published the first report on recordings of the electrical activity in the human cortex [5] and authored the term electroencephalogram. He never claimed having discovered 'brain waves' and his work could build up on electrophysiological contributions of researchers like Hales, Galvani, Volta, v. Humboldt and certainly the findings of Caton [7]. Nevertheless, Berger can be considered the father of the clinical EEG-method which he developed and he was the first to posit and prove that the electrical activity that governs the brains of animals must also govern the human brain. When Berger's work was confirmed by other scientists in the mid-nineteen-thirties, the study of EEG and its application in clinical neurology began to blossom.

While his work produced applicable techniques, Berger's motivation was driven by the hope that the electroencephalogram would uncover the mysteries of the human organism's most intriguing function: the mind. In the beginning of the nineteenth century the idea of the vitalistic dualism that thought to strictly divide the human nature to matter and soul was omnipresent. Attempting to bridge the realm of immaterial and empirical objectivity was and is one of the major challenges of modern philosophy. Berger envisioned science and philosophy as one discipline that seeks a dialectic thread aimed to heal the mind-body schism. Very similar hopes may still remain to be an additional but driving factor in the continued pursuit of all research related to the cognitive sciences. They silently accompany all efforts along the way of fostering an understanding of the inner workings of the mind, even when the primary focus of the work in question is pointed at a concrete application, as it is the case in this thesis.

1.2 Some biomedical background

In this section some basic biomedical background is provided, to enable the reader to follow further discussions.It is purely focussed on information directly relevant to the task at hand. Much of the information presented here was taken from [35], [21],[34] and can be found there for a more thorough review. A brief but accessible and highly relevant first overview can be found in [16].

1.2.1 Anatomy of the brain - a brief overview

The brain consists of multiple components, namely cerebrum, cerebellum and brain stem.The cerebrum itself is split into two hemispheres (cerebral hemispheres) by a characteristic longitudinal fissure across which there is a large connective band of fibres called corpus callosum. The cerebrum is the largest part of the human brain by volume, surface area and weight. It has a heavily wrinkled surface which gives the exterior of the brain the widely known walnut appearance. The wrinkles are important because they increase the overall surface area of the brain, without the evolutionary need of evolving a larger skull. Around $\frac{2}{3}$ of the surface area is hidden in the recesses of the wrinkles. The cerebral hemispheres are enveloped by a thin (2-3 mm thick) layer of gray matter, called cerebral cortex (or better known as neocortex). That gray tissue has a higher concentration of neuronal cell bodies[2] than the rest of the white brain tissue and is predominantly populated by pyramid cells. Aside from the longitudinal fissure there are two more especially deep ones: The central fissure divides each cerebral hemisphere into an anterior and a posterior region. The lateral fissure that courses along the side of the cerebrum is the other one. These major fissures serve as landmarks to identify the different divisions of the cortex, since their relative position is approximately the same among individuals. Beneath the cortex, nerve fibres lead into other parts of brain and body.

1.2.2 Specialized regions of the neocortex

One subject of discussion among neurologists and scholars of related fields has been the question of whether information processing in the brain occurs in a localized or completely

[2]The density of neurons is estimated to be around 10000 $\frac{neurons}{mm^3}$. The cortical area is about $2.3m^2$ and the total number of neural cells is around 10^9[36].

distributed manner. According to present knowledge both views are considered valid to some degree: on the one hand, the neural pathways are highly interconnected, while on the other, several areas of the brain that are assumed to be predominantly responsible for certain cognitive tasks have been identified. The assignment of those tasks to the respective brain areas has been done by surgery (i.e. patients with damaged or surgically removed brain tissue of the areas in question experienced certain disabilities or electrical stimulation of a specific region caused involuntary, repeatable reactions by the patients) or by today's spatially improved brain mapping methods [40] (i.e. the execution of certain mental tasks where accompanied by measurements that where able to identify areas of elevated activity in the test subject's brain).

Remark 1.2.1. Sometimes results of different studies regarding the question of localized processing of mental tasks lead to opposite results. One conceivable explanation of such contradictions lies in the difficulty of exclusively capturing the desired task during recordings, because the experimental fullfilling of a task involves the parallel execution of other tasks at the same time. For instance, if one monitors the brain activity in different regions while reading, it can be expected that brain parts responsible for vision, attention control,... are also highly active at the same time. Sometimes experimenters have used different monitoring techniques a circumstance that could also help explaining the disparate results. In general, one should always be aware of the hypothetical character of the current knowledge of the complicated inner workings that embody the human brain.

The precise locations and sizes of the identified brain areas may vary slightly from person to person, yet for human beings with healthy brains those differences are relatively small.[3] A typical assignment of areas for the cortical part of the brain is depicted in Figure 1.1. The reason why we focus on the neocortex is that all higher functions of the brain, including speech processing, are considered to take place there predominately. We now turn to some evidence on the participation of selected cortical regions in speech and language.

[3]The differences between the interconnections and activities among different genders are reported to be of higher significance, especially for tasks like spatial orientation and speech (see [14], for example).

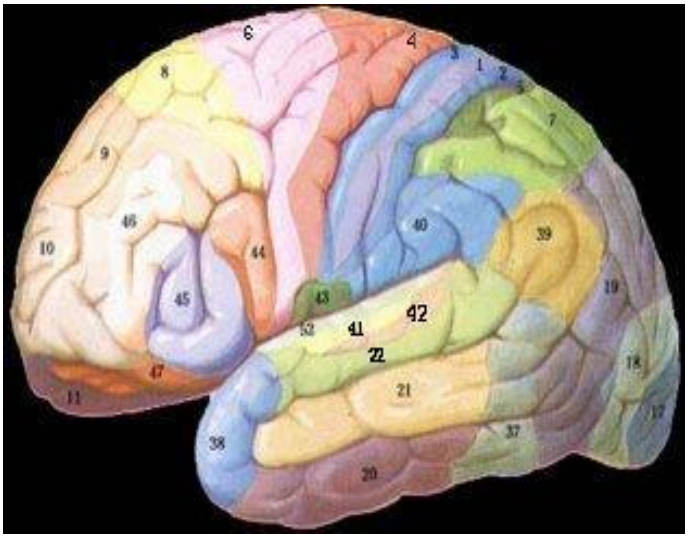

Figure 1.1: Side view of the left hemisphere of a typical brain. Various regions of identified areas are highlighted (so-called Brodmann areas) (modified from [41]). The four here described areas Broca's area, Wernicke's area, the premotor cortex and the primary motor cortex occupy Brodmann areas 44, {22 and 42 and 41},6 , 4 , respectively.

1.2.3 Cortical areas possibly specialized in speech production

For the production of speech there exist quite diverse models hypothesizing the flow of information through the different brain areas. These models are not subject to discussion in this subsection[4], instead three of the most prominent areas widely considered being involved in the production of speech, are briefly introduced.

Broca's Area

Broca's area was named after the French surgeon Paul Broca. He linked this cortical region to the production of articulate, fluent speech while examining a brain-damaged patient with a very limited speech vocabulary (he could only say the word "tan"). Later experiments using blood flow measurement techniques led to contradictory results [25],[20]. Still Broca's area is considered to be an important area in the planning and decision process for motor functions. It's role in speech production remains uncertain, although there are several strong indications that it is involved in the generation of words [31].

[4]For details on such issues, refer to [21],[34].

Wernicke's Area

According to the classic opinion *Wernicke's area* is responsible for speech comprehension and possibly the storage of the phonological representations of words. Some PET studies have come to contradictory conclusions once again. While one group reported increased blood flow during language related tasks, in [18] no single cortical region showed significant increase in blood flow while subjects either read words or repeated them. The authors concluded that the representation of semantic meaning of words is distributed over the entire cortex. In addition another group reasoned that during the recognition of words of different semantic categories different cortical regions would turn out to have an elevated level of activity.

Remark 1.2.2. This finding may be seen to be supported by the successful classification of the semantic category of words based on fMRI images[30]. In that work the classification was done directly on FMRI 3D-images as input. It appears reasonable to assume, the classification hardly would have been able to be successful if the semantic discrimination processes would have taken place exclusively in the relatively small Werknicke's area[5].

Areas for motor control

The production of audible speech involves a very complex sequence of facial muscle activation (e.g. lips,tongue,...). Therefore, two cortical regions which might be of interest for brain wave based speech recognition are believed to be highly responsible for steering muscle movement: the premotor area and the primary motor cortex.
Diverse studies of cerebral blood flow indicate an involvement of the *premotor area* in control, planning and regulation of complex, sequential and also learned movements. In another study, increased blood flow was detected in the premotor area during verbal and nonverbal naming tasks [21].
The *primary motor cortex ("motor strip")* is considered as the main area involved in the voluntary production of movements and is said to work together with other regions like the premotor area. The motor strip stretches alongside the anterior side of the central fissure (see Figure 1.1). Following the approximately symmetric construction of the brain,

[5]Note, fMRI measures the ratio of concentration of oxygenated and deoxygenated hemoglobin in each voxel, smeared over a time interval of 2-10 seconds. It is commonly assumed this ratio is caused by elevated activity of the according set of neural cells.

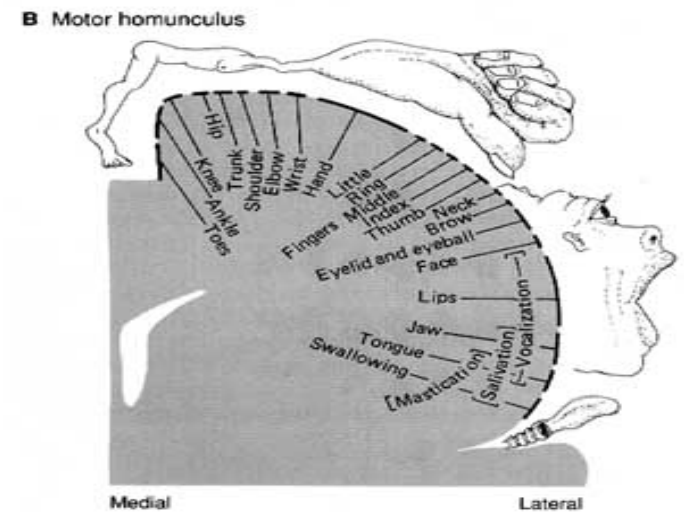

Figure 1.2: One hemisphere of the primary motor cortex. The size of the body parts of the homunculus creature reflect the size of the representation of the corresponding areas in the primary motor cortex. Note, the regions responsible for facial muscle activation are located in the lower part of the motor strip. (modified from [28])

there is a strip on the left, as well as on the right side of the brain, each responsible for the contralateral side of the body. Each strip can be divided into distinct sections responsible for the innervation of the muscles of the different body parts. The proportion of these sections assigned to the various body parts is not proportional to the size of the parts, but rather to the apparent need for precise motor information. Figure 1.2 shows a common way to depict these assigned regions of the motor strip using the homunculus. The size of it's body parts reflect the just mentioned proportionality. Therefore, fingers, tongue and lips have a disproportionate amount of allocated area of the motor strip considering their relatively small size. Hence, the lower part of the motor strip which is responsible for these parts (called *orofacial motor cortex*) is relatively large compared to the rest of the cortex, also due to the need of execution of complicated movements during speech vocalization. This circumstance could prove being helpful when placing scalp electrodes over that area of interest, especially given the poor spatial resolution of scalp EEG.

1.2.4 Neurons

All body tissue is composed of cells. The nervous system contains special cells, called neurons, which are considered as basic information processing units in the brain. They have various shapes and sizes but the structure of a typical neuron is depicted in Figure 1.3. The cell contains the nucleus and has a vast number of projections that are the means of communicating with other neurons, as well as other tissues such as muscles. The fine projections directly attached to the cells are called dendrites and conduct nerve impulses from other cells in the direction of the cell body. The main outgoing projection is the axon leaving the cell body has the objective of transporting signals from the neuron to other cells. It branches into smaller arms that eventually may connect to the dendrites of neighboring neurons via "bio-chemical adapters" called synapses.

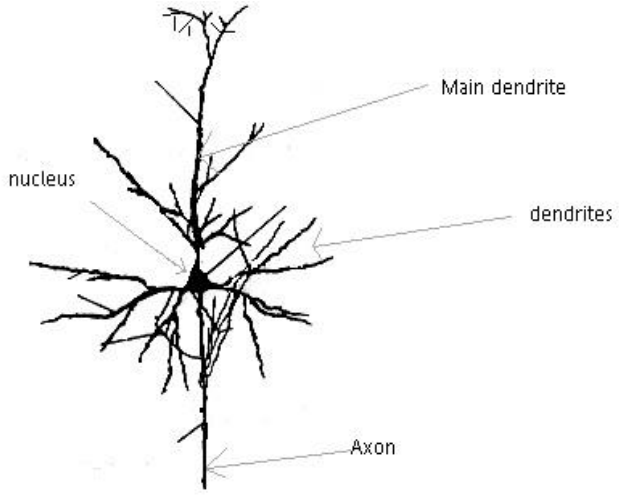

Figure 1.3: Structure of a typical neuron (here a pyramid cell). With modifications from [12].

In the human brain, a single neuron can be connected to thousands of other neurons and estimations of their total number in the human brain range from 10^9 up to $2 \cdot 10^{13}$, all of which are highly interconnected. Three basic types of neurons have been identified:

1) Sensory neurons : Conduct nerve impulses from a sensory receptor site into the brain or the spinal cord (afferent direction of the signal).

2) Motor neurons : Transfer signals from the brain to exterior parts of the body like glands and muscles (called efferent direction of signal travel).

3) Interneurons: Constitute the mass of the brain and spinal cord. They are held responsible for the intermediate information processing. They lie in the inner layers of the brain/spinal cord and exclusively connect to other neurons.

1.3 About electroencephalography

As mentioned in the beginning of this chapter solely electroencephalography[6] had the high enough temporal resolution while being portable and economically priced. These factors led to its selection from the available techniques of monitoring brain activity. It is also one of the most direct ways of measurement, since electrical excitement is the immediate nature of neural activity, as will become clear in subsection 1.3.2. On the other hand, one had to live with the drawbacks of EEG; Perhaps the most striking is the poor spatial resolution one has to live with when using scalp EEG. A more thorough treatment of this particular issue will be provided in subsection 1.3.3. Some other problems accompanying scalp EEG are mentioned in sequel.

1.3.1 Acquisition and properties

EEG is conventionally described consisting of patterns within four frequency bands, named δ ($<$ 4 Hz),θ ($4-7$ Hz),α (8-12 Hz), β (13-35 Hz) and γ (35- ca. Hz). Scalp recorded EEG in the waking state of the healthy brain typically reaches amplitudes of up to $75\mu V$ but can reach levels as high as $1mV$ in pathological cases like epileptic seizures [40]. The harmonic composition of the signals is usually complex and very rarely

[6]In this thesis electroencephalography is only meant in the context of scalp EEG, i.e. EEG obtained by placing electrodes on the scalp's surface. The other option, cortical EEG (i.e. taking measurements by surgically placing electrodes directly in the cortex bark), while having a fairly good spatial resolution, is certainly of no relevance to this work due to obvious practical considerations.

approaches a sinusoidal form.

To measure the EEG extracranially, electrodes are attached to the scalp with a conducting gel. Each electrode conducts the potential in the surrounding head area to an assigned channel of an amplifier where it is usually exposed to a low pass filter prior amplification. In most modern settings the signal is appropriately sampled and the digitized data is then forwarded to a computer for further processing depending on the individual application. The signal in each channel corresponds to the acute potential difference between the skin under the corresponding electrode and a reference.

The nature of this reference can be manifold. For instance, with differential derivation the reference of an electrode is the measurement of another electrode at every instant in time. However, in this work common reference derivation was used,i.e. all channels measure the potential in relation to the potential of one or the average of several common reference electrodes. The latter are advised to be placed at a position of the test subjects body that is far away enough from the neocortex so that none of its activity is captured. Also, the place should unlikely be influenced by artifacts. Those undesired disturbances can easily contaminate the EEG signal and may origin from various sources, including muscle movements, heart beats, poor electrode contacts, electrical interferences from electronic or electric devices near the recordings. For this work, as well as in [43], the ear lobes were used as references points.

Some characteristics of the signal may vary between subjects and even between sessions. In addition, brain signals are known to vary with age, gender, handedness, alertness, fatigue, habituation, level of autonomic arousal, use of alcohol, caffeine, drugs or consumption of nicotine [40].

The pattern of the electrode placements (montage) is in principle not restricted, although in most cases the widely accepted 10-20 system is used (see [47] for the technical details). The 10-20 system, as shown in Figure 1.6, is based upon spatial measurements from four standard points (nasion, inion, left and right pre-auricular points) on the head which can be easily found on any healthy person's head.

Starting from these four points the 10-20 systems describes a measurement procedure to locate standardized electrode positions whose locations relative to the brain is known. Of course, these positions can be filled up by additional electrodes but they also serve as good landmarks on the skull. For more details, refer to [47] or [16].

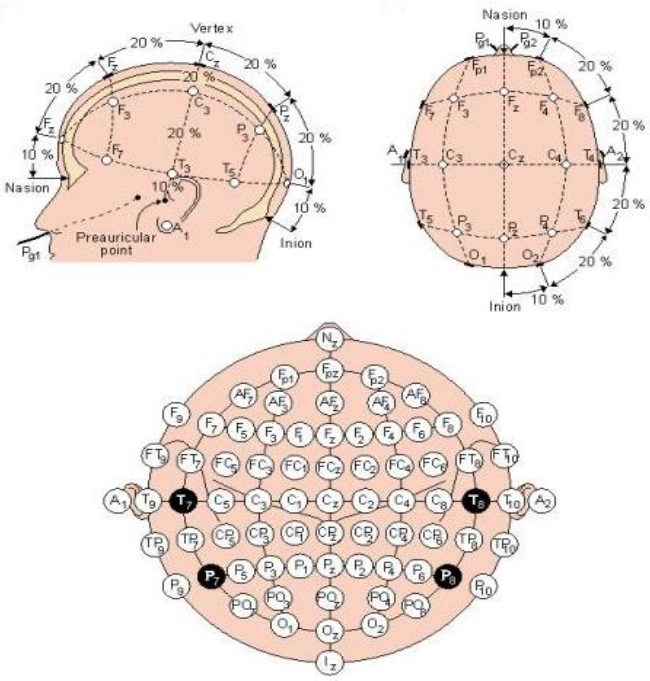

Figure 1.4: The standard 10-20 layout. (Picture taken from [45].

1.3.2 Electrophysiological explanation of the origin of EEG

The most popular and widely accepted theory regarding the origin of EEG signals is based on modeling cortical pyramid cells as dipoles. Pyramid cells which are estimated to constitute about 80 % of all cortical cells are neurons with a pyramid like shaped cell body and have the important property of having some very long main dendrites pointing in the direction of the pyramid's axis of symmetry. A neuron can be in a stable state (it does not emit a signal) or in an excited state (meaning it emits a signal along its axon). Each ingoing dendrite contains a certain electrical potential that influences the potential on the nucleus' exterior cell membrane. If that total potential exceeds a certain threshold, the neuron becomes excited and starts to spontaneously emit a signal. Depending on their contribution to the total potential the incoming potentials coming from the dendrites are classified as either excitatory or inhibitory. The potentials that encourage the excitement of a neuron are called EPSPs (**e**xcitatory **p**ost- **s**ynaptic **p**otential), while the potentials with opposite sign, which hence have a "calming" effect on the likelihood of excitement of the neuron, are called IPSPs (**i**nhibitory **p**ost- **s**ynaptic **p**otential). IPSP's usually have

a lower voltage than EPSP's.

The information transmission along a neuron's axon is controlled by the frequency of the emitted signal caused by electrochemical reactions. The signal travels down the axon until it reaches the synapses. The synapses "sense" the incoming signal and start sending positively charged ions from the end of the axon over to the dendrites of the connected neurons that have, depending on the kind of ions sent, either an excitatory (i.e. it generates an EPSP in the dendrite of the linked neuron) or an inhibitory (i.e. an IPSP in the dendrite of the linked neuron is caused) effect. The magnitude of that effect is regulated by the frequency in the axon (which is proportional to the rate of ions emitted in the according synapses).

The EEG recorded on the scalp surface is believed to be caused be EPSP's and IPSP's. An EPSP is caused when at the pre-synaptic membrane a larger quantity of positive ions moves into the post-synaptic dendrites and from there in the corresponding neuron. This makes the exterior part of this membrane segment relatively negative due to the lack of positive ions) compared to all other parts of the membrane of the same neuron where the number of positive ions can even increase due to capacitive effects. Therefore, the potential in the remote sub-synaptic membrane segment is comparably low to the potential in the neuron and its other membrane segments. Thus, if such a potential difference occurs the situation can be modeled by a dipole (Figure 1.5). An analogue process happens for IPSP's.

The resulting voltage of a single dipole is too small to be detected on the scalp surface. Instead it takes the synchronous excitement of a well localized bundle of a few thousand neurons, such that the potential differences between the point on the scalp above these neurons and a constant reference point can be registered which are commonly referred to as EEG.

Unfortunately the requirement that many pyramidal neurons are excited in concert is not yet sufficient to guarantee detectability. Let S be a set of such neural dipoles that are sufficiently close to each other and are synchronously excited at a given time instant. If one models a single dipole $s \in S$ as a vector v_s with a Euclidian length $\|v_s\|$ directly proportional to the potential difference between axon and the rest of the neuron, then the measurable influence of S to the EEG at scalp point P is proportional to the sine of the angle between $v := \sum_{s \in S} v_s$ and the vector perpendicular to the tangent on the

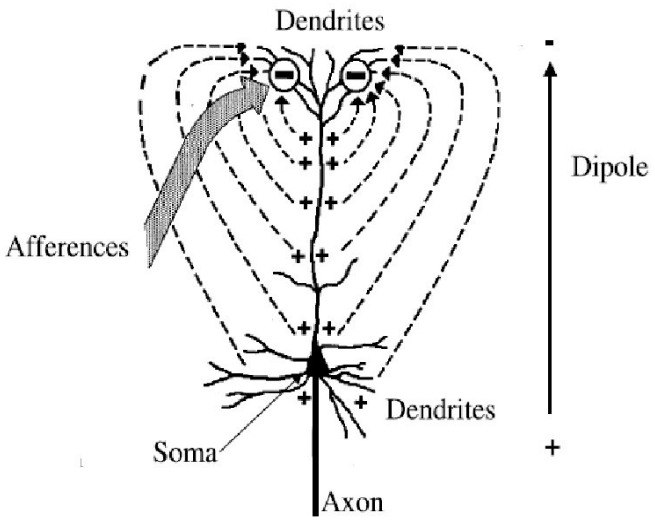

Figure 1.5: Pyramid cell as a dipole (from [16]. In this context, afference refers to information flow directed towards the neuron.

scalp going through P. It is believed that one scalp electrode records electrical currents generated in cortical tissue containing approximately 30-500 million neurons.

As a consequence, a neuronal dipole contributes to the measured EEG the more it is perpendicular aligned to the surface of the scalp and only if there are no close-by synchronous dipoles that eliminate it's influence by superposition due to having opposite directions. It is estimated that about $\frac{1}{3}$ of the cortical pyramid cells point in a direction perpendicular to the scalp. At this point it shall be remarked that three properties of pyramid cells encourage the belief that they are primarily responsible for EEG. First, they constitute about 85 % of the neocortex and hence may have the biggest impact purely by chance. Yet, a more sophisticated argument lies in the before mentioned fact, that if we choose to follow the dipole theory, the long, straight main dendrites of pyramid cells translate to better dipole properties that in effect result in higher voltages, when compared to other types of neurons. Finally, a large number of the pyramid cells axies are aligned nearly in parallel.

In summary, one has to be aware of the fact that an EEG signal measured to be around zero does not necessarily mean there is no activity underneath the corresponding electrodes' surface. For the same reason the measurement of a certain amplitude is not necessarily an exhaustive reflection of the actual activity.

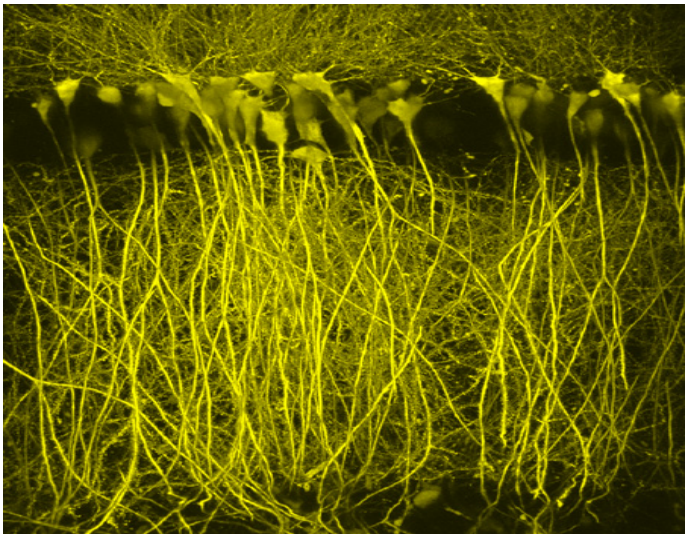

Figure 1.6: Image of a collection of pyramid neurons in the cortex of a rat (taken from [44]).

1.3.3 Spatial Resolution

As mentioned above, when EEG was selected as the monitoring method of choice this was done at the expense of spatial resolution. That is, the origin of the signal cannot be precisely localized when using scalp electrodes. In essence, this is due to the electrical properties of the head's tissue.

First, due to the conductivity of brain tissue and the head's liquids a superposition of potentials generated in different areas of the cortex is measured using scalp electrodes (Figure 1.7).

Second, the amplitude of the originally generated potential differences is attenuated because of the resistive properties of the tissue between the potential generators and the electrode. Finally, capacities caused by cell membranes and other inhomogeneities between potential generators and electrodes attenuate EEG of different frequency to a different degree.

More recent developments have led to sophisticated models and techniques regarding the localizations of evoked-potentials and other EEG-source localizations. However, the treatment of this is beyond the scope of this work. There are various approaches. For some input regarding these topics, refer to [37], [46], [8] and [13].

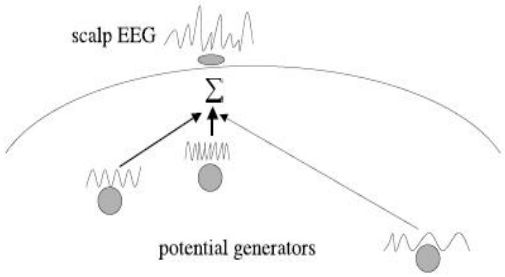

Figure 1.7: Volume conduction in the brain. Bold lines indicate a stronger influence on the measured signal, since the distance from the corresponding potential generator to the scalp electrode and thus the total resistance is smaller (from [16]).

Chapter 2

Extracting the signal of interest

2.1 Introduction

As will be explained in Chapter 3, the test subject was asked to mark the beginning as well as the end of the utterance of the word in question each with exactly one eyeblink. Hence, the signal in between these blinks corresponds to the brain waves during the utterance and is the signal we are interested in, which will be referred to as the signal of interest from now on.

The eyeblinks were chosen as markers for multiple reasons:

First, the facial muscle movements involved are of a higher magnitude than the background EEG and hence can easily be identified in the signal picked up by the electrode mounted just above the eye lid.

Second, every human being constantly feels the urge to close and open his eyelids from time to time to prevent the eyeballs from getting dry. Therefore it is a good idea to ask the test person to execute an eyeblink immediately before the short utterance since then the likelihood of the test person feeling the need to blink during the utterance is kept minimal. If the test subject would blink during the utterance the signal of interest would become heavily distorted which would have to be compensated by the application of rather sophisticated artifact removal techniques like ICA [1] for instance.

Third, the signal of the eyeblink has a very characteristic shape that allows us to discriminate it from other muscular activity with ease. Take a look at the signal depicted

[1] see [24] , [29] , [9], [16]

in Figure 2.1. This is an example for a very desirable signal in the sense that it contains no artifacts apart from the two intended blinks that can be easily spotted approximately between samples 2000-2250 and 2750-3000 respectively.

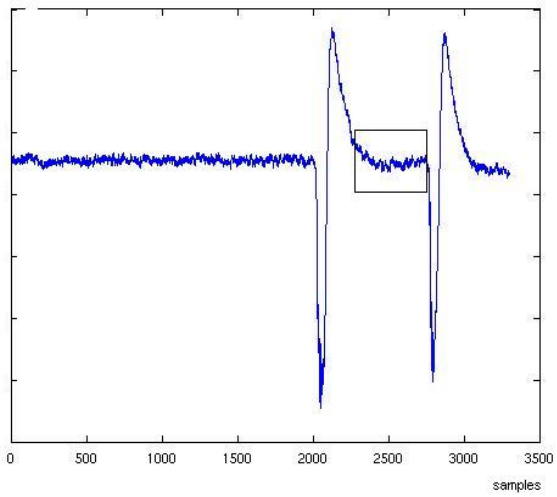

Figure 2.1: A good signal. The first eyeblink is the signal approximately between samples 2000 and 2250. The second blink can be observed roughly between the 2750th and the 3000th sample. The signal of interest is known to lie between these blinks and is highlighted by a box.

In [43] there was designed and implemented a procedure that is able to extract the signal of interest fairly well for the *unspoken speech* modality, provided each signal only contains the two purposefully inserted eyeblink artifacts.
Unfortunately many recorded signals cannot be guaranteed to meet all the requirements necessary to ensure that the just mentioned procedure is able to extract the right signal of interest. Many signals contain other artifacts that might mislead this procedure for various reasons, as can be seen in Figure 2.2
Having to strictly avoid the superimposition with that kind of unwanted artifacts by means of repetition can be tedious for both the test subject and the supervisor and unnecessarily prolong the recordings. That usually increases the number of recorded signals which then need to be repeated even further due to effects of mental and physical fatigue on the side of the participants of the recording sessions, since each session typically requires several hours even when everything goes smoothly. Also, the procedure used in [43] had problems with extracting the signal of interest for the modality *speaking* because of EMG-effects

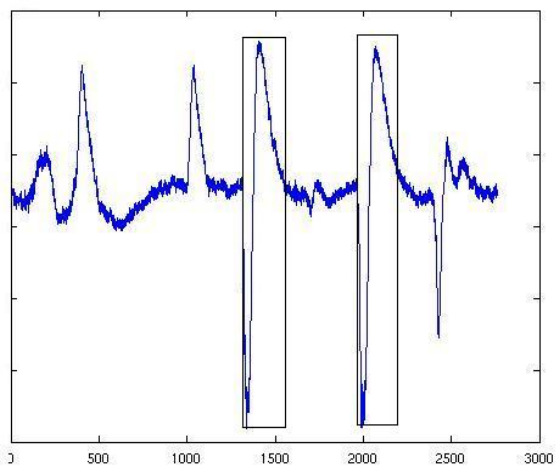

Figure 2.2: An example of a recorded signal heavily superimposed with undesired artifacts. The introduction of such artifacts can arise for various reasons, eg tiredness, dried out eyeballs, spontaneous muscle twitches, epileptic spikes, etc... Note that the two intentionally executed eyeblinks (framed by boxes) can still be easily discriminated from the other artifacts by visual inspection due to their unique shape.

(see discussion in Chapter 4) caused by the muscle use during the utterance of the audible words, spawing highly peaked signals of interest. Hence, extracting the signal of interest was not undertaken for that modality, which understandably could be seen as a drawback for principal reasons.

All of that caused the need to develop a new mechanism for the extraction of the signal of interest that was more robust and less dependent on idealistic assumptions. It shall be described and compared in greater detail in the subsequent sections after reviewing the procedure used in [43].

2.2 The initial extraction mechanism

2.2.1 A brief review of the algorithm

In this section the approach used in [43] for the extraction mechanism is briefly reviewed. After smoothing and uniformly adding a empirically determined constant *shift* to all samples to have the signal approximately centered around the abscissa, the following

were taken:
1) First, the algorithm takes an argument *offset*, stating from which sample no. on the signal will be examined. In effect, this is equivalent to cutting off the first *offset* samples from the signal prior further examination. The value of *offset* has to be predetermined and it relies on the user's experience with the procedure. 2) After that, the first 500 samples after *offset* are used as reference, i.e. their variance v is determined and later on, all following samples which have a magnitude suspiciously higher than v are considered possible candidates for being eyeblink points.
3) All those samples with an absolute value has a magnitude higher than v are added to a list C of candidates.
4) Next, a manually fine-tuned approach, with once again empirically determined parameters, is used in order to attempt to find a sample point l on the second last position of the signal that is above the variance (again in a absolute value sense) v and a point r on the last of those segments which is found the same way.

Remark 2.2.1. It is assumed that the signal is built up in a way that these two final segments of samples, with magnitude higher than the variance v of the first samples, are the positive / the negative hump of the first / second blink.

5) The final step consists of searching the first sample *start* of a higher index number than l with a magnitude lower than v, as well as of finding the first sample *end* with a lower index number than r that has a magnitude below v.
start and *end* are considered the beginning and the end of the signal of interest, respectively.

2.2.2 Critical assessment

It is to be pointed out that the approach discussed in this section was quickly developed intended to work exclusively in the given, particular setting. However the signal has to meet some required assumptions in order to let the procedure work properly, as the reader was certainly able to note. This consequences certain problems when these required conditions are not met.
One should note, that the algorithm involves many constant parameters, whose values were found empirically. Obviously using fixed parameters is always problematic as soon as characteristics of the signal, which were observed until implementation, change.

The key parameter of the algorithm is the variance v, which is found as the variance of set S of the first 500 samples. The assumption involved here is, that these samples have the same variance as the signal of interest.

Now two things can go wrong here: The contamination of S by artifacts of high magnitude[2] would cause the inclusion of too much eyeblink samples. In the opposite case , i.e. in case that the variance measured in S is significantly below the variance of the signal of interest, the signal of interest would probably not be properly extracted at all since much of it will be added to C the list of candidates of blink samples. The latter can easily happen when applying the extraction algorithm to signals recorded in the *speaking* modality, where facial muscle activity is responsible for high magnitude samples within the signal of interest. This was the primary reason that in [43] for recordings in modalities with the risk of such muscle involvement, no extraction of the signal of interest was attempted at all.

Remark 2.2.2. It is remarkable that the classification results still were superior over the ones in the modality *unspoken speech*, even though the former contained a lot of presumably irrelevant information. This issue will be talked about in Chapter 4 again.

As stated in Remark 2.2.1 another difficulty arises as soon as the final intended blink that is supposed to mark the end of the signal of interest is followed by another artifact (e.g. due to muscle tension or just a third unintended blink)that has an amplitude higher than the variance v. In practise this situation occurs quite frequently, especially when the participants of the recordings start getting tired. In summary, it appears to be doubtful wether the variance alone is a good indicator of the existence of an eyeblink. Take a look at Figure 2.3. It depicts such a situation where an additional, third blink occurred after the second blink, intended to mark the end of the utterance, was executed. It is interesting to observe that the shape of the third, spontaneous blink is quite different from the two intentionally executed ones. The explanation for that difference in shape is that the test subject fought against the urge to blink during its execution. This circumstance may give motivation to the newly developed cross-correlation based extraction mechanism, described in the subsequent sections.

[2]For an example of this situation refer to Figure 2.2.

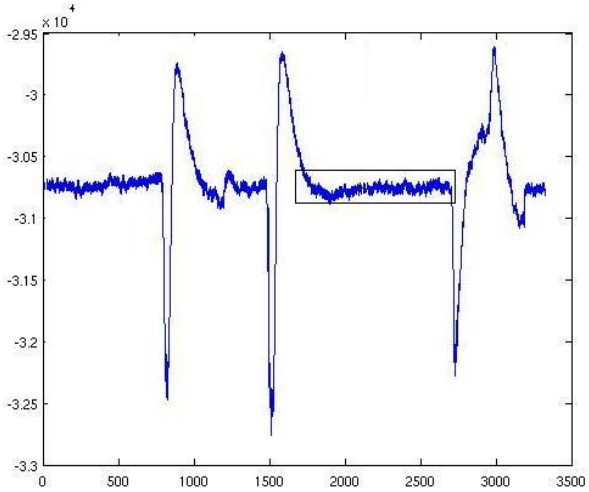

Figure 2.3: Two eyeblinks that are supposed to mark the start and end of the signal of interest are followed by an additional, unintended blink. Although the third blink is distorted due to its unintended execution and can be easily picked out, it contains numerous points that exceed the variance v no different than the other blinks do. Hence the extraction mechanism falsely declares the signal of interest to lie between bounds as indicated by the box.

2.3 An alternative approach based on cross-correlation

2.3.1 Description of the algorithm

The central part of the extraction algorithm is the localization of the two blinks, marking the start and end of the uttering process. Recall from the introduction that each of them is supposed to occur exactly once in the signal. As a consequence, the test subject is asked to blink exactly twice during the recording segment for one utterance. As we have seen our signal will still occasionally be exposed to undesired artifacts. On the bright side these artifacts have a different shape than the intentionally executed blinks and even the accidentally made ones tend to look quite distinct. Hence, if we manage to find a mapping M which computes a degree of similarity to the "typical blink"[3], we

[3]The question if a typical blink exists can be answered with yes. It turned out that all blinks of a test subjects look pretty much alike, even among independent sessions. While blinks of different persons seem to look slightly different, they appear to be still highly correlated.Hence, it is possible to use a single prototype blink with the signals of different subjects.

could search for the two fragments in the signal that are assigned the highest value by that mapping. Fortunately such a mapping can be realized by cross-correlation. A more thorough discussion of cross-correlation in the mathematical context will be provided in subsection 2.3.2. For now, we just give the definition:

Definition 2.3.1 (Cross-correlation)**.** Let $X := (x_i)_{i\in\mathbb{Z}}$, $Y := (y_i)_{i\in\mathbb{Z}}$ be two sequences of sampled signals. The cross-correlation cc of these sequences is given by $cc : (X, Y) \mapsto \sum_{i\in\mathbb{Z}} x_i y_i$.

Remark 2.3.2. (i)In practice we can expect to start with finite sequences of samples. As soon as we compute their cross-correlation we model them as infinite sequences $X : \mathbb{Z} \to \mathbb{R}$, $Y : \mathbb{Z} \to \mathbb{R}$ with finite support. This corresponds to concatenating sufficiently many zeros to the shorter of the finite, given ordered sets, so that the two sets have both the same length n. Then the cross-correlation can be computed as a finite sum of n component products.
(ii) An alternative way to define cross-correlation is to introduce a translation index b already in the definition. $cc_b : (X, Y) \mapsto \sum_{i\in\mathbb{Z}} x_i y_{i+b}$. In this case our definition becomes a special case, i.e. : $cc(X, Y) = cc_0(X, Y)$. Note, $cc(X, T_bY) = cc_{-b}(X, Y)$.

As a first step, we need to generate a prototypical blink. There are several ways to get such a blink. Since it was mentioned above that all blinks look pretty much alike, it is reasonable to cut out an arbitrary blink off a signal, normalize it and use the result sequence as the prototype P. Normalization of a signal F means the pointwise computation of $F = \frac{F-\mu(F)}{N(F-\mu(F))}$, where N is a suitable mapping, such as the $l_2(\mathbb{Z})$ norm. In practice $N(F - \mu(F)) = max|F - \mu(F)|$ works also very well.
The mapping M used as a similarity measure can now be defined as $M_P : (X, b) \mapsto cc(X, T_bP)$, where X is the normalized input sequence of samples, P is the sequence of samples of the prototype blink and T_b is the translation operator, defined by $T_b : (x_i)_{i\in\mathbb{Z}} \mapsto (x_{i-b})_{i\in\mathbb{Z}}$.
The additional parameter b defines the translation magnitude, i.e. the amount of shift in the direction of the abscissa of the prototype blink P. It can be used to localize the eyeblink in X, i.e. $s := argmax_{b\in\mathbb{Z}}\ M_P(X, b)$ marks the sample number of the start of the segment in X, having the same length as the support of P and the highest similarity to P. With that knowledge defining the rest of the extraction algorithm is easy:

1) Find $s_1 := argmax_{b\in\mathbb{Z}}\ M_P(X, b)$
(Refer to Figure 2.4, for an example.)

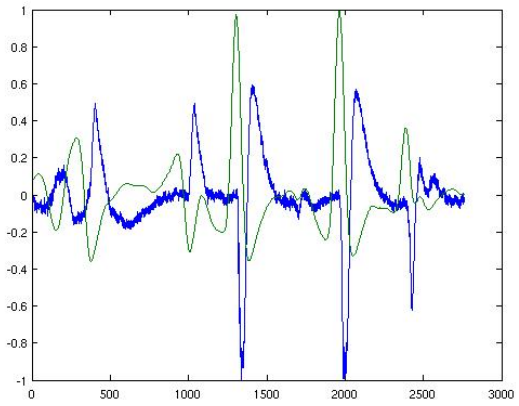

Figure 2.4: This image shows an example for the situation in the first step of the algorithm. X corresponds to the blue signal, while $M_P(X, b)$ is displayed as the green graph.

2) Erase the first blink from X that was detected and write the result in X':
$X'(t) := X(t) - \mathbf{1}_{[s,s+\ell]}(t) \cdot X(t)$, $\forall t \in \mathbb{Z}$, where ℓ denotes the length of $supp(P)$ and $\mathbf{1}$ is the indicator function, defined by $\mathbf{1}_S(t) = \begin{cases} 1, & t \in S \\ 0, & t \notin S. \end{cases}$, for any set S.

(Refer to Figure 2.5, for an example.)
3) Compute $s_2 := argmax_{b\in\mathbb{Z}}\ M_P(X', b)$.

(Refer to Figure 2.6, for an example.)
4) Determine, which of the values s_1 or s_2 corresponds to the beginning of the first blink and which corresponds to the start of the second blink and make sure that s_1 contains the number of the first sample of the first blink, while s_2 contains the number of the first sample of the second blink: If $(s_1 > s_2)$ then interchange the values of s_1 and s_2.

5) The first sample of the signal of interest is now estimated to be at $(s_1 + \ell)$ and its last at $(s_2 - 1)$. These two values are returned as the solution of the problem.

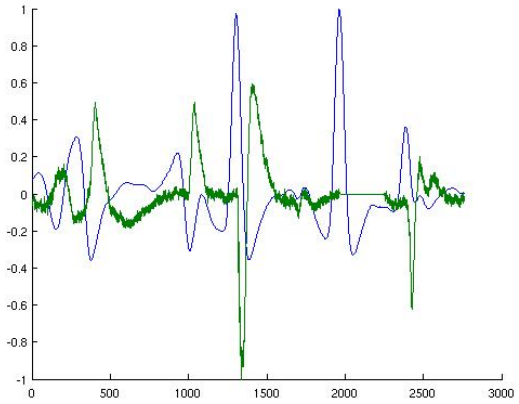

Figure 2.5: Here the signal was altered by the second step to become X' (green curve). Note, the second blink was replaced by a sequence of zeros, so it cannot be found again in the consecutive step.

Remark 2.3.3. If we were just concerned about the shape instead about both shape and scale of the pattern P that is to be detected, the algorithm could be altered in the following way: Instead of normalizing both the signal X and the pattern P once at the beginning of the algorithm we could instead leave them as they were and just replace the definition of the similarity measure by: $M_P(X,b) :=< \frac{(X-\mu(X))\mathbf{1}_{supp(P_b)}}{\left\|(X-\mu(X))\mathbf{1}_{supp(P_b)}\right\|_{l_2}}, \frac{T_bP-\mu(P)}{\|T_bP-\mu(P)\|_{l_2}} >_{l_2}$. This altered algorithm has some advantages and disadvantages we will not discuss here. In our pattern recognition environment the original procedure was found to be as successful while involving considerably less computation. Therefore, the original version was used in this work. However, some special cases (and also application domains) can be constructed where the just introduced alternative definition is more successful.

2.3.2 Mathematical justification - why it works

The core concept that makes the in this work developed extraction algorithm work, is cross-correlation. This section is not only meant to review some mathematical terminology, but also to view cross-correlation in the light of statistical correlation coefficients as well as projections in Hilbert spaces, which both are well known concepts that allow to make it intuitively understandable why cross-correlation turns out to be effective in being a measure of similarity. Some basic mathematical background as it is typically taught in undergraduate courses is favorable, apart from that, some basic terms are briefly reviewed

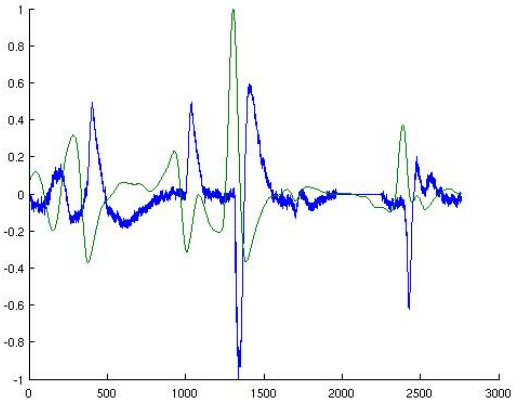

Figure 2.6: This image depicts an example for the situation in the third step of the algorithm, after the signal was altered to X'. X' corresponds to the blue signal, while $M_P(X', b)$ is displayed as the green graph.

[4] here.

Hilbert spaces

As a convention we focus our attention on complex or real vector spaces. Let $K \in \{\mathbb{C}, \mathbb{R}\}$.

Definition 2.3.4 (pre-Hilbert Space)**.** A normed K - vector space $(V, \|\cdot\|)$ is called pre-Hilbert space if there is a scalar product $< \cdot, \cdot >: V \times V \to K$ where $\|v\| = \sqrt{< v, v >}$, $\forall v \in V$.

Definition 2.3.5 (Hilbert Space)**.** A Hilbert space is a complete pre-Hilbert space.

Remark 2.3.6. (i) Completeness in $(V, \|\cdot\|)$ means here that every Cauchy sequence in V converges with a limit in V, with respect to the induced metric $d(x, y) := \|x - y\|$.
(ii) If S is a vector space with scalar product $< \cdot, \cdot >: S \times S \to K$, then the mapping $n : s \mapsto \sqrt{< s, s >}$ is a norm. Such a norm is called "induced by" or "associated to" the scalar product $< \cdot, \cdot >$.

In general it is advantageous to be in a position to model a given signal as an element of a Hilbert space, because this permits the use of the relatively advanced knowledge that Hilbert space theory has developed to further investigate signal properties.

[4] For a more detailed review of the functional analytic material it is recommended to refer to [2], [15] or [42]. The proofs that are omitted here can be found in these textbooks, too.

Modeling of signals

A well-known vector space in the domain of signal processing is $\mathfrak{L}_2(\mathbb{R}^d)$, the set of all square integrable complex functions[5] on $\mathbb{R}^d$.[6] In order to be able to apply mathematical tools to the recorded signals, one has to find a suitable mathematical model for that signal. It is very common to model signals as elements of $(L_2, \|\cdot\|_{L_2})$, where $\|\cdot\|_{L_2}$ is the norm on $L_2(\mathbb{R}^d)$ associated with the scalar product $<\cdot,\cdot>_{L_2}$: $(f,g) \mapsto \int_{\mathbb{R}^d} f\overline{g}d\mu$.[7] L_2 is a set of equivalence classes $\{[f]_\sim | f \in \mathfrak{L}_2(\mathbb{R}^d)\}$ on $\mathfrak{L}_2(\mathbb{R}^d)$, with respect to the equivalence relation $\sim$ defined by: $f \sim g :\Leftrightarrow \int_{\mathbb{R}^d}(f-g)d\mu = 0$. Hence, *f,g* are members of the same class in $L_2(\mathbb{R}^d)$, if and only if f and g are identical almost everywhere with respect to measure μ. The reason we have to rather work in $L_2(\mathbb{R}^d)$ than in $\mathfrak{L}_2$ is a formal, yet important one: $\|f\| := \sqrt{<f,f>} = \sqrt{\int_{\mathbb{R}^d} f\overline{f}d\mu} = \sqrt{\int_{\mathbb{R}^d}|f|^2 d\mu}, \forall f$ defines a norm on $L_2(\mathbb{R}^d)$ but not on $\mathfrak{L}_2(\mathbb{R}^d)$. This can be easily understood when considering an $z \in [0]_\sim$ which has a value equal to zero not everywhere but almost everywhere, i.e. there is a subset $N \subset \mathbb{R}^d$, where $(\forall n \in N, x \in (\mathbb{R}^d - N) : z(n) \neq 0 \wedge z(x) = 0)$ and $\mu(N) = 0$. Obviously $z \equiv 0_{\mathfrak{L}_2(\mathbb{R}^d)}$ but $\|z\| = 0$, hence $\|\cdot\|$ cannot be a norm on $\mathfrak{L}_2(\mathbb{R}^d)$.
Aside from a general, theoretical interest we do not have to be concerned with such issues. In this treatment, it is sufficient to model the input sequences of samples as discrete elements of another space, namely $l_2(\mathbb{Z})$.

Definition 2.3.7 (The Hilbert sequence space l_2)**.** Let S be an arbitrary set. The set of sequences

$$l_2(S) := \{X : S \to \mathbb{C} | \sum_{s\in S} |X(s)|^2 < \infty\} \tag{2.3.1}$$

is called Hilbert sequence space.

Remark 2.3.8. (i) On $l_2(S)$, the standard scalar product $<\cdot,\cdot>_{l_2(S)}$: $\begin{cases} l_2(S)\times l_2(S) \to \mathbb{C} \\ (X,Y) \mapsto \sum_{s\in S} X(s)\overline{Y}(s) \end{cases}$
induces a norm $\|\cdot\|_{l_2(S)} : x \mapsto \sum_{s\in S}|X(s)|^2$.
(ii) $(l_2(S), \|\cdot\|_{l_2(S)})$ is a Hilbert space.
(iii) It is interesting to note that for every Hilbert space H there exists a set S, such that

[5] $\mathfrak{L}_2(\mathbb{R}^d) := \{f : \mathbb{R}^d \to \mathbb{C} | \|f\|_{\mathfrak{L}_2} < \infty\}$

[6] $(\mathfrak{L}_2(\mathbb{R}^d), +, \cdot)$ is a vector space with respect to its operations $+, \cdot$ in a pointwise sense.

[7] With that induced norm $L_2(\mathbb{R}^d)$ forms a Hilbert space. Predominantly μ is set as the Lebesgue measure and the integral is a Lebesgue integral.

$H \cong l_2(S)$.
(iv) S countable $\Rightarrow$ $l_2(S)$ separable (i.e. it contains a countable, dense subset).
(v) For every Hilbert space H : H separable, iff it has a countable orthonormal basis.
(vi) From (iii) and (iv) one can conclude that $l_2(\mathbb{Z})$ has an orthonormal basis. A very simple one is $\{X_i : \mathbb{Z} \to \mathbb{C} | i \in \mathbb{Z}\}$,where $X_i(z) = \delta_{iz}$.[8]

As mentioned in the previous sections the input consists of finite sequences of samples of the continuous "real" signal in $L_2(\mathbb{R})$. These finite sequences can be canonically embedded as infinite sequences $X : \mathbb{Z} \to \mathbb{R}$ with finite support. Obviously such sequences are in turn elements of $l_2(\mathbb{Z})$ which is a Hilbert space. By this argument it is possible to understand the computation of the cross-correlation done in the algorithm developed in this chapter, as a projection on a vector of the Hilbert space $l_2(\mathbb{Z})$. This shall be done in sequel.

Cross-correlation in the light of projection

As a motivation, remember from Fourier analysis that the system of functions $\{\frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T}\}_{k \in \mathbb{Z}}$ is an orthonormal basis in $L_2[0, T]$. Therefore

$$\forall f \in L_2[0,T] : f \equiv \sum_{k \in \mathbb{Z}} < f, \frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T} >_{L_2} \frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T}. \tag{2.3.2}$$

If $f \in \mathfrak{L}_z[0, T]$ is continuous and T-periodic, the series converges pointwise.

The wide-spread, intuitive interpretation is to see f as a superposition of the basis waves $\frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T} (k \in \mathbb{Z})$. Each term $< f, \frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T} >_{L_2}$ can then be understood as a measure to what extent the corresponding wave $\frac{1}{\sqrt{T}} e^{i2\pi k \cdot /T}$ is contained in the signal f.
Another even more relevant example origins from wavelet theory:

Theorem 2.1. Let $\psi \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ be a Wavelet with wavelet constant $c_\psi = 2\pi \int_{\mathbb{R}_*} \frac{|F(\psi(\xi))|^2}{|\xi|} d\xi = 1$, $\forall x \in \mathbb{R} : x\psi(x) \in L_1(\mathbb{R})$, furthermore let $f \in L_2(\mathbb{R})$ be bounded and continuous, then the following limit converges pointwise for all $x \in \mathbb{R}$:

$$f(x) = \lim_{\varepsilon \to 0} \int_{|a| \geq \varepsilon} \int_{\mathbb{R}} < f, \frac{1}{\sqrt{c_\psi}} T_b D_a \psi >_{L_2} \frac{1}{\sqrt{|a|}} \psi(\frac{x-b}{a}) db \frac{da}{a^2}. \tag{2.3.3}$$

[8] δ denotes the Kronecker delta.

Remark 2.3.9. The operators T_b (translation by b) and D_a (dilation by factor a) are defined as $T_b\psi(x) = \psi(x-b)$ and $D_a\psi(x) = \frac{1}{\sqrt{|a|}}\psi(\frac{x}{a})$. F denotes the Fourier operator. For more details on wavelet theory it is advised to refer to [10]

If one recognizes integrals as infinitesimal sums it is again possible to perceive f as a superposition of functions $T_bD_a\psi$. One of the key differences to the case of Fourier series is the introduction of the translation parameter b that allows in combination with the good localization properties of wavelets (e.g. they can have compact support) the following observation : Each term $< f, \frac{1}{\sqrt{c_\psi}}T_bD_a\psi >_{L_2}$ can now be understood as a measure to what extent the corresponding wavelet $D_a\psi$is contained in the signal f around the abscissa point x = b. Therefore the projection $< f, \frac{1}{\sqrt{c_\psi}}T_bD_a\psi >_{L_2}$ of f onto $\frac{1}{\sqrt{c_\psi}}T_bD_a\psi$ contains not only information about that magnitude of containment, but because of the translation parameter b it also reflects the position of that magnitude in the signal.
In Section 2.3.1 a very similar thing was done for the case of l_2 signals by computing $M_P(X,b)$ for all translations b. Be reminded $M_P(X,b) = cc(X, T_bP)$. Assume X,P have zero mean (this can be guaranteed without loss of generality by subtracting the means if necessary). It is easy to link cross-correlation *cc* to projections on $l_2(\mathbb{Z})$ vectors by acknowledging
$cc(X, T_bP) =< X, T_bP >_{l_2}$. In order to be able to really compare the signals (the scalar product is sensitive to scale - we are more interested in shape instead), we had normalized prototype P and input signal X prior the computation (so the actual calculation was: $M_P(X,b) =< \frac{X}{\|X\|_{l_2}}, T_b\frac{P}{\|P\|_{l_2}} >_{l_2})$. Let U_b denote the subspace of $l_2(\mathbb{Z})$ spanned by T_bP. From Hilbert space theory we know $l_2(\mathbb{Z}) = U_b \oplus U_b^\perp, \forall b \in \mathbb{Z}$. Hence, each computation of $M_P(X,b) =< X, T_bP >_{l_2}$ can be understood as an orthonormal projection of X on U_P and in effect on its basis vector T_bP. Due to normalization, apparently $X = T_bP$ if and only if $< X, T_bP >_{l_2}= 1$ [9]. The more the normalized vectors T_bP, X deviate, the smaller $M_P(X,b)$s value.[10] If we used the alternative definition of M_P as described in Remark 2.3.3, we get $X\mathbf{1}_{supp(P_b)} = T_bP$ if and only if $M_P(X,b) = 1$.

[9]Note: $M_P(X,b)$ is a mapping into the interval $[-1,1]$.

[10]In this context it might help to imagine two arrows on the unit circle. In this setting the scalar product equals the cosine of their angle.

Cross-correlation in the light of correlation coefficients

An alternative way of understanding the purposefulness of cross-correlation and hence the calculation of $M_P(X, b)$ can be found in the statistical context of correlation coefficients, provided one can reasonably model the signals involved in the computations as being generated by a Gaussian distribution. The latter requirement is may be questionable in our current application, however this point of view is still quite interesting and may foster an intuitive understanding of the cross-correlation based algorithm. However, for notational convention and fluent readability purposes, some statistical terminology is briefly reviewed first.

Definition 2.3.10. (Variance *var*). Let v be a discrete random variable which can take on the values $V := \{v_1, ..., v_m\}$.Let P_v be the probability density of v.

$$var(v) := E((v - \mu_v)^2) = \sum_{x \in V} (x - \mu_v)^2 P_v(x) \tag{2.3.4}$$

is named variance of v. (E denotes the expected value operator, μ_x denotes E(x).)

Remark 2.3.11. (Marginal density). Let r,v be two discrete random variables which can take on the values $R := \{r_1, ..., r_n\}$ and $V := \{v_1, ..., v_m\}$, respectively. Let P denote their joint probability density, P_v the so-called marginal density of v (i.e. the density of v alone). P_v can be obtained in the following way:

$$P_v(x) = \sum_{r \in R} P(r, x) \tag{2.3.5}$$

The calculation of the equation is called marginalization.

Definition 2.3.12. (Covariance σ_{rv}). Let r,v be two discrete random variables which can take on the values $R := \{r_1, ..., r_n\}$ and $V := \{v_1, ..., v_m\}$, respectively. Let P denote their joint probability density.

$$\sigma_{rv} := E((r - \mu_r)(v - \mu_v)) = \sum_{r \in R} \sum_{v \in V} (r - \mu_r)(v - \mu_v) P(r, v) \tag{2.3.6}$$

is called the covariance of r and v.

Remark 2.3.13. The covariance σ_{rv} is one measure of the degree of statistical dependence between r and v. Statistical independence of r and v implies $\sigma_{rv} = 0$. The reverse direction does not hold in general. If $\sigma_{rv} = 0$, r and v are said to be uncorrelated.

Definition 2.3.14. (Correlation coefficient ρ_{rv}). Let r,v be two discrete random variables which can take on the values $R := \{r_1, ..., r_n\}$ and $V := \{v_1, ..., v_m\}$, respectively. The correlation coefficient ρ_{rv} is defined by:

$$\rho_{rv} := \frac{\sigma_{rv}}{\sigma_r \sigma_v} \tag{2.3.7}$$

, where $\sigma_r = \sqrt{var(r)}, \sigma_v = \sqrt{var(v)}$.

Let again P be the Pattern we wish to detect in the signal X and M_P be defined as in Remark 2.3.3. So, how do we bridge the gap between computing $M_P(X, b)$ and correlation coefficients? To obtain such a construction, we assume the input signal X and the translated prototype T_bP were each generated by an underlying normally distributed, stochastic process. Let m be the number of samples contained in T_bP, assume $|suppT_bP|$ = m. Let $Y : t \mapsto \mathbf{1}_{suppT_bP}(t) \cdot X(t)$. Note, $M_P(X, b) = M_P(Y, b)$. Hence, we may continue the further treatment with Y, which will turn out to result in some simplifications of the involved formulae.

The random variable of Y will be denoted by y and the according one of T_bP by p. For Gaussian distributed samples the maximum likelihood estimate $\widehat{\mu}$ of expected value, $\widehat{var}$ of variance and $\widehat{\sigma_{yp}}$ of covariance is each given by: [11]

$$\widehat{\mu_y} := \frac{1}{m} \sum_{i=1}^{m} y_i \tag{2.3.8}$$

$$\widehat{\mu_p} := \frac{1}{m} \sum_{i=1}^{m} p_i \tag{2.3.9}$$

$$\widehat{var}(y) := \frac{1}{m} \sum_{i=1}^{m} (x_i - \widehat{\mu_y})^2 \tag{2.3.10}$$

$$\widehat{var}(p) := \frac{1}{m} \sum_{i=1}^{m} (p_i - \widehat{\mu_p})^2 \tag{2.3.11}$$

$$\widehat{\sigma_{yp}} := \frac{1}{m} \sum_{j=1} (y_j - \widehat{\mu_y})(p_j - \widehat{\mu_p}) \tag{2.3.12}$$

.

[11] See also [11].

With that model in mind, we can compute the correlation coefficient ρ of the joint stochastic process:

$\rho_{yp} = \frac{\widehat{\sigma_{yp}}}{\sqrt{\widehat{var}(y)}\sqrt{\widehat{var}(p)}}$.
With equations 2.3.8 - 2.3.12, we get :

$\rho_{yp} = \frac{\frac{1}{m}\sum_{j=1}(y_j - \widehat{\mu_y})(p_j - \widehat{\mu_p})}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - \widehat{\mu_y})^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(p_i - \widehat{\mu_p})^2}}$

$= {}^{12}\frac{\sqrt{m}\cdot\sqrt{m}\cdot < Y, T_b P >_{l_2}}{m \cdot 1 \cdot 1} =< Y, T_b P >_{l_2} = M_P(Y, b) = M_P(X, b).$
Hence, for the Gaussian case it can be observed that $M_P(X, b)$ is a maximum likelihood estimate of the correlation coefficient of the joint process of $Y, T_b P$. From this point of view finding $argmax_b M_P(X, b)$ means finding the beginning of a signal fragment in the normalized input signal that most statistically correlated to P and hence, most "similar".

2.3.3 Summary and final remarks of this Chapter

There is one point of criticism that has to be mentioned regarding the mechanism described in 2.3.1. It lies in the use of a fixed prototype blink. As was said before, this difficulty is a minor one, because all blinks seem to be very similar in shape. As far as the length of the blink is concerned (which translates to the support of the pattern that is to be found / the support of P), the temporal variation is very low,too, if the test subject is advised to blink just briefly, so the impact of the variations appear to be quite insignificant with the sampling rate of 300 Hz, used in the experiments described in this work. Still, it would be conceivable to adjust the algorithm in a way to make it more adaptive. Such efforts could be pursued in future work.
In summary, the algorithm just presented has proven to be enormously successful in robust detection of the bounds of the signal of interest, provided there are only two properly executed blinks visible in the signal. The verification of its ability to do so was done by testing over several sessions by means of visual inspection. One of such testing outputs is depicted in Figure 2.7 for our worst-case scenario. It surely is an improvement over the algorithm previously used with regard to reliability, accuracy and the fewer assumptions

[12] Remember,according to section 2.3.1, we start off with normalized versions of the signals in the algorithm.

it requires that must be guaranteed to be met by the signal (which translates to less effort during and after the recordings).

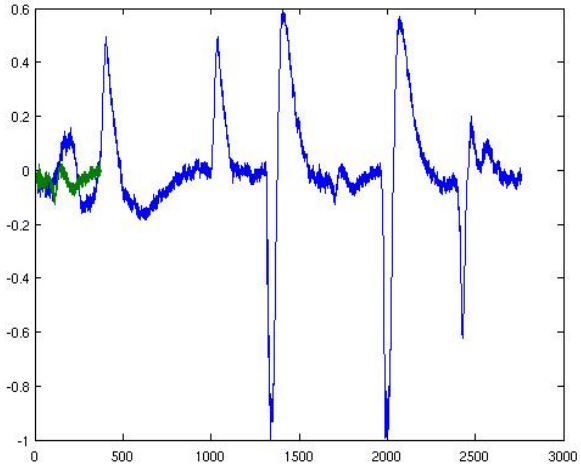

Figure 2.7: The green signal is what the algorithm estimates to be the signal of interest.

Furthermore, note the mechanism certainly can be used to find any patterns in a 1-dimensional signal and not just eyeblinks. This would only require replacing the prototype pattern P accordingly.

Chapter 3

Experiments

This work is a continuation of [43] designated to compare the our results with the previous results using the same methods and tools. Therefore, constant references to that thesis need to be made in this chapter. After the description of the task that was aimed to be fulfilled the following section outlines some of the technical aspects of the tools that were developed in the preceding work and used in this one,too. The experiments are then described and the results are presented. The final two sections are concerned with a critical discussion of the concepts and findings made up to this point, finalizing with a summary and an outlook of possible, future work.

3.1 Description of the task

The general aim intrinsic to the work in [43] was to adjust a speech recognizer and apply it on electroencephalographic data that was recorded during the utterance of words of a known vocabulary in the hope to be able to correctly classify them. The utterances were executed in different modalities. Among those are *speaking*, the normal vocalization of a word and *unspoken speech*. By the latter we mean that the test subject was asked to focus on the word while thinking it, as if they were uttering the word in the *speaking* modality. The results were promising in the sense that the cross-validation estimated, generalized classification error (around 50 %) was decidedly better than random guessing (80 % for the five word vocabulary). The idea guiding the current work was to reposition the available 16 electrodes from the standard configuration used before to a higher density over the

primary motor cortex, in order to apply the previously mentioned speech recognizer on the data acquired with the new setting. The underlying hope this current work sought to fulfill was that the neural-electric processes in the primary motor cortex could already contain enough information for successful recognition. It was encouraged by first findings of the preceding work where the training and testing with a subset of the of electrodes along the motor strip yielded an average recognition rate of 35.5 % on a vocabulary of five words for the modality *unspoken speech*, thereby being still 15.5 % better than random guessing [43].

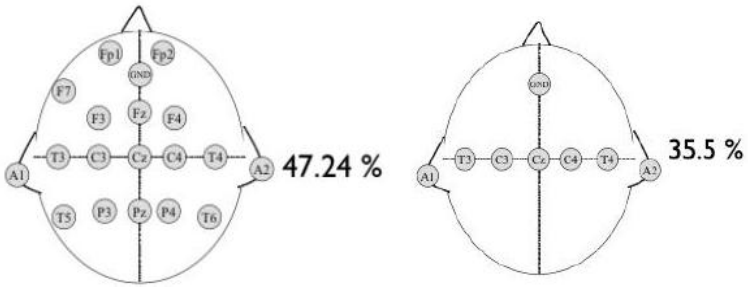

Figure 3.1: Left: Standard layout that was used in [43] with word recognition accuracy for *unspoken speech*. Right: Layout only containing a subset alongside the motor strip was used and obtained accuracy [43].

While pursuing the task a well performing algorithm for blink detection was developed and implemented in Matlab and Janus. It is quite thoroughly described in chapter 2.

3.2 The experimental procedure

The experimental procedure initially followed the protocol used in [43].

3.2.1 Hardware setup

For the most part, the hardware was the same as used in the previous work described in [43] and [17], where the only difference was the additional use of a new high density cap for some of the recording sessions, as shall be described.

Caps and electrodes

We used Ag/AgCl electrodes. The electrodes were attached to an elastic cap in fixed positions so that each electrode's position on the wearer's scalp was known [1]. The elasticity allowed slight variations of the test subjects head circumference and also established tight contact between head and electrode. Because the amplifier restricted us to 16 input channels, only 16 electrodes of a cap could be recorded.
The cap already being used in the previous work was a standard low-density cap with electrode positions selected in a standard way with compliance to the 10-20 system. The manufacturer was Electro-Cap International, Inc. The cap's electrode placement is depicted on the left side of Figure 3.1.
For the new cap, the choice eventually fell on a 128 - electrode, high-density cap produced by the same manufacturer. The layout is depicted in Figure 3.2.

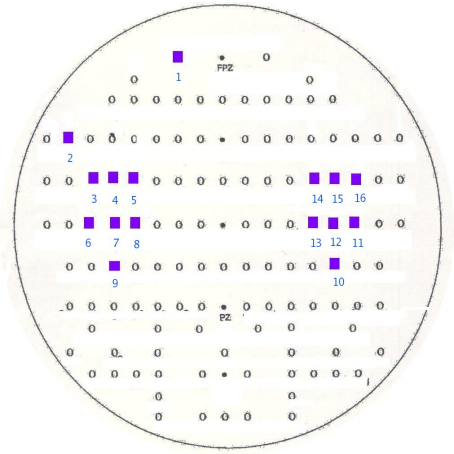

Figure 3.2: The layout of the new, high density cap. The subset of electrode positions actually recorded are marked by squares. Position 8 and 13 correspond to C_3 and C_4 (according to the 10-20 system), respectively. The position numbers equal the channel numbers the corresponding electrodes where connected to.

As you can see there, the selected subset of 16 electrodes was located around the orofacial motor cortex with higher density. The only two electrodes recording signals over other areas where one over the left eyebrow placed there for blink detection and the other one

[1] A prospective test subject's head had to be measured to guarantee the cap was suitable for the head size.

A/D Conversion	12 Bit
Input Range	$\pm 450 \cdot 10^{-6} V$
Resolution	0,22 V / Bit
Frequency Range	0,9 ... 60 Hz
Amplification Factor	2775
Input Channels	16

Table 3.1: Specifications of $VarioPort^{TM}$. For further details, refer to [4].

was placed in a position as far as possible from the motor strip while still picking up a great fraction of signals from Broca's area of the left cortical hemisphere [2].

Remark 3.2.1. As it turned out, the extensive wiring under the cap's surface due to the higher electrode density was sub-optimal. It prolonged the experiment's preparations because after the cap was put on the test subject's head, contact between head and various electrodes frequently tended to be disturbed by cable. For the future, we recommend caps with outside-wiring.
Also, the electrodes now were surrounded by plastic that had a different diameter than before. Therefore, the two electrodes above the eyebrows did not fit into the two sponge discs anymore that had to be glued at this position to keep the electrodes in that place and prevent the cap from contracting due to its elasticity. Since no fitting sponge discs could be ordered the problem was overcome by a make-shift solution that had to be set up before each session and prolonged the preparation time even more. It is to be hoped that sponge discs of the appropriate size will become available in the nearer future.

$VarioPort^{TM}$

The amplifier/ADC and recording device used in the experiments was the $VarioPort^{TM}$ [4]. The amplifier/ADC's specifications are listed in Table 3.1.
The device was linked to a laptop via an interface. The latter was connected to the device with a fiberglass cable in order to minimize interference. The interface was hooked up to one of the laptop's USB-ports with an RS-232 to USB adaptor.

[2]The left hemisphere was chosen because it reportedly seems to be dominant for speech production (especially with males) [31].

3.2.2 Selection of the vocabulary domain

While a variety of vocabulary domains were used in [43] [3] we only chose the *Alpha* vocabulary domain this time, i.e. a set consisting of the five words {*alpha, bravo, charlie, delta, echo*} was used being identical to the first five letters of the international radiotelephony spelling alphabet (NATO alphabet). Several reasons led to the selection of this corpus. First, NATO alphabet is designed to have words that are easily distinguished when spoken. Second, the words have not a familiar semantic meaning,i.e. the possibility that the test subject visualized the semantics behind the word on which we would do classification was limited. Third, they do not fall into different semantic categories. Since it was intended to do speech recognition on brain waves and not classification of different semantic categories this point is also important [4].

3.2.3 Data acquisition

The Setting

The recording sessions were done in quiet rooms in the evening. The participants were the test subject and a second person responsible for the recordings to whom will be consequently referred to as the supervisor. The test subject was sitting in a chair in front of a monitor. After the test subject was outfitted with the cap and the electrodes were filled with a conductive paste it was carefully instructed. The supervisor was sitting on another table in the same room controlling the recordings on a laptop which was attached to the CRT display the subject was facing. The screen showed instructions which the subject was asked to follow. If the subject or the supervisors noted any mistakes during an utterance, this single one was repeated. The actual data accumulation and online control of the experiments could be done on a laptop in front of the supervisor.
The subject was told he could quit the experiment without any consequences at any time and was also allowed to ask for as many breaks as he wanted. The possibility of interruptions or an abortion of the session was favored over the conceivable possibility that an unconcentrated or no longer willing test subject could produce misleading data.

[3] In [43] the word corpus was used for a vocabulary domain. However, in this work we follow the general convention of referring to a corpus as a collection of transcribed recorded data.

[4] For works, that explicitly use such effects, refer to [6] or [30].

The test subjects where all male, German native speakers for those recordings included in the analysis of this work

Remark 3.2.2. During the first recordings, a few sessions where done where the test subject was female and Japanese or English native speaker. However, these recordings could not be included in the analysis of the results because either for technical reasons, or the session was aborted based on the explicit wish of the test subject. The test subjects uttered the words in a German tone.

Corpus Production

Each recording session followed a certain protocol that was defined by a session's word list. By word list of length n, we will mean an ordered set of pairs $((w_i, m_i))_{i\in\{1,..,n\}}$, where each w_i is word in our vocabulary domain and m_i is a modality. The recording process was divided into several steps (utterances). In the i. step, the pair w_i, m_i was crossed off from the word list and displayed on the test subject's monitor for a few seconds. Then, the screen turned white indicating to the test subject that the actual recording of this utterance had begun. It then uttered the word displayed on the screen before, in the according modality. About 2 seconds before uttering the subject marked its beginning by a single blink with both eyes. Immediately after uttering the word the test subject marked that event by a second blink. After the supervisor saw this blink vanishing in the signal he stopped the utterance step and indicated this to the subject. In between these two blinks the subject was asked to remain still except for the facial muscle movements necessary for speech production in the *speaking* modality. The supervisor was also held to make sure the test subject did not blink more than exactly twice during each utterance step.
Since the word list controlled the procedure of the recording session its structure shall be described in narrower detail. We have tried recordings with different word list structures:
1) Following [43], in the first sessions the structure of the word list was built up in a way that all training examples of the same class were recorded directly back-to-back:
For the recording of two different modalities, the word list was set in a way, such that
$\forall i \in \{2, ..., 31\} : (w_i, m_i) = (alpha, speaking)$,
$\forall i \in \{33, ..., 62\} : (w_i, m_i) = (alpha, unspokenspeech)$,
$\forall i \in \{64, ..., 93\} : (w_i, m_i) = (bravo, speaking)$,

$\forall i \in \{95, ..., 124\} : (w_i, m_i) = (bravo, unspokenspeech)$,...
...,$\forall i \in \{281, ..., 310\} : (w_i, m_i) = (echo, unspokenspeech)$.
Hence, the structure of the word list determined a corpus that consisted of ten blocks containing each 30 consecutive identical (word,modality) pairs.[5] The type of the resulting corpus will be called *homogenous blocks 1*.
In some sessions, only one modality was used while still recording in a blockwise mode. In this case, the structure is similar to *homogenous blocks 1* the only difference being that one modality is missing and thus, two neighboring block consisting of recordings of utterances for different words of the same modality (in this case, the session contains only 5 blocks with each 30 words of homogenous word type). This type of corpus will be called *homogenous blocks 2*.

Remark 3.2.3. As will become clear shortly, it is important to point out that between each block the test subject was usually given a short break of several minutes (the duration was varying with the condition of the test subject and its willingness to quickly proceed with the recording of the utterance steps of the next block).

2) In later experimental sessions, two other word list compositions were used where we called the resulting corpora *mixed blocks(sequential)* and *mixed blocks (randomized)*. In both, the blocks of the corresponding were still composed of recordings of utterance steps for words uttered in the same modality. In the first case, the word structure was built up like (alpha,bravo,charlie,delta,echo,alpha,bravo,...). In the second case, the word list was completely randomized for each block of the same modality, subject to the constraint that in the end, every word had been uttered exactly 30 times.
3) For each corpus of type mixed blocks (randomized or sequential) an additional corpora was formed by relabeling the training data, such that the structure was as in the case of homogenous blocks. The structure of a homogenous block corpus was obtained by intentionally assigning all recordings in the first block (utterance step 2-31) the (in most of the cases wrong) label *alpha*, to all elements of the second block of the same modality the label *bravo*,... . Such an artificially relabeled corpus will be consequently called *relabeled*.

[5]Note, between the blocks a circa two seconds long record file was made that recorded the test subject silently, motionlessly sitting, being asked to think of nothing. This utterance step was marked on the word list for all $i \in 1 + 31\mathbb{Z}$ until the end of the word list.

Remark 3.2.4. One can say (especially with Remark 3.2.3 in mind), if two elements from a relabeled corpus have got the same label, it expresses rather a temporal closeness of the utterance steps (ie the time when they were uttered was close) than that they would contain any similar information about the class of the word that was uttered (i.e. whether alpha, bravo,..., etc was uttered).

3.3 The recognition procedure

In the preceding work [43] a HMM -based[6] speech recognizer was modified and used for training and classification on sampled electroencephalographic data. The implementation was done with the Janus recognition toolkit (*Jrtk*) ([27]). The technical aspects are not subject of this thesis[7], however, some elements of the used recognizer shall be mentioned: In a preprocessing step, for every sample i for each of the 16 input channels twelve features were extracted and concatenated to a 192 dimensional vector l_i. For the purpose of dimensionality reduction and feature selection linear discriminant analysis (LDA) was applied on the l_i. By doing so, feature vectors x_i with reduced dimension were obtained. The choice of the initial features being used to form l_i fell on ones commonly applied in the field of speech recognition, including the short time fourier transform power spectrum, delta, delta delta and delta mean coefficients. Most of the parameters involved were determined empirically, the same goes for the topology of the HMM. They were trained using Gaussian Mixture Models (GMM) for the output probabilities. The training and testing procedure followed cross-validation.

3.4 Results

The recognition results are listed in this section using the terminology introduced in 3.2.3. The sessions are denominated by a code of the form *test subject code_session number*. The first number has the purpose of anonymization while the second identifies the number of the session being done with the specific test subject. In the following, recognition rate

[6]HMM stands for Hidden Markov Model. For a first, general introduction to HMM's and their role in speech recognition, refer to [32].

[7]For details on the aspects of implementation, refer to [43]. Nice material of the theoretical background of speech recognition is provided in the textbooks [19] and [33].

is a term denoting the percentage of successfully recognized words estimated using cross-validation.
Recog. Rate 1 denotes the recognition rate of the recognizer applied on the data after preprocessing it with the eye blink detector used in [43], while *Recog. Rate 2* is the recognition rate achieved when the cross-correlation based algorithm developed in Chapter 2 was used for preprocessing.

Remark 3.4.1. The sessions 13_01,13_02 have been made on the same day without removing and reattaching the cap between the sessions. The same goes for 11_08 and 11_09.

Note, for the modality speaking, the results *Recog. Rate 1* and *Recog. Rate 2* are always equal. The reason for that is, it turned out that in the preprocessing implementation, no extraction of the signal of interest took place at all, i.e. the whole signal was used for training and classification. The reason for it could be found in the old eye blink detector's inability to distinguish blinks from other high-amplitude muscle movement artifacts which certainly occur during the utterance of vocalized speech (see Chapter2). It is interesting that the training still seemed to be especially successful with *speaking*.

Remark 3.4.2. As one can see in table 3.2 the recognition rate *Recog. Rate 1* is frequently better than the recognition rate achieved when we used the new cross-correlation based extraction algorithm for *unspoken speech* corpora. However, we consider the latter to deliver more accurate cut-outs of the signal of interest, so we chose to think the results more accurately reflect the actual information content regarding the utterance of the words themselves. This is why this recognition rate was chosen as a reference for the later experiments exhibited in Tables 3.3 - 3.8.
In addition, it shall be noted that the results on the data used in [43] in the vocabulary domain *alpha* where very often better when using the cross-correlation based extraction algorithm than when using the old one. On average, *Recog. Rate 1* was 36.8 %, while *Recog. Rate 2* was 38 % for the modality *unspoken speech* (with a dimensionality reduction onto 35 dimensions with LDA).

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking	86.2	86.2
11_02	high density	homogenous blocks 1	unspoken speech	73.7	74.8
11_03	high density	homogenous blocks 1	speaking	92.7	92.7
11_03	high density	homogenous blocks 1	unspoken speech	87.3	82.7
11_04	low density	homogenous blocks 1	speaking	86.0	86.0
11_04	low density	homogenous blocks 1	unspoken speech	78.7	77.7
11_05	low density	mixed blocks(seq.)	speaking	41.3	41.3
11_05	low density	mixed blocks(seq.)	unspoken speech	27.3	20.7
11_08	low density	homogenous blocks 2	unspoken speech		61.3
11_09	low density	mixed blocks(seq.)	unspoken speech	35.6	34.8
13_01	high density	homogenous blocks 2	unspoken speech		43
13_02	high density	mixed blocks(rnd) 2	unspoken speech		26.8

Table 3.2: Results when training and testing was done with all 16 channels and the LDA is set to reduce the feature space to 35 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_05	low density	relabeled	speaking	83,8	83,8
11_05	low density	relabeled	unspoken speech	56,7	58
11_09	low density	relabeled	unspoken speech		52,3
13_02	high density	relabeled	unspoken speech		46.3

Table 3.3: Results for the relabeled data correspond to the rearrangements of 11_05 , 11_09, 13_02, respectively. Training and testing was done with all 16 channels and the LDA was set to reduce the feature space to 35 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking		71
11_02	high density	homogenous blocks 1	unspoken speech		50.6
11_03	high density	homogenous blocks 1	speaking		70
11_03	high density	homogenous blocks 1	unspoken speech		60
13_01	high density	homogenous blocks 2	unspoken speech		41
13_02	high density	mixed blocks(rnd) 2	unspoken speech		21.7

Table 3.4: Results when training and testing was done with the channel input captured by electrodes on positions 1,7, 12 (refer to Figure 3.2) and the LDA was set to reduce the feature space to 35 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking		88.9
11_02	high density	homogenous blocks 1	unspoken speech		66.4
11_03	high density	homogenous blocks 1	speaking		83.3
11_03	high density	homogenous blocks 1	unspoken speech		73.3
13_01	high density	homogenous blocks 2	unspoken speech		37.5
13_02	high density	mixed blocks(rnd) 2	unspoken speech		20.8

Table 3.5: Results when training and testing was done with the channel input captured by electrodes on positions 1,2,7, 12 (refer to Figure 3.2) and the LDA was set to reduce the feature space to 35 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking		67.6
11_02	high density	homogenous blocks 1	unspoken speech		46.6
11_03	high density	homogenous blocks 1	speaking		68
11_03	high density	homogenous blocks 1	unspoken speech		60
13_01	high density	homogenous blocks 2	unspoken speech		38.5
13_02	high density	mixed blocks(rnd) 2	unspoken speech		29.5

Table 3.6: Results when training and testing was done with the channel input captured by electrodes on positions 1,7,12 (refer to Figure 3.2) and the LDA was set to reduce the feature space to 16 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking		85.5
11_02	high density	homogenous blocks 1	unspoken speech		63.9
11_03	high density	homogenous blocks 1	speaking		87.3
11_03	high density	homogenous blocks 1	unspoken speech		70
13_01	high density	homogenous blocks 2	unspoken speech		36.8
13_02	high density	mixed blocks(rnd) 2	unspoken speech		20.3

Table 3.7: Results when training and testing was done with the channel input captured by electrodes on positions 1,2,7, 12 (refer to Figure 3.2) and the LDA was set to reduce the feature space to 16 dimensions.

Session Id	*Cap Type*	*Corpus Type*	*Utterance Type*	*Recog. Rate 1*	*Recog. Rate 2*
11_02	high density	homogenous blocks 1	speaking		86.2
11_02	high density	homogenous blocks 1	unspoken speech		62.9
11_03	high density	homogenous blocks 1	speaking		74
11_03	high density	homogenous blocks 1	unspoken speech		52.7
13_01	high density	homogenous blocks 2	unspoken speech		32.8
13_02	high density	mixed blocks(rnd) 2	unspoken speech		24.7

Table 3.8: Results when training and testing was done with the channel input captured by electrodes on positions 1,2 (refer to Figure 3.2) and the LDA was set to reduce the feature space to 16 dimensions.

3.5 Experiments on attenuation estimation

As mentioned in 1.3.3, the scalp electrodes pick up a superposition of EEG signals coming from various sources of the cortex. The degree to which each source has an impact on the electrode's signal depends on the signals attenuation along the way from the origin of the signal to the position on the scalp where the electrode is attached. That attenuation in turn, is influenced by the conductivity of the tissue in between, as well as by the distance. Investigations regarding the conductivity of brain tissue are described in [1]. The therein described findings indicate a quite rapid reduction of voltage as the electrode is moved away from the source. However, even if small, the amplitude measured at distant locations was still found to be significant.
Since the underlying idea of this work was the intention to pick up signals from the motor cortex area in a more fine-grained manner by the use of additional electrodes over it close to each other two questions arise:
1) To what extent do the additional electrodes translate to more information compared to the standard layout where only one electrode is located above the motor strip on each hemisphere ?
2) If one places an electrode of the orofacial motor cortex, to what degree does this sensor also pick up EEG signals from the speech related areas like Broca's and Wernicke's ? This question is particularly relevant because successful classification results may also be derived from that two areas and not necessarily from information contained in the signals emitted from the motor cortex. This is especially self-evident since Broca's and Wernicke's area are located in direct neighborhood to the orofacial motor strip (Figure 1.1).

3.5.1 Estimating the attenuation as a function of distance

The conductivity of the head may vary from person to person between sessions, too. In order to get a feeling for our test subject's brain conductivity and the attenuation an own investigation was carried out.
The first observation that was used is the fact that the eye blink's amplitude can be spotted in every channel more and more attenuated with increasing distance to the electrode above the eyelid.

For simplification, it was assumed that in the data obtained in the experiments, the source of the blink S was located between the two eyes (nasion) [8] and that the upper half of the test subjects head could be modeled by the upper half of an ellipsoid.
In order to sketch attenuation as a function of distance through the head (i.e. ellipsoid) the coordinates of the electrode with respect to S had to be determined.
The ellipsoid was thought to be produced by rotation of an ellipse with parameters a,b (see Figure 3.3) with the rotational axis going through nasion S and inion. The constant a equals the length of the semimajor axis, the constant b equals the length of the semiminor axis of the ellipse.

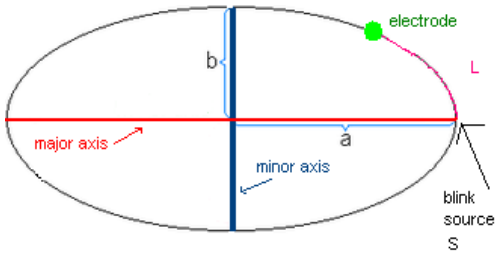

Figure 3.3: Each electrode was modeled to be on an ellipse. L is the arc length from the electrode's position to the assumed blink origin between the eyes.

The two ellipse parameters a and b were determined by measurement of the test subject's head. Also, for each electrode i, the arc length L_i was measured with a tape measure, i.e. the tape measure was spanned alongside the head starting at S and ending at the electrode i.
Let P_i denote the position of electrode i on the ellipse, then the coordinates of P_i had to be determined. Unfortunately, there is currently no method known to exactly calculate P_i from L_i. Instead the coordinates of P_i were computed with a reasonably fast, numerical forward approach, for which an algorithm was designed devised that iteratively computed the arclength L of a candidate point P so long until the error $|L - L_i|$ was smaller than a tiny, predefined threshold.
The attenuation of the blink amplitude was determined by taking the difference between the maxima of the amplitudes of the blinks in the corresponding channels and dividing it

[8] In reality, the test subject had blinked with both eyelids.

by the the maximum of the amplitude of the blink measured over the eyebrow. This was done for every electrode i averaged over 30 utterance steps.
This procedure was done with recordings from two test subjects yielding quite similar results. A representative result is depicted in Figure 3.4. The attenuation is given as a

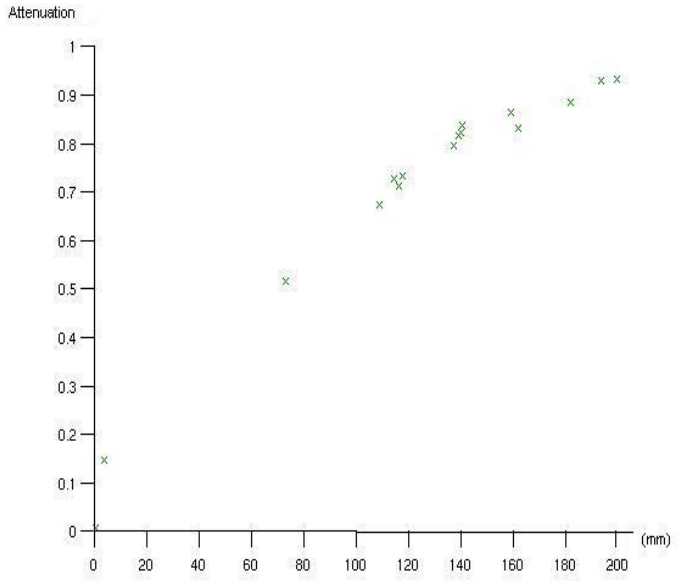

Figure 3.4: Attenuation $[\frac{percentage}{100}]$ as a function of the distance from the amplitude's origin. The functional relation would become explicitly apparent with regression.

percentage of the maxima of the amplitudes in the various channels (corresponding to distances from the source) relative to the amplitude measured directly over the eye brow. The results differ from the findings in [1] in that they indicate a less rapid attenuation with increasing distance. This may be in part due to the invalidities of the model assumptions that have not been met in the actual experiments producing the data.

3.6 Correlation between the scalp electrodes

Given the fairly good conductivity and slow attenuation of signals traversing the head, it is a reasonable question to ask whether the close concentration of an increased number of electrodes is not just leading to redundant information. To rephrase this concern in a simple question : If one measures the signals from seven electrodes, all located very close

to each other (e.g. around the orofacial motor cortex) does one essentially measure seven times the same signal then?

To get an impression of a possible answer to that question, Correlation was used as a similarity measure, again. We repeatedly computed the average cross-correlation between various channels over a block of 30 utterances. For comparability reasons this was done only for the modality *unspoken speech* [9] and only for the signals of interest. In a first step, a matrix $M := (cc(i,j))_{i,j}$ was computed for each utterance step, where cc(i,j) denotes the Correlation of the signal of interests in channels i and j. Then, these matrices where averaged. An example for such an averaged cross-correlation matrix are visualized in Figure 3.5.

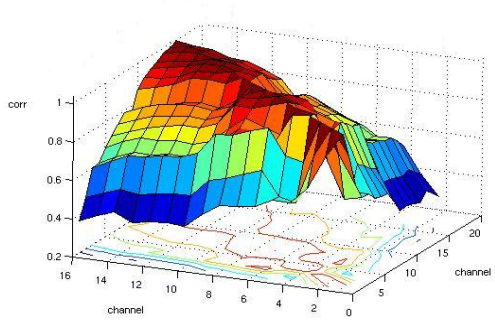

Figure 3.5: Visualization for the average correlation matrix for session 11_03 (averaged over all utterances labeled as *bravo*).

In a next step, the average cross-correlation between subsets of the electrodes was determined for the high-density cap (refer to 3.2 for the layout and the electrode numbering). The following subsets of electrodes where defined: $L := \{3,4,5,6,7,8,9\}$, $LH := \{4,7,9\}$, $LV := \{6,7,8\}$, $RH := \{10,12,15\}$, $RV := \{11,12,13\}$. The average cross-correlation for the high-density cap for the set of 30 utterances equaled 0.7731. The other average cross-correlation values for the different subsets are given by the entries of Table 3.9. As you can see, the electrodes down the motor cortex are highly correlated. The fact, that the average correlation between electrodes within LH, RH is slightly lower than within LV, RV (which are alongside vertical axes) may be explained by that the distances between two

[9] The contamination with artifacts due to muscle movements would adulterate the results.

Electrode set	L	LH	LV	RH	RV	{7,12}	{2,7}	{2,3}
Average correlation	0.8961	0.9099	0.9772	0.9599	0.9883	0.8395	0.7017	0.8435

Table 3.9: The average correlations between the electrode subsets, computed for one block of session 11_03.

horizontally neighboring electrodes was 3.2 cm while the distance between two vertically neighboring electrodes was 2 cm and hence smaller. Of course, that higher correlation could alternatively be explained by that the subsets LV,RV contain only electrodes placed above the motor strip, while the subsets LH,RH cover different, functional cortical areas. The same computations where done with sessions made with the standard, low-density cap. In this case the electrodes where not only located over more different cortical areas, but also spread out further apart. This circumstance is reflected in a lower, average correlation that was found to be 0.6116 for the example of session 11_04.

Remark 3.6.1. The average correlation matrices were different for each session but also varied to a certain degree between utterances and different blocks. However, the average correlations where always approximately of the same magnitude for the high-density and the low-density cap, respectively.

As a conclusion, we can say that the electrodes within the two sets L,R , densely located over the orofacial motor cortex of the left and right hemisphere, are highly correlated. Hence, it could support doubts regarding whether there actually is a huge benefit to the this layout over the standard one. However, the signals are not identical and it cannot be excluded the possibility that valuable information for word discrimination could be contained in exactly that fraction of the signals, that embodies their distinctness.

Chapter 4

Discussion and Final Remarks

4.1 Discussion

In the following, the classification [1] results of the experiments will be compared. The significance test of the classification results' comparisons follows [23] (chapter 5) as a modified hypothesis test. Whenever we use the term *significance* in the context of a comparison of two recognition rates, it means that the relevant relation holds with a confidence of at least 95%.

4.1.1 Comparison of the recognition rates

Word recognition rate estimation has been done with cross-validation. As you can see from Tables 3.2 - 3.8 the rates were estimated for different adjustments to the learner's input, ie we trained with different subsets of electrodes as well as with using different numbers of dimensions we projected the input data on using LDA.

Dimensionality reduction

The comparison of the estimated classification rates between learning with a reduced dimension of 16 and learning with 35 dimensions leads to the observation of no significant differences with the new cap. In contrast to the results in [43], a reduction of dimensionality onto 16 instead of onto 35 leads even to a numerically worse recognition rate for

[1] We treat our word recognition task as a classification problem with five (word-) classes.

corpora of the *unspoken speech* modality.

Training with subsets of electrodes and implications

It can be seen, that for the *speaking* modality channel 2 seems to have an important role for the recognition of words. It's exclusion from training resulted in a significant decrease of the recognition rate. Note, channel 2 contained the signal measured by the electrode in the vicinity of the scalp surface above Broca's area. Therefore, a neurological explanation would be that Broca's area contained helpful information for word discrimination. In [31] a PET-study is described investigating the cortical activation during different speech related tasks. As depicted in Figure 4.1,

Figure 4.1: PET-study of cortical activation during speech related task (from [31]).

the study found Broca's area to be involved in word-generation tasks. This indicates that we also should observe a similar, significant decrease in recognition rate for the *unspoken speech* modality, which is the case.
However, an alternative explanation of the decrease is the position of electrode 2 being especially frontally located, close to the source of electrical signals originating from facial muscle movements during audibly speech production[2]. Therefore, if EMG was highly responsible for good classification results the effect of excluding channel 2 would cause a smaller fraction of EMG information to be accessible to the learning algorithm and hence,

[2]The electrical signal originating from muscle contraction is called electromyogram (EMG).

could lead to the worse classification performance we observed. It would also explain the fact, that the estimated recognition rates for *speaking* utterances are significantly better than for *unspoken speech* utterances.
It is to be pointed out, that all investigations trying to link the usefulness of a scalp electrode for classification to the function of the cortical area underneath the electrode, are advised to be accepted with reservation: Even if certain classification results are achieved in the absence of an electrode above a certain cortical area it does not mean that neighboring electrodes do not pick up the signal from that area as well as should have become clear in 3.5. The situation becomes especially tricky with reasoning about the role of the orofacial motor cortex, because it is very close to prominent speech related areas like Wernicke's and Broca's.
Seen in this light it becomes clear, why there is no significant decrease in recognition rate when only electrodes 1,2,7,12 are used. Hence, these electrodes can be seen to be important. The surrounding electrodes seem to contribute more or less redundant information for classification.

New cap and layout versus old ones

At first glance, the recognition results in Table 3.2 show appealing results for *homogenous blocks* corpora, being clearly significantly higher than a recognition rate that would have been produced by random guessing (which would most likely be around 20 %).
However, if the corpora recorded with the new, high-density cap (with the layout according to Figure 3.2) are compared to the ones recorded with the standard, low-density cap, on average, no significant differences in recognition rates can be observed for the same test subject between corpora of the same modality. Therefore, the new layout could not be shown to have a significant advantage over the standard layout, at least with our learning approach.

Remark 4.1.1. On the other hand it is remarkable, that the high redundancy of the electrodes' signals in the new layout (see Section 3.6) which is probably explainable with the good conductivity properties of the head (see Section 3.5) does not translate to poorer classification results.

Of course, the observations are based on the idea that the learning algorithm actually learned the proper concept that leads to good generalization abilities in classifying words.

4.1.2 Did we really recognize words?

In [43], all recognition rates had been obtained with *homogeneous blocks* corpora, using cross-validation estimation. Although these numbers looked promising, the trained classifier had not generalized well in the field, i.e. all attempts to build a demo-system capable of online-classification of completely new, unseen examples after completed training, with a recognition rate significantly higher than random guessing, had failed. Biological and instrumental reasons had been identified as possible explanations [43].
However, this phenomena inspired us to investigate in an alternative direction [3]: overfitting.
By definition, a learner's output-hypothesis h is said to overfit the giving training data if there exists some alternative hypothesis h' from the same hypothesis space, such that h has a smaller error than h' on the set of training examples, but h' has a smaller error than h over the entire distribution of instances (i.e. h will have a lower expected classification performance than h' in the classifier's real environment) [23]. The causes for overfitting can be manifold.
While it was attempted to counter overfitting with cross-validation, it is still conceivable that the training data may have contained regularities accidentally changing in correlation with the labels of the corpus' instances. If that was the case, it might have misled the training algorithm spawning it to learn the wrong concept. We conjectured, that (eg due to non-stationarity of EEG-signals, slow head- volume conduction changes,...) long temporal changes of some EEG-signal properties were temporally aligned with the structure of the *homogeneous blocks* corpora.

Remark 4.1.2. This structure has the following property: If one assigns to each utterance a time stamp indicating the relative (to the beginning of the experiment) time when the utterance was recorded, then the temporal difference between two recorded utterances of the same class will be on average relatively small, while this time difference between two instances of distinct classes will be relatively large. The argument for this property becomes even stronger in the light of Remark 3.2.3. Therefore, the fraction of the temporal within-class scatter over the temporal between-class scatter is low. If our features that were gained in the feature extraction process directly or indirectly contain the time stamp

[3] Remember we were using the exact same training/ classification algorithm as in [43], hence would certainly share the same problems.

information by detecting the suspected temporal changes in the EEG-signals, it seems reasonable to assume the LDA would choose this information as a good transformed feature to project on.

If this conjecture was true, then the recognition results would drop down if we designed experiments creating corpora without the just mentioned property. To try this, experiments leading to the *mixed blocks* corpora where executed. The results support the conjecture (refer to Figure 4.2 and 4.3).

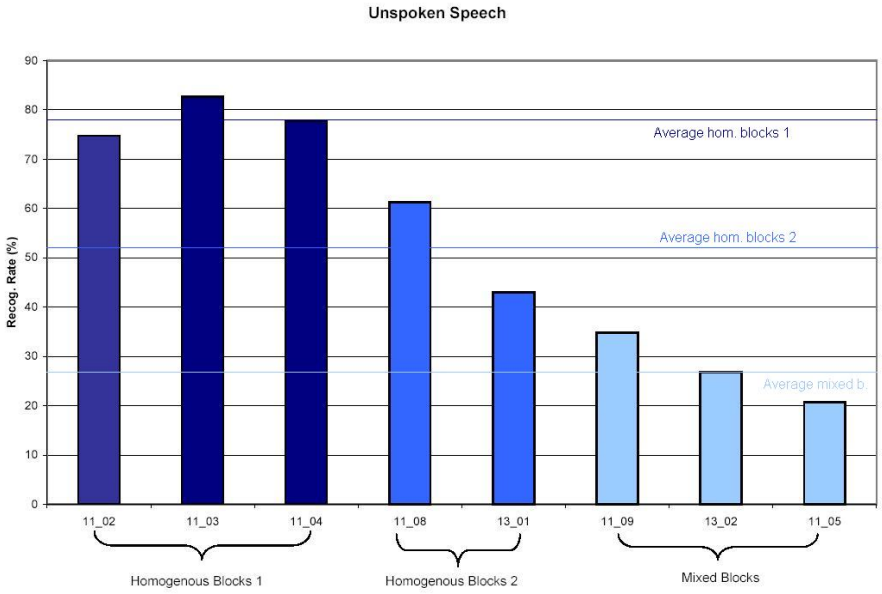

Figure 4.2: Recognition rates for the modality *unspoken speech*. Note, the average recognition rate for the *mixed blocks* corpora is not even significantly higher than chance.

They have been obtained with the settings described in Table 3.2.
The conjecture is even more supported by the observation, that the property described in Remark 4.1.2 is valid with an even higher magnitude for *homogeneous blocks 1* corpora than for *homogeneous blocks 2* corpora. This could be due to the circumstance, that the average temporal difference between elements of two blocks of the same modality is larger, because two blocks of the same modality (and different classes) are temporally separated by at least one block of the other modality. So if the conjecture is true the, average recognition rate should be higher for *homogeneous blocks 1* corpora than for corpora of type *homogeneous blocks 2*. Indeed, this observation can be made as can also be seen in

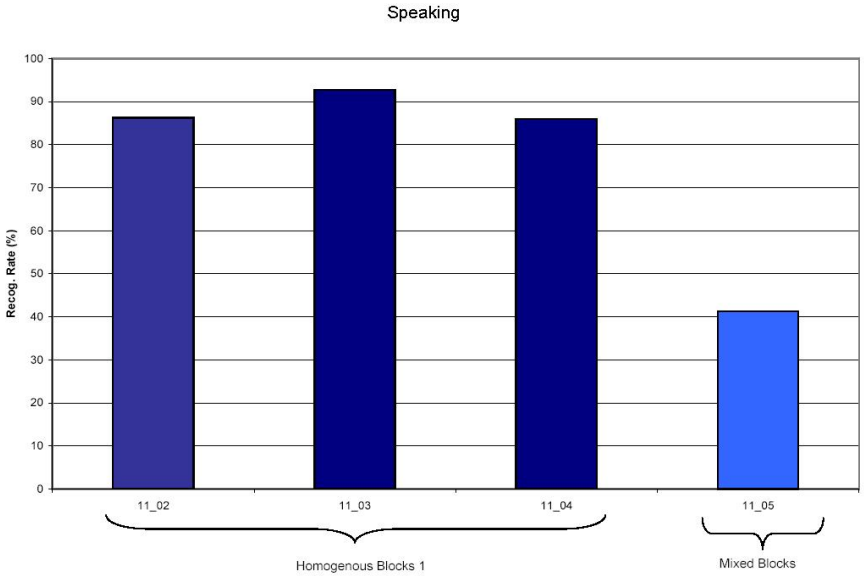

Figure 4.3: Recognition rates for the modality *speaking*.

Figure 4.2.

Other observations also contribute to an acceptance of the conjecture, in that they suddenly make sense when seen in the light of the it. By referring to Table 3.2 one is able to observe (for the sessions held in this work, for modality *unspoken speech*, on average) slightly lower classification rates when the signal of interest was extracted with the new extraction algorithm presented in Chapter 2, compared to the case when the previous mechanism was used. This was astonishing at first, since we believe the new algorithm to be more reliable and more precise in extracting the correct signal. At first, the higher precision was expected to translate to containing less falsely included signal parts of the first blink within the extracted signal on average. Hence, this was expected to translate to less irrelevant, potentially misleading information in the signal of interest.

In the same spirit, the fact that the recognition rates were significantly better for the utterances made in the modality *speaking* is remarkable when noting that we discovered that the eye blink detection and hence, the extraction of the signal of interest did not take place for that modality at all. Therefore, most of the signal of a *speaking* utterance is supposedly irrelevant and again, potentially misleading. One conceivable explanation was to assume the more frontal electrodes picked up a lot of signal originating from the facial

muscle movements during the speaking of the words (EMG-effects). If we now choose to view it in the light of our conjecture, both observations could be explained by assuming that the classification was still successful because what was actually learned/recognized was some signal component that is not confined to the support of the signal of interest as we would expect from the information regarding the word-classes. The fact that the recognition results were even better could then be interpreted to be caused by the increased length of the signals which means that potentially more of the temporal information could be extracted and also, if more of the whole recorded instance's signal is used as a reference, the relative distance between two members of the same block shrinks.
To get more clearness, the *mixed* corpora were transformed into *relabeled* corpora. With the argument in Remark 3.2.4 it becomes obvious, that a good, estimated recognition rate of these falsely labeled utterances could not be explained by a successful use of information that is anyhow related to the word- class that was uttered. It could be related to the improved property described in Remark 4.1.2 caused by the artificial homogenization, instead and therefore, give the strongest indication that our conjecture is true.

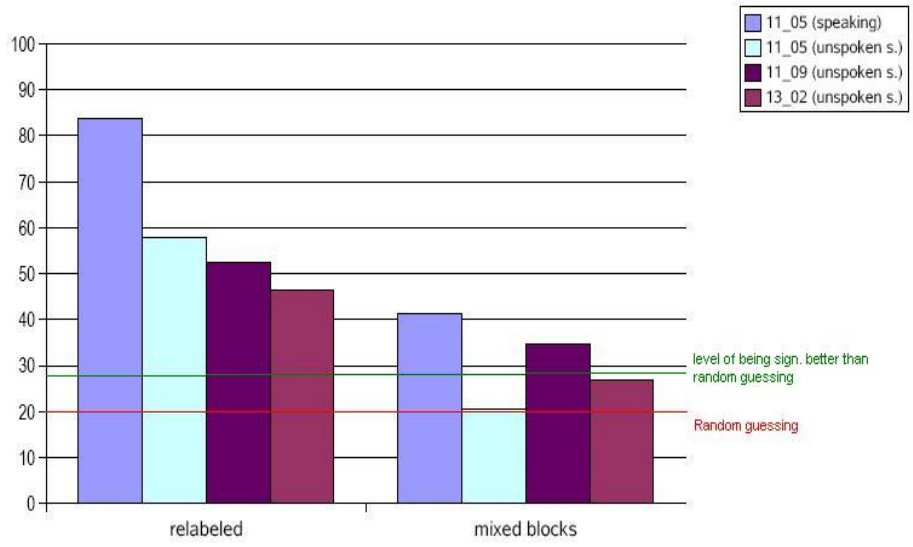

Figure 4.4: The *mixed* corpora experience a boost-up in recognition rate, after transforming them into *relabeled*.

As a matter of fact, this observation can be made with the data, as is depicted in Figure

4.4.

In summary, our conjecture has been demonstrated to be likely true. There seem to be strong indications making it evident, that our EEG-based word recognizer has a tendency to make classification decisions based on a so far unknown feature reflecting temporal effects rather than different word-classes, if trained inappropriately. Inappropriately means in this context, that the labeling of the training examples reflects temporal information (in the sense of Remark 4.1.2, rather than the true concept which would allow actual word recognition). Inappropriate training was done for all *homogeneous blocks* corpora. As a result, the production of *mixed* corpora for training is strongly recommended. However, it is to be pointed out, that this does not necessarily imply that the recognizer has no potential in being capable to learn the concept leading to successful word recognition in a real environment, at all. It simply means that there is a danger of achieving overly optimistic estimates of the real environment recognition rates under the current settings (ie with the current features that are extracted and parameters of the learning algorithm) if the training set is chosen inappropriately. Therefore, we recommend using training corpora of type *mixed* for future efforts which are better candidates for not causing the learner to overfit.

In effect, there have been some *mixed*-type corpora showing estimated recognition rates being significantly higher than random guessing. These are 11_05 (41.3%) for the *speaking* modality and 11_09 (34.8 %) for the *unspoken speech* modality if one chose to use all channels for training and projected on the first 35 dimensions of the LDA's image space (refer to Table 3.2). For the setting described 3.6, 13_02 exhibited an estimated recognition rate being on the verge of being significantly higher than random guessing (29.5 %).

This gives reason to still believe that there could be a potential in our recognizer to learn a concept allowing EEG-based word recognition if trained appropriately. However, this should be investigated more thoroughly in future work.

4.2 Summary

Our main task was to aggregate electrodes on the scalp above the orofacial motor cortex and compare the recognition results to the ones achieved using the standard layout already used in [43] for the modalities *speaking* and *unspoken speech*. For this purpose,we

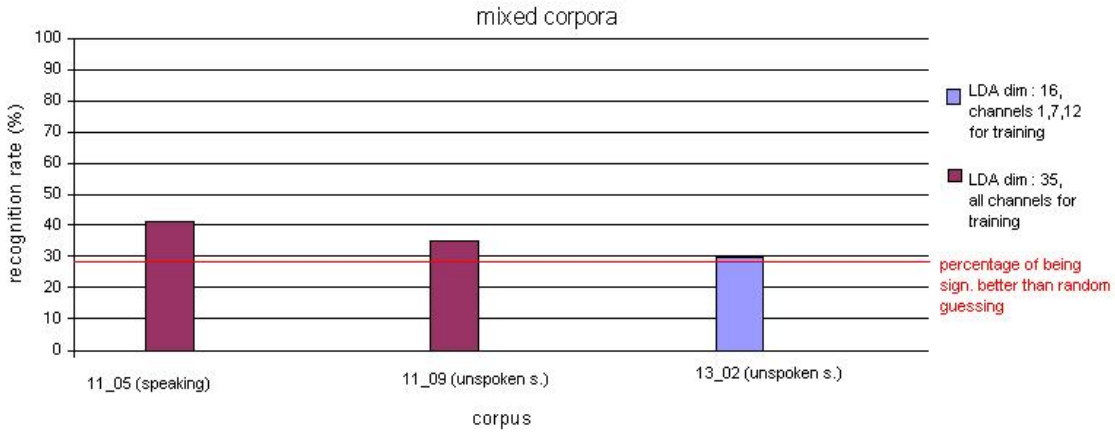

Figure 4.5: The *corpora* yielding to classification results being significantly better than random guessing.

acquired a new cap for EEG - recordings that was, due to a higher electrode density, better suited for realizing such arrangements of electrodes, adapt the implementation of the work mentioned above and executed experiments. The electrodes of were found to be highly correlated when using the new layout. However, we found no significant differences between the estimated recognition rates on data produced with the aggregated, layout on the new cap and the data acquired with the old, standard layout and cap.

However, we executed investigations making it doubtful whether the estimated recognition rates obtained in [43] and at the beginning of this work are meaningful. Instead, we believe to have demonstrated that, with the current setting, the structure of the recorded training examples actually caused the speech recognizer to learn a wrong, temporal concept that misleadingly was correlated to the labeling of the first training data sets.

The corpora that we expect to have no such problems and are therefore advised to be generated in future experiments, are the ones of modality *mixed*. Unfortunately there was no time to record more sessions in that modality. The estimated recognition results for the few available corpora were inconclusive. For *unspoken speech* utterances, some were significantly higher than random guessing, some were not. We only had one unaffected training data set with instances of the *speaking* modality. It was significantly higher than random guessing. However, further investigations producing more unaffected data will have to be made.

In the course of the work, additional improvements and ideas were developed. These

include the derivation and implementation of a cross-correlation based pattern detection algorithm that was then used for blink detection and proved to be very accurate and reliable for that application. Also, methods were investigated for the purpose of being able to estimate the test subject's brain tissue conductivity properties, which translate to an estimation of the degree of ability to assign meaningfulness to some aspects of the experiments. As we have seen, scalp electrodes pick up signals from a huge variety of cortical areas and not just from the area directly underneath the electrode.
We observed significantly better estimated recognition rates for the utterances of *speaking* modality. We prefer to hypothesize this effect to be due to influence of EMG.

4.3 Ideas for future work and final remarks

In any case, the number of available corpora should be increased to get more cogent results which was not possible to realize in this work due to the given constraints in time and budget. For the same reasons additional experiments regarding the learning algorithm's settings could not be done. One of the more obvious things to do would be to further investigate the better recognition rates for *speaking* corpora. If this is due to EMG- effects, it would be interesting to try to exclude the frontal electrodes from training and observe if the recognition rate decreases. If it does this might be explained by the fact that the EMG - waves will be more attenuated in the remaining channels (that are picking up signals originating from scalp location farther away from the moving, facial muscles). Also, the algorithm for the extraction of the signal of interest should be applied to *speaking* corpora. Furthermore, the nature of the conjectured temporal effects could be investigated in narrower detail.
As an idea for preprocessing, it could be considered to dilate all the signals to obtain uniform length. By this, it can be excluded that on the one hand, signals of different classes can be discriminated due to their length, on the other hand signals of the same class might become more similar. In this context it might be also worth trying to use Wavelet based features (eg MRA). When MRA is used, it may be interesting to try to detect patterns on different scales which would solve the just mentioned issue of different utterance length of the same word.
Conceivably improving results could develop from the attempt to execute experiments

with a combination of MEG and EEG, so that the activity of neurons oriented to various directions could be measured, not only from the ones close to perpendicular to the scull's surface.

Another idea that was casually pursued in this work was to compute a matrix $M := (cc(i,j))_{i,j}$ for each utterance step, where cc(i,j) denotes the cross-correlation of the signal of interests in channels i and j it as a feature for classification. While nearest-neighbor classification has not worked for discriminating the words of our vocabulary domain the following situation could be tried:

For every utterance step, compute M and connect its entries to the input layer of a neural network. Perhaps, training of such a neural network could lead to results, if the channels correspond to electrodes of different, suitable cortical areas and the vocabulary is composed of words of different semantic categories. This appears a worthwhile experiment, because the thought on words of different semantic categories may lead to different patterns of synchronous activation of cortical areas.

In the field of brain-activity signal classification, there may be a variety of ideas that may lead to results, eventually, if one is able to discover suitable features that allow successful classification of the different classes in question. Such discoveries may either be inspired by neurological progress or inversely, a coincidental discovery of such features may as well foster an understanding of the human brain in return. Up to today, classification attempts of cortical activity share the common problem of being speculative about what features may be good to extract and what information needs to be contained in the measured signal in the first place. The main reason for this is, that present knowledge of the mind's inner workings is still very limited and remains in the shadow of uncertainty. However, this ostensible weakness,in concert with the philosophical implications of neuroscience, are an intriguing challenge, making brain-activity classification a domain particulary worth pursuing.

Bibliography

[1] Allison, T. , *Calculated and empirical evoked potential distributions in human recordings*, Multidisciplinary Perspectives in Event-related Brain Potential Research. Ed. by Otto, A. D. , Washington DC, US Environmental Protection Agency, 1978.

[2] Aubin, J-P. , *Applied Functional Analysis*, 2nd ed., Wiley-Interscience, 2000.

[3] Baillet, S., Mosher,J.C., Leahy,R. M., *Electromagnetic Brain Mapping*, IEEE Signal Processing Magazine, 2001.

[4] Becker, K., $Varioport^{TM}$- *Gebrauchsanweisung*, 2004.

[5] Berger, H., *Über das Elektroencephalogramm des Menschen (On the human electroencephalogram).*, Archiv für Psychiatrie und Nervenkrankheiten, 1929.

[6] Berlin Brain-Computer Interface .

[7] Caton, R., *The electric currents of the brain.*, Br.Med. J.,1875

[8] Cuffin, B.N. , *EEG dipole source localization*, Engineering in Medicine and Biology Magazine, IEEE , Sep/Oct 1998.

[9] Cichocki,A. and Vorobyov,S. . *Application of ICA for automatic noise and interference cancellation in multisensory biomedical signals*, In Proceedings of the Second International Workshop on ICA and BSS, pages 621-626, June 2000.

[10] Daubechies,I. *Ten Lectures on Wavelets*,SIAM, May 1992.

[11] Duda,R.O.,Hart,P.E.,Stork,D.G. *Pattern Classification*,2nd.ed., Wiley-Interscience, 2000.

[12] *http://www.erzwiss.uni-hamburg.de/personal/hoffmann/lehre/sose2000/*

[13] Gutierrez, D. Nehorai, A. Muravchik, C.H. , *Estimating Brain Conductivities and Dipole Source Signals With EEG Arrays* IEEE Transactions on Biomedical Engineering, vol. 51, no. 12, Dec. 2004.

[14] Harasty, J.,Double,K.,Halliday,G.M.,Kril,J.J.,McRitchie,D.A., *Language-associated cortical regions are proportionally larger in the female brain*, Archives in Neurology, 54.

[15] Heuser, H. , *Funktional Analysis*, 3.Auflage, Teubner, 1992.

[16] Honal, M. , *Determining User State and Mental Task Demand From Electroencephalographic Data*, Diplomarbeit, Institut für Theoretische Informatik, Universität Karlsruhe(TH),2005.

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

[18] Howart,D.,Patterson,K.,Wise,R.,Brown,W.D.,Friston,K.,Weiller,C.,Frackowiak,R., *The localization of the lexicons.*, Brain,115,1992.

[19] Huang, X., Acero, A., Hon, H. Spoken Language Processing, Prentice Hall PTR, NJ, 2001.

[20] Ingvar, D.H., Schwartz, M.S., *Blood flow patterns induced in the dominant hemisphere by speech and reading*, Brain, 97, 1974.

[21] Kandel, Schwartz, Jessell, *Principles of Neural Science*, 3.ed., 2000.

[22] Kull, L. , *Kathleen Mears Lecture: Twentieth Century EEG, Technology for the 21st Century: Considering our Future by Examining our Past*, Lecture notes, 1996.

[23] Mitchell, T. , *Machine Learning*, McGraw Hill, 1997.

[24] Knight, J. N., *Signal Fraction Analysis And Artifact Removal In EEG*, Master thesis, Department of Computer Science, Colorado State University, 2003.

[25] Lassen, Ingvar, Skinhoj, *Brain function and blood flow.*, Scientific American, 239, 1978.

[26] Mayer, C, *UKA EMG/EEG Studio v2.0*, 2005.

[27] Metze, F., et. al., *JRTk online documentation*, http://isl.ira.uka.de/ jrtk/janusdoku.html.

[28] http://calder.med.miami.edu/pointis/

[29] Jung T-P, Makeig S, Bell AJ, and Sejnowski T. J, *Independent Component Analysis of Electroencephalographic and Event-related Data*, In: (Poon P. and Brugge J, ea), Auditory Processing and Neural Modeling , 189- 197, 1998.

[30] Pereira,F., Gordon,G., *The Support Vector Decomposition Machine*, Proceedings of the 23rd International Conference on Machine Learning,Pittsburgh, 2006.

[31] Petersen, S.E., Fox, P.T., Posner, M.I. , Mintun , M., Raichle, M. E. *Positron emission tomographic studies of the cortical anatomy of single-word processing.*, Nature, 331:585-589, 1988.

[32] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.*, In Proceedings of the IEEE, Vol.77, No. 2, Feb. 1989.

[33] Rabiner, L. and Juang, B. *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.

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

[35] Schmidt, R. F. and Thews, G., *Physiologie des Menschen*, Springer, 1997.

[36] Scholl,D.A., *Organization of the Cerebral Cortex.*,Methuen, 1956

[37] Soong, A. C. K. , Koles, Z. J., *Principal-Component Localization of the Sources of the Background EEG*, IEEE Transactions on Biomedical Engineering, VOL. 42, NO. 1, Jan. 1995

[38] Spitzer,M., *Geist im Netz. Modelle für Lernen, Denken und Handeln*, Spektrum, akad. Verlag, 2000.

[39] Suppes, P., Lu, Z., Han, B., *Brain wave recognition of words*, Proc. Natl. Acad. Sci. USA., vol. 94, 1997.

[40] Toga, A.W. ,Maziotta, J.C. *Brain Mapping*, 2.ed., Academic Press, 2002.

[41] http://bss.ewi.utwente.nl/research/neurostimulation

[42] Werner, D., *Funktional Analysis*, 3.Auflage, Springer, 2000.

[43] Wester, M. , *Unspoken Speech*, Diplomarbeit, Interactive Systems Lab, Carnegie Mellon University, Universität Karlsruhe(TH). , 2006

[44] *http://www.zeiss.de* .

[45] *http://www.psychologie.unizh.ch/neuropsy/Forschung/eeg*

[46] Zhukov,L. , Weinstein, D. , Johnson, C., *Independent Component Analysis for EEG Source Localization*, IEEE Engineering in Medicine and Biology, 2000.

[47] Zschocke, S., *Klinische Elektroenzephalographie.*, Springer, 1995