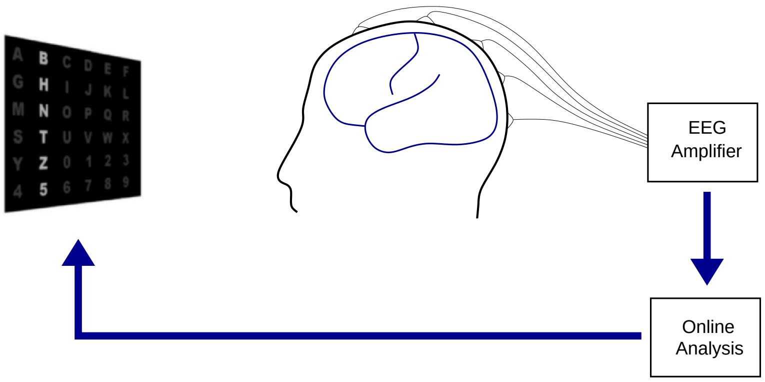
P300-Based
Brain-Computer Interfacing



Matthias Kaper

Reviewers: Helge Ritter, Niels Birbaumer, Horst M. Müller

Der Technischen Fakultät der Universität Bielefeld
vorgelegt zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften

Printed on acid-free paper (DIN-ISO 9706).

Acknowledgements

Above all, I wish to thank Helge Ritter for his spirit and support. Working in his Neuroinformatics Group was not only very inspiring but also a great pleasure. Therefore, thanks to the members of the Neuroinformatics and Applied Neuroinformatics Group, particularly Bert Arnrich, Holger Bekel, Tim W. Nattkemper, Jörg Ontrup, Axel Saalbach, and Thorsten Twellmann for providing this pleasurable and supportive social environment. Speaking of support, I am very much obliged to Risto Kõiva and Oliver Lieske for their technical support and to Petra Udelhoven for administrative aid. I highly appreciate the help of Bert Arnrich and Thorsten Twellmann, who faced the unpleasant work of proofreading.

This thesis was completed while working in the Collaborative Research Center 360. I learned a lot within this interdisciplinary cooperation and I am grateful for the fruitful cooperations with Horst M. Müller and Sabine Weiss.

Special thanks go to Florian Hoppe with whom I had a great time developing the basics of this work in our Diploma Thesis. The work would further not have been possible without Peter Meinicke who introduced us into the depths of Machine-Learning.

I further wish to thank the Cognitive Psychology Group, headed by Odmar Neumann, for introducing me to EEG research and providing the experimental environment for the offline experiments. Particularly Ulrich Ansorge, Manfred Heumann, and Werner Klotz promoted my scientific development in this area.

For their invaluable assistance, I wish to thank Andrea Finke, Alexander Lenhardt, and Christian Stollenwerk.

Sincere thanks to my mother, my family, my friends, and my Verena for all the care and all the joy. Great to be among you!

Thank you for the music, Ben Sowton.

Table of Contents

Chapter 1

Introduction

The field of *Brain-Machine Interfacing* has received much attention and has become a prospering domain for research in recent years, mainly caused by the progresses in hardware and data analysis techniques.

The main idea behind Brain-Machine Interfacing is to connect the brain with an artificial device. Therewith, it should be possible to gain *control* over the artifical device, which could especially be interesting for people suffering from severe motor diseases like *Amyotrophic Lateral Sclerosis* (ALS). In the worst case, such patients experience a *locked-in syndrome*, such that they are literally locked-in in a motionless body although they are still conscious and have an awake mind. On the other hand, the brain can also be stimulated by the artificial device to, e.g., *restore lost sensory functions* as in an auditory prosthesis like the cochlea implant. Third, it is even possible to stimulate the brain for therapeutical purposes. It is for instance possible to inhibit symptoms from *Parkinson's disease* (PD) by stimulating the subthalamic nucleus in the brain.

According to the notation of Nicolelis et al. (2004), Brain-Machine Interfaces can be divided into *invasive* and *non-invasive* approaches. While the former approaches penetrate the body, the latter ones do not require incisions into the body. Non-invasive Brain-Machine Interfaces are also denoted as *Brain-Computer Interfaces (BCIs)*. At present, such Brain-Computer Interfaces can only be designed to *control* an artifical device. The by far most frequently employed technique for assessing brain signals for such an interface is *Electroencephalography (EEG)* due to its excellent temporal resolution and its comparable little costs.

Several approaches in Brain-Computer Interfacing exist, but they all have in common that they aim to increase their current unsatisfying speed for transferring information. This means, within a specific amount of time, as much information as possible should be transferred to the artificial device in order to become able to, e.g., write letters or steer a wheelchair with a reasonable speed.

Among the Brain-Computer Interfaces, the P300 speller paradigm, introduced by Farwell and Donchin (1988), produces high transfer rates. This paradigm relies on the so-called *P300 component* which can be elicited by rare (*oddball*) stimuli, a subject is directing attention to. This component is well-studied and can be induced in human subjects without prior training, making it appealing for Brain-Computer Interfacing which often requires to train the subject who is intending to use the device.

This thesis focuses on the P300 speller paradigm and aims to improve its transfer rates by employing Machine-Learning techniques for data analysis. Such techniques *learn* from given data to perform, e.g., classification tasks. They can therefore easily adapt to a given data structure and do not depend upon assumptions about the structure of the problem. While common Model-Based techniques rely on data from one electrode, the ability of Machine-Learning techniques to adapt to a data structure is exploited in this thesis by making use of data from a *set* of electrodes without the necessity to extend some model assumptions. *Support Vector Machines* as state-of-the-art Machine-Learning classifiers as well as the more simple and computationally less demanding *Fisher's Linear Discriminant* are utilized. In order to become able to directly work with the latter technique, appropriate feature vectors of low dimensionality need to be constructed first.

Own experiments are conducted with the P300 speller paradigm and analyzed *offline*. This means that data are recorded from subjects who perform the experiment without actually *operating* the device. Only after the experiment itself, the data are analyzed and classified. With this kind of experiment, reliable data are produced on which the data analysis techniques are optimized with the goal to increase the transfer speed.

Another topic which is investigated in this thesis is the generalization capability to new subjects, from which no data were previously assessed to train the classifiers. This enables to perform classifications without prior training of the classifier with respect to the individual subject. Rather, a more *general* classifier can be constructed.

In a second step, the classification strategies as derived from these experiments are employed in order to construct an *online* Brain-Computer Interface for the P300 speller paradigm, which can actually be operated by a user. The system is built up from the scratch, and own software is designed which incorporates the optimized classification strategies as derived from offline experiments. An own online experiment is conducted with this system, which proves to work with a high speed and yields high information transfer rates compared to existing P300-based Brain-Computer Interfaces.

This thesis is organized as follows: After introducing into brain anatomy and Electroencephalography in chapter 2, an overview of current Brain-Machine Interfaces with an emphasis on non-invasive approaches is given in chapter 3. This chapter further introduces into the concept of *information transfer rates.* Previous work with the P300 speller paradigm, which is the central concept in this thesis, is reviewed in chapter 4 and considerations about transfer rates in this context are made. Afterwards, data analysis techniques for preprocessing and classification of the BCI data in this thesis are discussed in chapter 5. Thereby, an introduction into Statistical Learning Theory is provided within this chapter.

Offline experiments are conducted in chapter 6, and it is investigated how to improve the speed of the P300 speller Brain-Computer Interface. Additionally, possibilities for generalizations are examined. In chapter 7, the way from offline to online Brain-Computer Interfacing is drawn, and it is described how such a system is created from the scratch incorporating the findings from the foregoing chapter within an efficient and powerful data analysis procedure. Finally, the results obtained are summarized and an outlook is given in chapter 8.

Chapter 2

Electroencephalography

Electroencephalography (EEG) is the measurement of electrical activity in the brain as recorded by scalp electrodes. After Richard Caton discovered the electrical activity of the brain on the exposed cortex of rabbits and monkeys (Caton, 1875), Hans Berger recorded electrical human brain signals from the scalp and published the first study about the so-called "Elektroenkephalographie"(Berger, 1929). Berger made several fundamental discoveries about EEG signals, such as α and β waves (see section 2.4), amplitude changes during epileptic seizures, and altered signals in *Alzheimer's disease* and *Multiple Sclerosis*.

Due to the high temporal resolution and the non-invasive nature of the EEG, i.e., no incision of the body is necessary, it became a valuable tool for investigating human brain activity. While other techniques like *functional Magnetic Resonance Imaging* (fMRI) offer a higher spatial resolution, the EEG exposes a very high temporal resolution making it an important technique for analyzing cognitive processes in the brain and the method of choice in Brain-Computer Interfacing at present.

After a brief introduction into brain anatomy in section 2.1, this chapter introduces into Electroencephalography by explaining on which processes in the brain the EEG relies on (section 2.2) and how these signals are technically acquired (section 2.3). When performing EEG measures, artifacts and ways to reduce them always have to be considered (see section 2.3). Typical ways of analyzing EEGs are discussed in section 2.4. A common way to analyze EEG data is to calculate averages in terms of Event-Related Potentials (see section 2.5). In this section, the *P300 component*, essential for this work, is also discussed in detail. The reader is referred to the German books of Birbaumer (1990), Birbaumer and Schmidt (2005), and Zschocke (1995), or the English articles of Coles and Rugg (1995), Coles et al. (1990), Picton et al. (1995), and Näätänen (1982) for more detailed information about Electroencephalography.

2.1 Brain Anatomy

Since Electroencephalography measures signals from the brain, it is useful to introduce into its basic structures and mechanisms. Thus, in the following, a brief introduction is given into the *phylogenetic structure* and the *organization of the neocortex*, which hosts the most complex functions in the brain. More detailed information can be received from Pinel (1990), Birbaumer and Schmidt (2005), and Nieuwenhuys et al. (1988).

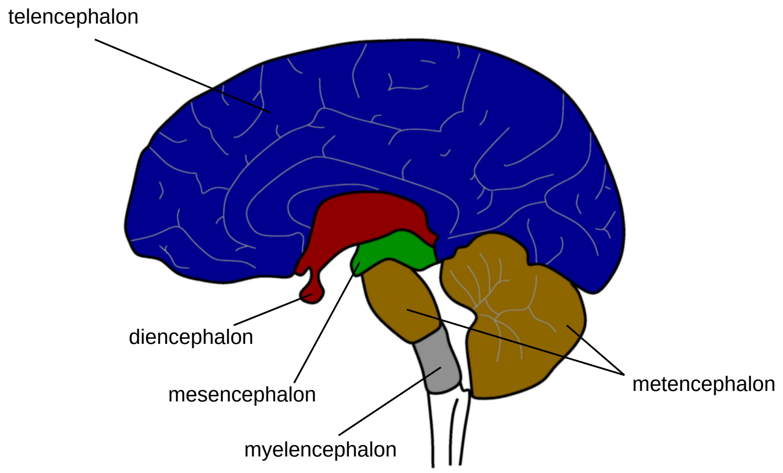


Figure 2.1: Sagittal section of the human brain, illustrating its phylogenetic structure. While the telencephalon is the phylogenetic youngest structure, the myelencephalon is the oldest. The ensemble of diencephalon, mesencephalon, metencephalon and myelencephalon is also referred to as the *brain stem*.

2.1.1 Phylogenetic Structure

The vertebrate's brain can be divided into its phylogenetic structures *telencephalon, diencephalon, mesencephalon, metencephalon* and *myelencephalon* as depicted in the sagittal section of Figure 2.1. The latter four structures are also referred to as the **brain stem**. While the myelencephalon is the phylogenetic oldest structure, the telencephalon is the youngest. Depending on the evolutionary age of animals, the different phylogenetic parts expose different expansions; for example, phylogenetic younger animals expose an enlarged neocortex, which is part of the telencephalon. In principle, elder structures are concerned with ensuring basic bodily functions, while younger structures are more important for higher cognitive functions and are less specialized. The **telencephalon** includes the *neocortex*, the *limbic system* and the *basal ganglia*. While the neocortex plays a crucial role in complex reasoning, perception and consciousness, the limbic system is involved in the regulation of motivated behavior, and the basal ganglia play a major role in performing voluntary motor responses. The **diencephalon** is composed of the *thalamus*, the *hypothalamus*, the *pituitary gland* and the *optic chiasm*. The thalamus processes signals from sensory receptors and direct them to sensory cortex areas. On the other hand, the hypothalamus is involved in the regulation of motivated behaviors by regulating the release of hormones from the pituitary gland; the optic chiasm processes information from the optic nerves to the visual cortex. The **mesencephalon** processes visual and auditory information in its *tectum* for fast responses, while another structure, the *tegmentum*, is related to simple sensomotoric coordinations and reflexes. In the **metencephalon**, including the *cerebellum*, movement and body position regulations happen; in the **myelencephalon**, which leads to the spinal cord, the *reticular formation*, a network of about 100 nuclei, regulates essential functions like sleep and attention as well as cardiac, circulatory and respiratory reflexes (Pritzel et al., 2003).

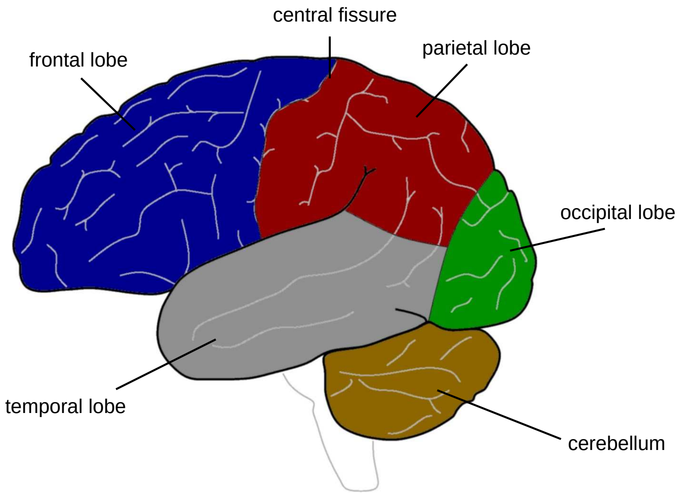


Figure 2.2: Cerebellum and the organization of the neocortex. The neocortex can roughly be divided into the frontal, parietal, occipital and temporal lobe. The cerebellum belongs to the metencephalon, and is involved in motor processes.

2.1.2 Organization of the Neocortex

The neocortex is divided into the left and the right hemisphere by the *rolandic fissure*. Each hemisphere is concerned with the processing of motor and sensory information of the opposing side of the body. Furthermore, the left hemisphere is commonly regarded to be mainly responsible for analytic and language-related thinking, while the right hemisphere is to a higher degree involved in more holistic aspects of thinking like melody and spatial perception (Pritzel et al., 2003).

The neocortex can roughly be divided into the regions *frontal lobe, parietal lobe, occipital lobe* and *temporal lobe* within each hemisphere as depicted in Figure 2.2 (Pinel, 1990). The central fissure separates the **frontal lobe** from the parietal lobe. The former lobe hosts areas for preparing and producing motor actions in its premotor and motor cortex areas, as well as areas for emotional, motivational and social behavior in the prefrontal cortex. Furthermore, *Broca's area* is located in the frontal cortex, in 97% of the humans in the left hemisphere. Damages to this area result in strong deficits in the production of fluent speech (Goodglass and Geschwind, 1976).

The **parietal lobe** includes the somatosensory cortex which captures incoming sensory information from all parts of the body and is organized in a somatotopic fashion, i.e., neighboring body parts project onto neighboring locations in this part of the brain. Thereby, not the size of the specific body part determines the expansions of their corresponding areas in the brain, but the number and types of the receptors. Thus, the hand is represented by a larger region than, e.g., the chest. Within the **occipital lobe**, visual stimuli are processed via the dorsal stream into the parietal lobe, and over the ventral stream into the temporal lobe (Ungerleider and Mishkin, 1982). While the former stream is said to prepare motor actions and is sometimes assigned as the *where-system*, the latter one is likely to be involved in object recognition and conscious representation (*what-system*). Beside these functions, the **temporal lobe** includes auditory processing structures as well as, usually in the left hemisphere, the language-related *Wernicke's area* Müller (2006). Damages to this area result in deficits in language comprehension and lead to the production of non-sense sentences by the patient. The temporal lobe hosts furthermore parts of the olfactory system.

2.2 Electrophysiological Mechanisms behind EEG Signals

In Electroencephalography, voltages in terms of electrical potential differences between electrodes are measured by an EEG amplifier. While one electrode (or a set of certain electrodes) provides a reference signal, e.g., acquired from the ear, other electrodes measure electrical activity from the scalp. Due to the long distances to generators of the electrical signals in the cerebral cortex and the weakness of the signals from single neurons, mainly only synchronous activities from *assemblies* of several hundred thousands of neurons can be assessed by this technique. While signals from the brain's direct surface as measured inside the head by the *Electrocorticogram* (ECoG) reveal amplitudes up to 100μV, EEG amplitudes are 3 to 10 times smaller (Cooper et al., 1965). Compared to ECoG, the spatial resolution of the EEG is also worse (1cm^2 to 6cm^2) because the distance in the EEG to the electrical sources is prolongated and the electrical signals are volume conducted from their origin to the electrode (Vaughan and Arezzo, 1988).

An obvious source for EEG signals seems to be the neuron's action potentials, which produce signals up to 80-100μV. But since action potentials are very short (1-2ms), summation of signals from a large group of neurons becomes very improbable. Thus, their contribution to the EEG signal is only minor (Zschocke, 1995). Studies assessing intracerebral and EEG data simultaneously did not find correlations between single neurons and the surface EEG (Schandry, 1981). Researchers agree that most electrical potentials recorded from the scalp stem from extracellular current flow, caused by electrical dipoles between the soma and apical dendrites of the *pyramidal cells*. Since these cells are oriented parallel in an *open field* manner, they are able to produce these potentials in a way that the potentials of single neurons can accumulate and be registered by EEG recordings.

The dipoles in turn are generated by excitatory (EPSP) and inhibitory (IPSP) postsynaptic potentials and emerge as follows: When an **excitation** takes place at an apical dendrite (see Figure 2.3, right) by, e.g., contralateral callosal neurons, positive Na^+ ions accumulate within the cell body at the apical dendrite. This will be compensated by negative ions in the extracellular space resulting in a negative environment within this area. These ions were recruited from the extracellular media and also from the soma's region. Thus, the positive potential in the extracellular space around the soma is a result from the flow of ions. On the other hand, if an excitatory synapse connects to the pyramidal cell nearby the soma, like thalamocortical neurons, the opposite effect results and the dipole changes its direction (see Figure 2.3, left). At an **inhibition**, negative ions flow into the cell or positive ions leave the cell and the dipole reverses polarity. Thus, a positive potential results for the electrode on the scalp when an inhibition takes place in higher cortex layers, or an excitation in deeper cortex layers (Schandry, 1981). Pyramidal cells have their origin mainly in layer 5 of the cerebral cortex. Therefore, predominantly events in this cortex area are registered by the EEG (Martin, 1991) and subcortical processes can only be inferred indirectly.

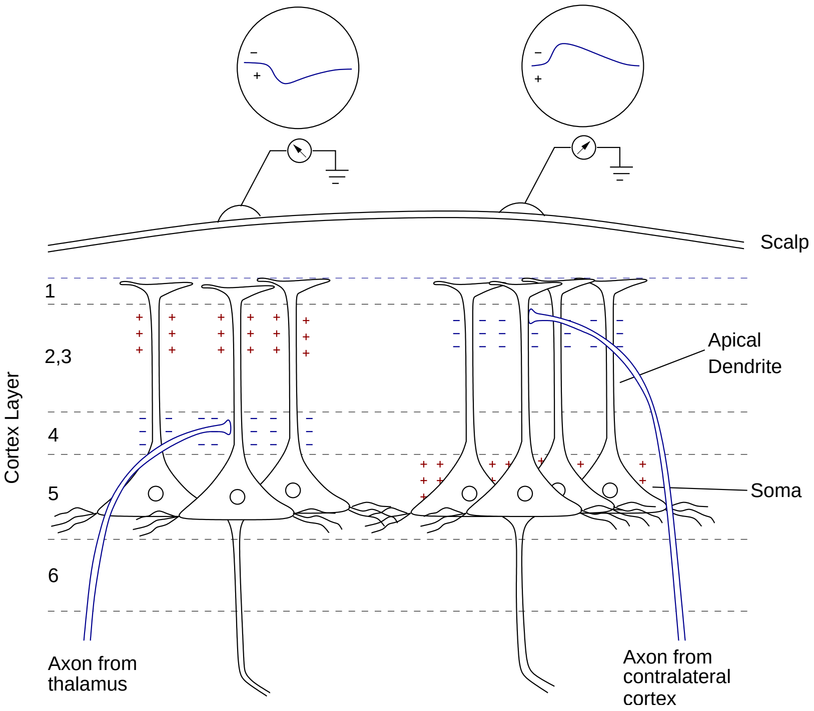


Figure 2.3: Constitution of EEG signals. **Left**: Electrical potential as recorded by the EEG following thalamic excitatory inputs. The terminals of the thalamocortical neurons connect to the pyramidal cell mainly in layer 4. This leads to an Na^+ influx in this region into the cell body. Reversely, the extracellular space becomes negative. Since the ions for this negativation stem from upper layers, these layers become more positive. Taken together, an electrical field in the extracellular space results. **Right**: Axons from the contralateral cortex connect to the apical dendrite in upper layers, and the same dynamic results for the opposite direction such that the dipole changes its direction. Adapted from Martin (1991).

2.3 Signal Acquisition and Artifacts

As stated above, in Electroencephalography, differences between electrical potentials of electrodes are measured. In a simple case, one electrode serves as a reference, and the potential from each other electrode is compared to that electrode's potential. While the reference electrode is usually attached to the ear, the other electrodes are located on the scalp - hence, their name *scalp electrodes*.

The most common system for electrode placement is the international 10-20 system (Jasper, 1958). According to the reference points *Inion*, *Nasion*, and left and right *preauricular points*, the head's surface is divided by steps of 10% and 20% of the whole length between reference points into characteristic locations (see Figure 2.4). For more detailed topographical analyses, more electrodes can be used, e.g., within a 10-10 or a 10-5 system.

In order to gain high quality data, a number of aspects has to be considered when assessing EEG data. EEG data reflect electric voltages and are thus very sensitive to electrical noise from the environment, such as the 50Hz or 60Hz power supply fre-

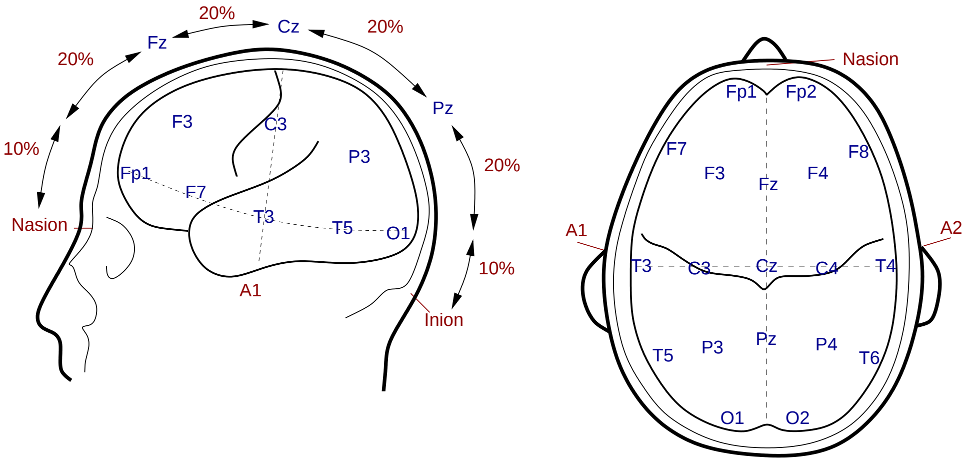


Figure 2.4: Electrode placement according to Jasper's international 10-20 system (Jasper, 1958). The distances between the reference points Nasion, Inion and preauricular points A1 and A2 are divided into steps of 10% and 20% resulting in electrode locations correlated to specific brain regions, like, e.g., the occipital lobes at $O1$ and $O2$. Adapted from Birbaumer and Schmidt (2005).

quency. Various kinds of those *artifacts*, i.e., observable signals which are not related to the biosignal the researcher is interested in, can be identified. They can roughly be divided into the 5 sources *electromagnetic induction, eye movements, muscular artifacts, movement artifacts*, and *skin and sweating artifacts* (Zschocke, 1995).

Electromagnetic Induction

Since EEG signals are voltages with only little amplitudes, they can easily be polluted by electromagnetic influences. These can result from any electrical source like computers, monitors, and power supply. Especially the 50-60Hz frequency of the power supply can be filtered out by *Notchfilters* (Zschocke, 1995). On the other hand, with this technique, also signals the researcher might be interested in, can get lost. However, electromagnetic influences can be reduced by using a *Faraday cage*. In general, a good way to enhance the signal quality is to keep the impedance between scalp and reference electrodes low, i.e., commonly below $5k\Omega$. Additionally, the subject and the electrical devices should be grounded.

Eye Movements and Blinking Artifacts

The front of an eye ball has a positive potential as compared to its back. Thus, the eye builds a dipole and movements of the eye ball can influence the scalp potentials. In particular, they affect signals as measured from frontal sites. These artifacts primarily stem from vertical eye movements as they happen by, e.g., closing the lid. It is well-known as the *Bell-Phenomenon* that the eyes move upwards when the lid closes. Thus, lid closing commonly results in artifacts. A related mechanism exists for blinking, which should therefore be prohibited by instruction and the experimental design. One way to detect these artifacts is measuring the *Electrooculogram* (EOG, see Figure 2.5): Electrodes are applicated left and right or above and below the eye balls. They register shifts in the electrical potentials which can directly be interpreted as eye movements. A common method is to exclude EEG trials where the EOG exceeded a specific threshold

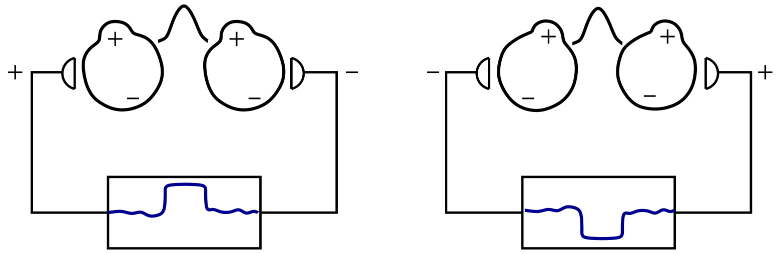


Figure 2.5: Detecting eye movements with an Electrooculogram. Since the back and the front of eye balls expose different electrical potentials, eye movements result in voltages as recorded by electrodes located left and right or above and below the eyes.

(e.g., 100μV). Other methods subtract the weighted EOG signal from the scalp electrodes (Gratton and Coles, 1989) or eliminate their influence by utilizing *Independent Component Analysis* (Jung et al., 2000).

Muscular Artifacts

The innervation of muscles is an electrical process, hence all muscular movements are possible sources for interferences. In particular movements of the heart as measured by the *Electrocardiogram* (ECG) and muscles nearby the EEG electrodes like in the neck and at the forehead are important sources for artifacts. Neck muscle artifacts are mostly caused by a lack of relaxation and can therefore be reduced by relaxation techniques or pharmaceuticals. The ECG influences are vital, but can be recognized in the EEG by their rhythmic structure.

Pulse and Movement Artifacts

Pulse artifacts are closely related to ECG artifacts but are not based on the electrical character of the heart beat. Instead, they are based on pulsating blood vessels in close proximity to the EEG electrodes. This can result in potential shifts within the electrodes. Pulse artifacts can be reduced by slightly moving the electrodes, putting in more electrode gel, or changing the pressure of the cap. Beside movements through pulse artifacts, they can also be caused by the mechanical changes of the subject's position, which are also accompanied by muscular activations. A common kind of movement artifacts is caused by breathe, where breathe-synchronous head movements are carried forward to the EEG electrodes and their cables. Similar to ECG artifacts, breathe artifacts can be identified by their frequency of about 0.25Hz. Furthermore, movements of the chest can be recorded. By changing cable or head positions, these artifacts can be reduced.

Skin and Sweating Artifacts

Further artifacts may arise from skin anomalies and sweating. Skin diseases, hairspray, and lardy hairs are likely to cause malfunctions by affecting the contact to the electrode. Thus, a head wash prior to the experiment can be useful for the latter two cases. Sweating artifacts are the most common skin related artifacts. They can be caused by electrical potentials of the sweating glands, by changing the skin's electrical resistance, and by the emission of sweat itself. This must not necessarily be accompanied by visible sweating. Sweating artifacts predominantly occur at frontal sites and are characterized by slow and very high potential changes. To avoid sweating artifacts, the experimental room should be cooled. On the other hand, sweating is often a psychological reaction caused by stress or fear. In order to reduce these symptoms, the experimenter should talk, explain and try to provide a less threatening, comfortable social situation for the subject.

2.4 Analyzing EEG Data

As discussed in the previous section, EEG data are likely to be polluted by artifacts and are affected by electromagnetic noise and therefore expose a low *signal-to-noise ratio* (SNR). Thus, specific data analysis techniques are usually necessary to receive interpretable information. Depending on the focus of the researcher or therapist, different techniques for EEG data collection and analysis are suitable for different purposes.

2.4.1 Data Collection Techniques

Among the EEG data collection techniques, either *spontaneous* or *evoked* EEGs can be recorded. The latter one can further be analyzed as *Event-Related Potentials* (ERPs).

The **spontaneous EEG** relies on unprovoked neural activity in absence of an identifiable stimulus (Salek-Haddadi et al., 2003). This kind of EEG data can mainly be categorized on the basis of amplitude and frequency and is widely used for monitoring purposes in anesthesia or for cerebral death identification. But also epileptic discharges and sleep stages can be identified from the spontaneous EEG. In recent years, several efforts have also been conducted to employ the spontaneous EEG for the purpose of Brain-Computer Interfacing, as discussed in detail in section 3.2.

In contrast to the spontaneous EEG, **evoked potentials** are EEG data following a stimulus. They can be analyzed as single trials or as a collection of trials as in ERPs. The goal of single trial analysis is to analyze EEG data following single stimuli, which can especially be useful for preprocessing purposes in cognitive science by, e.g., calculating the correct latency of each trial performed in a series of trials in order to perform temporal shifts (Jaskowski and Verleger, 2000). As in the spontaneous EEG, single trial analysis can also target on Brain-Computer Interfacing (cf. section 3.2).

A common way to overcome the bad signal quality of EEG data is to calculate **Event-Related Potentials**, which will be discussed in detail in section 2.5. By repeating stimulus expositions and recording the resulting EEG time series, the ERP can be calculated by averaging these time series. Thus, the SNR is enhanced and those parts of the signal which are correlated with the (psychological) properties of the stimulus accumulate, while non-systematical voltage changes (noise) are averaged out. Data can either be averaged according to the stimulus onset (*stimulus-locked*) or the response of the subject (*response-locked*).

ERPs only reflect evoked potentials with a close temporal relationship to the event. In order to be able to analyze components with temporal variations, so-called *induced potentials* can be analyzed by performing power spectra calculations, and average the power spectra. Afterwards, another transformation back to the amplitude domain can be performed to receive a time series signal (Tallon-Baudry and Bertrand, 1999).

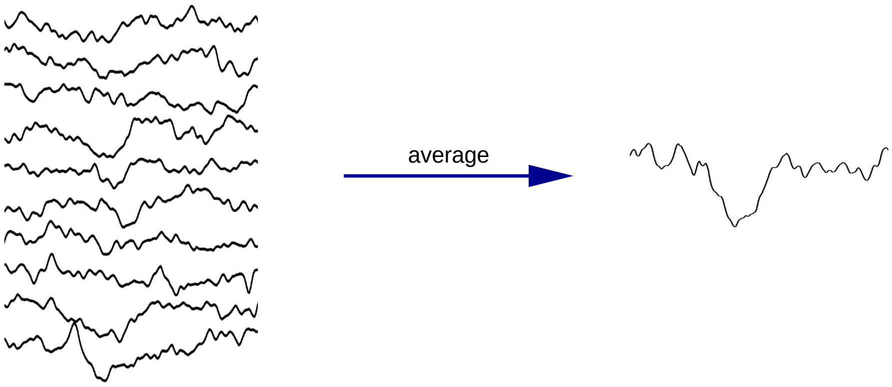


Figure 2.6: EEG recordings belonging to repetitions of a certain stimulus (event) can be averaged to compute the Event-Related Potential. Thereby, parts of the signal correlated to the event remain while non-systematical influences are regarded as noise are reduced with this procedure.

2.4.2 Data Analysis Techniques

EEG data analysis commonly relies on *amplitude changes* or *frequency properties*, which can be further processed to assess *coherence* and *phase relationships*. In order to become able to perform even more sophisticated analyses like, e.g., source localization, the EEG data can be *combined with other imaging techniques*. In the past decade, modern data analysis techniques inspired by *artificial neural networks* became increasingly important. These different aspects of EEG data analysis will shortly be discussed in the following; further information is provided by Coles et al. (1986).

First, an apparent characteristic of EEG signals is their **amplitude** changes. They are foremost analyzed in Event-Related Potentials. Section 2.5 gives an introduction into analyzing these components.

A complementary aspect of EEG data is their **frequency**, which importance was early recognized: Berger (1931) discovered decreases in frequencies about 10Hz (so-called α waves) during sleep, anesthesia and cocaine stimulation. Table 2.1 gives an overview about the five different frequencies which are commonly distinguished today[1]: δ (0.5-4Hz), θ (5-7Hz), α (8-12Hz), β (13-30Hz), and γ (>30Hz). Each frequency band is correlated with certain aspects of cognitive processes: While, e.g., α waves correspond to deep relaxation, β waves occur in awake humans and γ band activity can reflect memory processes. A common way to perform the transformation into the frequency domain is *Fourier transformation* (FT). For instationary signals like EEG, *Short-Term Fourier transformation* (STFT) can be employed on windows of EEG data. Since the size of the STFT time window directly affects the frequency resolution, in recent years *Continuous Wavelet transform* (CWT) became popular for power spectra calculations - they expose different resolutions for different frequencies.

Large-scale integration, i.e., communication processes between cell-assemblies at dis-

[1] Depending on the focus of research, further subdivisions of either frequency, e.g., β-1 (13-18Hz) vs. β-2 (19-31Hz) or predominant location, like in α (occipital, 8-12Hz) vs. μ (motor cortex, 8-12Hz) can be found. The specific frequency ranges vary between researchers.

Table 2.1: EEG frequency bands and related cognitive states or processes.

Frequency Band	**Frequency (Hz)**	**Cognitive State / Processes**
δ	0.5-4	deep sleep, coma
θ	5-7	drowsiness
α	8-12	relaxed but awake, esp. with closed eyes
β	13-30	active, busy or anxious thinking
γ	>30	e.g., perceptual binding and memory processes

tant scalp locations can be investigated by analyzing their **coherence** properties (Weiss and Müller, 2003; Weiss et al., 2005). Coherence is reflected by power spectra correlations of data from two electrodes and is commonly computed for specific frequency bands (Rappelsberger and Petsche, 1988). Thus, frequency analysis can be employed to calculate coherence. Furthermore, phase relationships provide information about the direction of the communication between neurons, whether a certain cell assembly sends information to or receives information from another one (Weiss and Müller, 2003; Varela et al., 2001). Beside correlating power spectra, alternative sophisticated approaches exist, like estimating coherence and phase relationships from adaptive autoregressive moving average (ARMA) models (Schack and Weiss, 2005). In this model, electrode couplings are calculated by computing cross-correlations between electrodes based on autoregressive functions.

In order to overcome the low spatial resolution of the EEG and its restriction on cerebral cortex activity, efforts have been made to **combine** this technique with other **imaging techniques** like fMRI. This combination enables researchers to perform, e.g., localizations of generators of EEG signals (Bledowski et al., 2004).

The rise of **artificial neural networks** in all its variations permitted to analyze EEG data in new ways (Kaper et al., 2006). For example, Hidden-Markov-Models were successfully utilized to identify sleep stages (Flexer et al., 2005). Independent Component Analysis was employed to, e.g., perform artifact removal (Jung et al., 2000), feature extraction (Meinicke et al., 2004), and to determine independent EEG components (Makeig et al., 2004). Multi-Layer Perceptrons classified mental states (Anderson, 1997) and Self-Organizing Maps allowed classifications and convenient exploratory data analyses (Heuser et al., 1997; Kaper et al., 2005). For Brain-Computer Interfaces, a large variety of such approaches has been applied, foremost for classifying EEG data (see section 3.2).

2.5 Event-Related Potentials

Components in Event-Related Potentials are classified by the polarity and latency of amplitudes. Thus, negative deflections after about 100ms are assigned as a *N100* component and positive deflections with a latency of about 300ms are designated as a *P300* component[2] (cf. Figure 2.7). Based on this classification scheme, a number of components can be distinguished. The most common components are the N100, N200, N400, and P300 components. These components can further be divided into subcomponents.

[2] By omitting the trailing zeros, components can also be abbreviated to, e.g., N1 and P3 in this case.

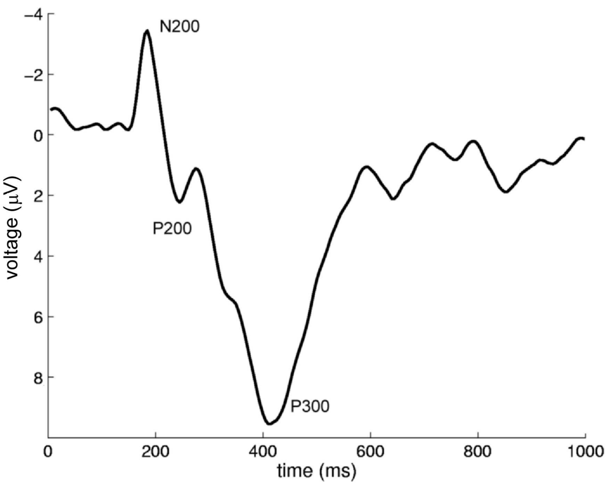


Figure 2.7: ERP components are classified by their polarity (P/N) and their latency. Positive deflections at about 300ms after stimulus exposition are therefore designated as P300 components while a negative deflection at about 200ms would be regarded as a N200. Please note that it is common in EEG research to depict reversed polarities.

A more detailed overview of the different ERP components is provided by Patel and Azzam (2005), Coles and Rugg (1995), and Coles et al. (1990).

N100 Component

The N100 component's peak latency is between 90ms and 200ms. The component is elicited by novel or unexpected stimuli. Its occurrence does not depend on the attention to the stimulus, thus it is regarded as being related to the *orienting response (OR)* as introduced by Sokolow (1963). Nevertheless, its amplitude can be enhanced by directing attention to the stimulus (Hillyard et al., 1973). The N100 habituates with repetitions up to its disappearance. It is suspected that the psychophysiological basis for the orientation response is a comparison of new incoming stimuli with previously stored stimulus features, such that an OR results whenever no suitable neural representation of the incoming stimulus exists.

N200 Component

The N200 component is correlated to stimulus evaluations by the subject and is divided into the subcomponents N2a, N2b, and N2c, respectively. One way to elicit a **N2a** is to present rare deviating stimuli in a series of similar stimuli - thus, the alternative designation *Mismatch Negativity* (MMN) (Näätänen, 1982). Since this component particularly results when attention is *not* directed to the stimulus, it is assumed that this component is related to preattentive processes and represents an automatic novelty-sensing process (Picton et al., 2000). The **N2b** exposes a prolongated latency and is elicited under similar circumstances as the N2a. But in contrast to the N2a, it only occurs when subjects selectively attend to the rare stimuli and is therefore supposed to reflect deviations to mentally-stored expectations of the standard stimulus. The N2b commonly precedes a P300 (see section 2.5). If the stimuli are quite similar, the latency of this component is prolongated. While the N2b has its maximum amplitude at

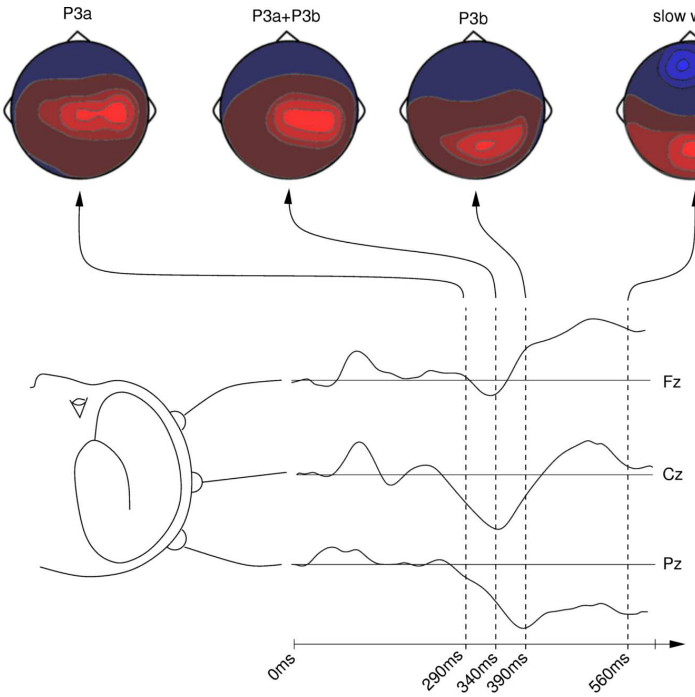


Figure 2.8: Typical topographical distribution of ERP components. While the topographical maximum of the early P3a components is located at central sites, the maximum shifts to dorsal sites with prolongated latency of the components P3a+P3b, P3b, and slow wave. Adapted from Picton et al. (1995).

parietal sites, the **N2c** maximum amplitude is fronto-central. A N2c results when the subject's task is to classify stimuli into categories. It is suspected that the N2c consists of components for memorizing, recalling and categorizing stimuli.

N400 Component

In nonsense sentences, where the last word does not fit to the semantic context, like in *"the pizza was too hot to cry"*, a N400 results for the word "cry"(Kutas and Hillyard, 1980). Furthermore, this component occurs when the word is semantically correct, but makes no sense in the broader context. Thus, not the grammatical properties of the word are responsible for eliciting the N400, but the probability of the target word's occurrence in that context. Hence, the N400 is associated with expectation violations. This phenomenon is not restricted to linguistic material: The modification of well-known melodies also result in this component.

P300 Component

Sutton et al. (1965) discovered a positive wave with a latency of about 350ms which occurred when few target stimuli (*oddballs*) are presented in a series of background stimuli. Therefore, this general experimental setup is also known as the *oddball-paradigm*, which is also the most common way to elicit a P300 component: Within a series of background stimuli, rare stimuli (i.e., the oddball-stimuli) are presented. subjects are instructed to concentrate on the oddball-stimuli by, e.g., counting their occurrences. Then, for these stimuli, a P300 results in the subject's EEG pattern. This finding was first utilized by Farwell and Donchin (1988) in order to design a Brain-Computer In-

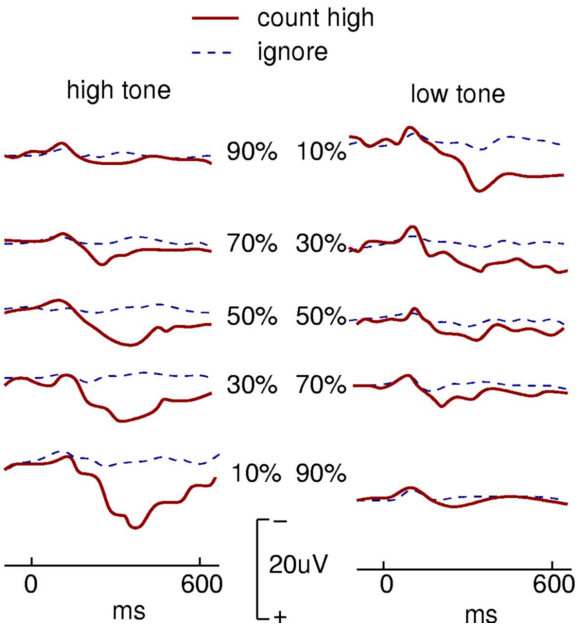


Figure 2.9: Probability dependence of the P300 amplitude. In this experiment, two different auditory stimuli, a high and a low tone, were presented. The probability of their occurrence was varied between conditions and the subjects were instructed to either count the high tone (bold, red line) or ignore the stimuli (dashed blue line, control condition). No remarkable effects could be observed for the latter condition, while in the former condition, the P300 amplitude decreases with the likeliness of the stimulus. This effect happens for attended (left column) as well as for unattended stimuli (right column). The P300 amplitude is higher for the attended stimuli (Duncan-Johnson and Donchin, 1977).

terface. Even though most experiments utilize auditory stimuli for P300 research, the component is multi-modal and can also be elicited by visual or even tactile stimuli. Extensive overviews about the P300 are given by Polich (1998) and Pritchard (1981).

Although the P300 is a well-studied component, researchers differ about its psychological meaning. Furthermore, it is more accurate to talk about a P300 *group*. This group is divided into the subcomponents P3a, P3b and *slow wave*. Figure 2.8 illustrates their temporal and spatial characteristics. For novel target stimuli, a pronounciated **P3a** with a latency between 250ms and 350ms will result. The amplitude's maximum is at Fz (fronto-central, see Figure 2.4). It also occurs when the attention is not directed to the stimuli and it can easily habituate. The component is therefore regarded as being correlated to an automatized orienting response. The “classical”P300 component is the **P3b** component. This component is elicited by the oddball-paradigm and has its maximum amplitude at Pz (centro-parietal cortex, see Figure 2.4). Depending on the modality, its latency lies between 340ms and 700ms after stimulus onset. The P3b is *endogenous*, i.e., it depends on the subject's interpretation of the stimulus rather than its physical properties (which in turn would be correlated to *exogenous* components). For example, Klinke et al. (1968) demonstrated that the P3b component can also be elicited by *missing stimuli*. In a regular series of “click”-sounds, an expected “click”was missing, which resulted in a P3b. Thus, the presence of “no stimulus”with obviously no

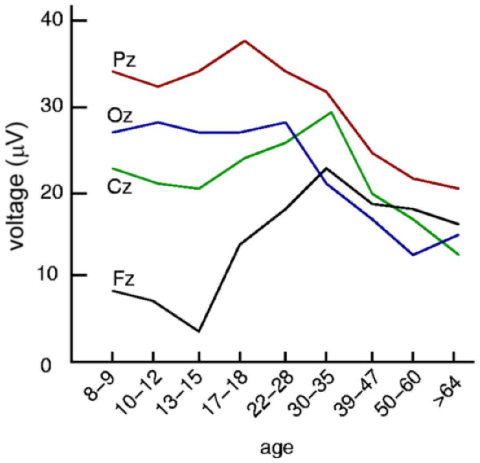


Figure 2.10: Age effects on P300 Amplitude. While strong differences of the P300's maximum amplitude can be observed in children and teenagers between scalp sites, these voltages converge to a medium level with increasing age. Adapted from Mullis et al. (1985).

physical properties is able to induce the P3b. A mandatory precondition for eliciting a P3b is directing the attention towards the target stimuli. Some researchers believe the P3b reflects *context updating*, i.e., the adaptation of a mental model to the incoming oddball stimulus (Donchin, 1981). Others are convinced that it reflects a *context closure* procedure, i.e., the termination of a waiting process for an anticipated stimulus (Verleger, 1988). More details about P3a and P3b components can be found in Polich (2003).

Slow waves, also denoted as *slow cortical potentials* (SCPs), with a latency of 600ms up to 1400ms occur whenever the stimuli are relevant for the solution of a demanding task. They regulate thresholds of excitability of cell assemblies and results for, e.g., complex thinking processes under time pressure. By operant conditioning, subjects can learn to gain voluntary control over their SCP which led Birbaumer et al. (1999) to design a Brain-Computer Interface, the *Thought-Translation Device* (TTD, see chapter 3.2.3), utilizing this component. Birbaumer et al. (1990) provides a review of SCPs.

Effects on the Amplitude and Latency of the P300

The P300 highly depends on the **probability** of the target stimuli. The rarer the stimuli, the higher is its amplitude. A clear P300 amplitude results with target probabilities between 15% and 20%, but its amplitude wont increase any further below a probability of 10%. Figure 2.9 illustrates the findings of Duncan-Johnson and Donchin (1977) who conducted an experiment with varying probabilities of a high tone (1500Hz) relative to a low tone (1000Hz). Either the high tone should be counted (red line) or every tone should be ignored (blue line). With sinking probability, the amplitude of the P300 increases. It is important to distinguish between *global* and *local* target probability. The latter refers to a series of few succeeding stimulus expositions. Even with a low global target probability, the amplitude for a second target stimulus decreases when it is exposed shortly after a target. The **interstimulus interval** (ISI) denotes the temporal distance of two stimuli. The amplitude of the P300 is negatively correlated with this temporal distance. It is possible to induce a P300 with an ISI even below 300ms, which is very helpful in designing a P300-based Brain-Computer Interface as described in chapter 4. The component's amplitude and latency does not stay constant throughout life. With increasing **age**, its latency prolongates (with about 1.4ms per year) and its scalp distribution shifts after about 30 years of age. In elder people, the P300 appears to display a more equipotential scalp distribution across the scalp midline for visual, auditory and somatosensory stimuli (Mullis et al., 1985; Friedman et al., 1989). The Pz amplitude ranges between 25μV and 40μV (see Figure 2.10). These are important facts to keep in mind, when targeting to design a classifier working for different subjects (see chapter 6.5). A subject needs to be awake and to be able to focus his **attention** towards the target stimuli. Lowered attention results in decreased amplitudes, which was experimentally proven by introducing a distracting task within an oddball-paradigm. As a result, the amplitude decreased with increasing attention towards the distracting task. The higher the reward for correctly recognized stimuli, the higher the amplitude. Thus, the **relevance** of the stimuli also plays an important role. On the other hand, this might just mediate attention. After **meals** and within **summer time**, the amplitude increases. The latency of the P300 decreases of about 30ms for an increase of 1° Celsius in **body temperature**. Gender and menstruation cycle do not affect the P300.

2.6 Summary

Electroencephalography (EEG) measures potential differences between electrodes on the scalp. Due to the small currences at single neurons, only synchronous activities from assemblies of several hundred thousand neurons can be assessed by the EEG. Nevertheless, the signal-to-noise ratio remains low, and the signal is likely to be polluted by artifacts like, e.g., electromagnetic induction and muscular activity. One way to overcome this handicap is to calculate so-called Event-Related Potentials (ERPs), i.e., averaged EEG time series belonging to certain experimental conditions. Resulting components are classified by the latencies and polarities of peaks in the ERP. The P300 component is a positive deflection after about 300ms. It occurs when in a series of background stimuli few target stimuli (oddballs) which are relevant to the observing subject, are presented. Even on the basis of very few trials, a P300 can be identified which led Farwell and Donchin (1988) to design a Brain-Computer Interface utilizing this component.

Chapter 3

Brain-Machine Interfaces

Brain-Machine Interfaces (BMIs) establish a connection between the brain and an artificial device to stimulate the brain or to receive information from it (Nicolelis, 2001). In the former case, the BMI provides *input* to the brain, which can be employed to restore lost sensory functions as in an artificial auditory prosthesis, or to suppress symptoms from brain diseases like *Parkinson's disease* (Nicolelis, 2001; Donoghue, 2002). In the latter case, the BMI receives *output* from the brain, such that control over an artificial unit can be achieved, which can especially be useful for paralyzed patients. An example for intended users are thereby so-called *locked-in* patients. This syndrome denotes a state where people are literally locked-in in a motionless body, but in contrast to *coma*, these people are conscious and have an awake mind (Kübler et al., 2001). Jean-Dominique Bauby, a former editor-in-chief of *Elle France*, wrote about his experiences in such a state. He suffered a stroke and was since then only able to perform residual movements of his head and left eye. He wrote a whole book by employing just movements of the eye (Bauby, 1998). At worst, locked-in patients are not able to move *any* muscle and are therefore unable to communicate or express desires. As Patterson and Grabois (1986) describe, *"these patients are aware of both internal and external stimuli but are able to carry on only an internal monologue"*. Such patients could be equipped with a BMI system to open up a communication channel when using their brain signals to e.g., control a spelling device.

As depicted in Figure 3.1, Brain-Machine Interfaces can either be **invasive** or **non-invasive**. While invasive methods rely on penetrating the body and depend on surgery, non-invasive methods do not require any incision into the body. Electroencephalography (see chapter 2) is the most common method to acquire data among the latter BMI approaches. Non-invasive BMIs are also called *Brain-Computer Interfaces* (BCIs)[1]. Another subdivision of BMIs relies on the direction of the information transfer: While the brain is stimulated in **input** BMIs, brain signals are employed to deliver information to a device in **output** BMIs. At present, non-invasive approaches can reasonably only be designed as output devices. Although non-invasive stimulation is in general possible by, e.g., *transcranial magnetic stimulation* (TMS), where strong magnetic fields affect brain activity, this method is too rough and offers only a very low spatial resolution, such that it is not suitable for BMI purposes at present.

[1] This categorization of BMIs and BCIs follows the distinction of (Nicolelis et al., 2004) and the reader might find deviating declarations. For example, "BCI"can also subsume invasive approaches (Wolpaw et al., 2002), and instead of invasive vs. non-invasive BMIs, some authors talk about *direct* vs. *indirect* BMIs (Donoghue, 2002).

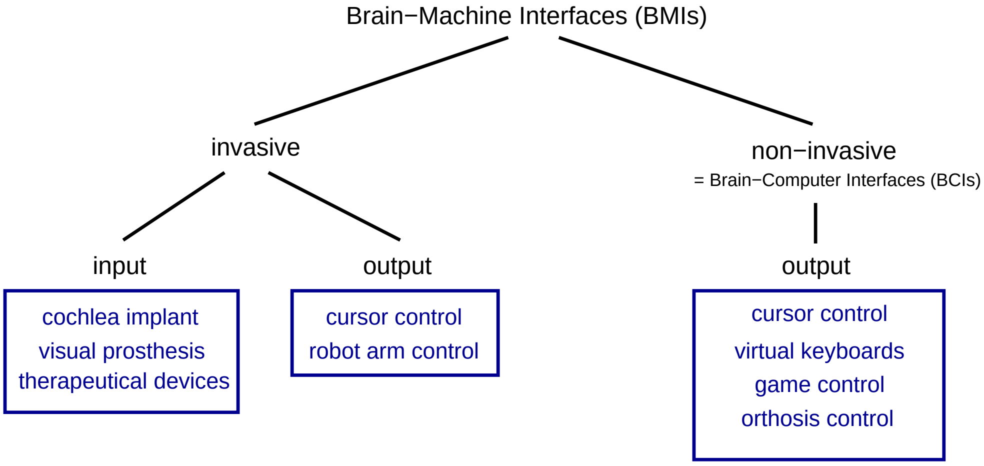


Figure 3.1: Brain-Machine Interfaces can be divided into invasive and non-invasive approaches. While invasive approaches require the penetration of the body, non-invasive approaches do not depend on surgery. The direction of the information flow constitutes another dimension: BMIs can either be input or output devices. The former stimulate the brain and can be employed to, e.g., restore lost sensory functions, while the latter ones receive information from the brain to enable the subject to, e.g., control a computer. Exemplary applications are given in the blue boxes.

This chapter gives a review about different BMIs with a focus on EEG-based Brain-Computer Interfaces. Examples for invasive input BMIs, mainly auditory and visual prostheses as well as therapeutical devices, are discussed in section 3.1. Output BMIs, enabling to control cursors or robot arms are considered in the same section. Non-invasive BMIs or Brain-Computer Interfaces, mostly providing spelling devices or *virtual keyboards* intended for paralyzed people, are presented in chapter 3.2. Finally, the concept of information transfer rates (ITR), which is important for performance comparisons of BMI approaches is introduced in section 3.3.

Further introductions and overviews can be received by, e.g., Kübler et al. (2001), Alfa (2005), Wolpaw et al. (2002), Wickelgreen (2003), Nicolelis (2001), Nicolelis (2003), and Donoghue (2002). Progress in invasive recordings are discussed in Engel et al. (2005) and a detailed introduction into neuroprosthetics in general is given by the book of Horch and Dhillon (2004).

3.1 Invasive Brain-Machine Interfaces

Invasive Brain-Machine Interfaces establish a direct connection to the brain. Electrodes are implanted into or in close proximity to the brain and therefore provide good signal qualities, temporal and spatial resolutions. These factors allow for large bandwidths compared to non-invasive approaches. On the other hand, these approaches face the risks of bioincompatibility, often preventing long-term studies. Furthermore, ethical problems have to be considered, and the benefits of the surgical treatment should clearly countervail its risks. In the following, examples of realized input and output devices are presented.

Input Invasive Brain-Machine Interfaces

The first BMI application was the **cochlea implant** (Clark et al., 1981). This auditory prosthesis was designed for deaf people and translates features of acoustic signals as recorded by microphones into electrical stimuli which are delivered to implanted electrodes nearby the auditory nerve fibers on the basilar membrane of the ear's cochlea (Pfingst, 2000). Within the cochlea, adjacent frequencies correspond to neighboring locations. Although about 40000 nerve fibers are located within the cochlea, present day implants only provide as few as 6-22 electrodes. Thus, the quality of the auditory stimulation remains low. Nevertheless, such implants allow to perform, e.g., telephone calls (Qian et al., 2003). Cochlea implants for adults were approved by the *U.S. Food and Drug Association* (FDA) in 1985 and for children in 1990. In 2002, the *National Institute on Deafness and Other Communication Disorders* (NIDCD) indicated that 59000 people have received a cochlea implant, aging from 12 months to 80 years (NIDCD, 2006).

Similar interfaces were also invented for blind people: A **visual prosthesis** is based on neuronal electrical stimulation at specific locations along the visual pathways. Those prostheses exist for the different locations *retina*, *optical nerve*, and *visual cortex*. Implants providing stimulations to one of the first two locations are used when the visual loss is caused by outer retinal degeneration, while visual loss caused by inner or whole thickness retinal diseases, eye loss, optic nerve diseases, or diseases of the central nervous system can be reversed by a cortical visual prosthesis (Maynard, 2001; Margalit et al., 2002). A *cortical* visual prosthesis utilizes the finding of Löwenstein and Borchart (1918) that electrical stimulations of the visual cortex result in visual impressions, so-called *phosphenes*. Dobelle reported a first visual cortex prosthesis, allowing blind people to recognize simple patterns (Dobelle et al., 1974, 1976). Such a cortical visual prosthesis uses signals of a video camera and stimulates the visual cortex by a 64 channel platinum disk electrode array on the surface of the visual cortex. This device enabled patients to recognize 6-inch characters at 5 feet distance. One of his patients was wearing this implant for more than 20 years (Dobelle, 2000).

Recent research for cortical visual prostheses focuses on using penetrating intracortical microelectrodes (Fernandez et al., 2005). Schmidt et al. (1996) reported the implantation of 38 microelectrodes into the right visual cortex of a 42-year old woman who was blind for more than 22 years. 34 of the electrodes produced phosphenes. Since the implant was not designed for long-term usage, the experiment was performed for only 4 months. At present, long-term viability and biocompatibility are two of the main frontiers of invasive Brain-Machine Interfacing (Vetter et al., 2004; Fernandez et al., 2005).

Beside compensations for sensory loss, a further direction of BMI research targets on developing **therapeutical devices** to alleviate the symptoms of brain disorders like Parkinson's disease or epilepsy by electrical stimulation of certain brain regions or nerves. In *deep brain stimulation* (DBS) for example, electrodes are stereotactically implanted into certain regions of the patient's brain. Although also considered for, e.g., Huntington's disease and epilepsy (Fawcett et al., 2005; Hamani et al., 2005), the technique is most commonly used for Parkinson's disease. Thereby, electrodes are placed into the *subthalamic nucleus* (STN) or *globus pallidus interna* (GPi) within the basal ganglia (Kumar et al., 1999). Afterwards, electrical stimulation of these regions can alleviate most Parkinson symptoms like tremor, slowness of movements, dyskinesia and difficulty with balance and walking (Anderson et al., 2005; Kumar et al., 2003). DBS for Parkinson treatment was FDA approved in 1997 for unilateral thalamic regions and in 2002 for STN and GPi bilaterally. Worldwide, more than 14000 Parkinson patients have received a DBS implant (Medtronics, 2005).

In order to suppress epileptic seizures, *vagus nerve stimulation* (VNS) was successfully employed (Schachter, 2002). A recent long-term study on 48 patients reported a decrease of the mean seizure frequency from 74% after one year to 48% after 12 years with VNS treatment (Uthman et al., 2004). VNS was FDA approved in 1997, and since then VNS devices were implanted to more than 30000 patients (Cyberonics, 2005). VNS has also been considered as a therapy for depression (Hoppe et al., 2001) and the treatment of pain disorders (Kirchner et al., 2000). Some findings suggest that it could improve cognition and memory (Clark et al., 1999).

Beside these medical attempts, efforts have been conducted to **gain control over the brain**: By stimulating the *medial forebrain bundle* (MFB), intense pleasure can be produced. Utilizing this phenomenon as a reward system, and simultaneously stimulating sensorimotorcortex areas in the rat's brain, which receive sensory inputs from the whiskers, Talwar et al. (2002) succeeded in literally steering rats with implanted electrodes. They see applications in, e.g., land mine detection and search-and-rescue missions.

Output Invasive Brain-Machine Interfaces

In order to control a cursor, a prosthesis or a robotic device, the vast number of invasive output BMIs employ motorcortex signals. Such techniques could offer paralyzed patients the possibility to gain control over an artificial device substituting lost motor functions to control a computer or an artificial device like an orthosis.

Chapin et al. (1999) performed pioneering work in this area by deriving signals from the brains of rats in order to steer a robot arm. In the following, Wessberg et al. (2000) analyzed brain signals as recorded by microwire arrays from two owl monkeys. They applied microwires to several cortex areas, primarily in the motor cortex. For 12 and 24 months, the monkeys were trained to perform one-dimensional hand movements with a lever. Afterwards, also three-dimensional hand movements were trained by grasping food from one out of four positions in front of the monkeys. The goal of the researchers was the real-time approximation of the original hand movements of the monkeys (as measured by the lever position) as precisely as possible. For this purpose, in a first step, they employed coherence analysis, and, afterwards, the microwire data were analyzed by either linear models or multilayer artificial neural networks, both continuously updated. In the one-dimensional case as well as for three-dimensional movements, the researchers achieved highly significant real-time predictions for both monkeys. The predictions were utilized to control a local and a remote robot arm as well, indicating that it is possible to gain real-time control over artificial limbs by brain signals. Another outcome of the study was that chronically implanted microwires arrays can yield reliable signals in the BMI context for as long as 24 months.

In a subsequent experiment, two macaque monkeys were trained to perform not only reaching, but also grasping (Carmena et al., 2003). In this *closed-loop* BMI, the monkeys received visual feedback about their actions. First, in a *pole control mode*, the monkeys operated a pole to move a cursor on a computer screen. One task was to move the cursor towards a randomly located disc on the screen. Further tasks also

Table 3.1: Order of magnitudes of spatial and temporal resolutions of different imaging techniques. Note that although NIRS offers higher temporal resolutions in principle, the assessed hemodynamical signals induced by brain activity result with a few seconds delay.

method	**EEG**	**MEG**	**fMRI**	**PET**	**NIRS**
spatial resolution	cm	cm	mm	cm	cm
temporal resolution	ms	ms	s	min	s

included the application of gripping force to the pole to perform grasping behavior. The screen delivered visual feedback about their actions and after a number of training trials, the pole was plugged off, and the monkeys were required to move the cursor by solely using their brain signals. In this *brain control mode*, the animals first produced arm movements but realized soon that those were not necessary and discontinued them for periods of time. The investigators therefore physically removed the pole after the animals ceased to produce the movements in a session. Muscle activity as measured by *electromyography* (EMG) at the three different locations *wrist flexors*, *wrist extensors* and *biceps*, indicated that both animals were able to operate the system without muscle activity.

Several other groups conducted studies proving that motor neurons from monkeys can provide BMI control signals for 2D and 3D control (Taylor et al., 2002; Serruya et al., 2002; Mehring et al., 2003). But also first studies with humans were conducted: Kennedy et al. (2000) implanted neurotrophic electrodes in three patients suffering from motor diseases. The electrodes were implanted into the motorcortex and the growth of neural tissue into the hollow electrode tip was encouraged by using trophic factors. Two wires within the electrode registered intra-cortical *local field potentials* (LFPs). The patients learned to perform binary decisions to control a software system for communication (*TalkAssist*). Despite the technical efforts, the most successful patient produced only about 3 letters per minute. One patient used this interface for as long as 4 years. More recently, in order to minimize surgical complications, the group applied skull screws to assess extra-cortically LFPs in two paralyzed patients suffering from *Amyotrophic Lateral Sclerosis*. Another group applied microarray electrodes into the human motorcortex in order to provide control over steering devices. FDA approved a test series on five patients, but only popular science publications can be obtained so far (Duncan and Friedman, 2005). Both groups are involved in first commercial spin-offs: *Neural Signals Inc.* (NSI) and *Cyberkinetics Inc.* each distribute invasive Brain-Machine Interfaces to enable paralyzed patients to communicate (NeuralSignals, 2006; Cyberkinetics, 2006).

3.2 Non-invasive Brain-Machine Interfaces

In non-invasive BMIs, so-called Brain-Computer Interfaces, no penetration of the body is performed. Since Electroencephalography is easy to use, comparably cheap, and, most of all, offers a very high temporal resolution (see Table 3.1), this technique is by far the most frequently employed method for BCIs, although some work exists employing *functional Magnetic Resonance Imaging* (fMRI) (Weiskopf et al., 2004; Hinterberger et al., 2004b), *Magnetencephalography* (MEG) (Laitinen, 2003; Nykopp et al., 2005), and

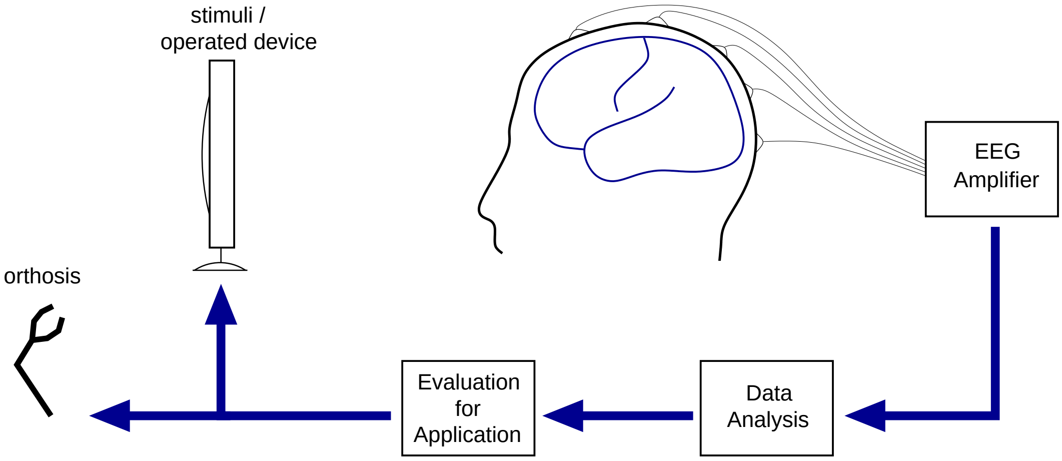


Figure 3.2: Illustration of an EEG-based Brain-Computer Interface. Depending on the paradigm, stimulus presentation can be necessary to induce specific brain activities or to provide feedback. EEG signals are recorded from the subject's scalp and a data analysis procedure yields features which can be evaluated for specific applications like controlling an orthosis or operating a virtual keyboard.

Near-Infrared Spectroscopy (NIRS) (Coyle et al., 2004). Figure 3.2 gives an illustration of a general EEG-based Brain-Computer Interface.

Since spontaneous EEG data itself does *a-priori* not provide interpretable information to operate a BCI, any BCI is closely related to a specific paradigm. The different approaches within non-invasive Brain-Computer Interfacing can be divided into the basic approaches *mental tasks*, *steady-state visual evoked potentials (SSVEP)*, *slow cortical potentials (SCP)*, *sensorimotorcortex activity* and *P300 evoked potentials* which will be explained in detail in the upcoming sections.

Beside the distinction of **invasive vs. non-invasive** Brain-Machine Interfaces, particularly the non-invasive approaches can be categorized further: Along a first dimension, BCIs can be distinguished according to whether they require **training of the subjects** or not. Training time can last from almost no training (P300 and SSVEP) up to several months (SCP). Another dimension is the **stimulus-dependence** of a BCI. Some BCIs depend upon the presentation of a stimulus to be able to induce, depending on the subject's attention, a certain brain signal (P300 and SSVEP) which can be problematic for designing real-time devices. A similar distinction lies in **synchronous vs. asynchronous** BCIs. The former BCIs depend upon the presentation of a cue as a trigger, while the latter ones need to analyze spontaneous EEG signals without time markers. Finally, BCIs can be divided into **dependent and independent** devices: *"an independent BCI does not depend in any way on the brain's normal output pathways"* (Wolpaw et al., 2002). Such a "normal output pathway"could be eye movements to direct the gaze. The SSVEP BCI is therefore a dependent BCI, since the eyes are directed towards a specific location which then result in modulations of the EEG signal. The current P300 speller device is probably dependent (Kaper et al., 2004) because the subjects commonly direct their gaze to specific locations of a screen. On the other hand, the confoundation of fixation and locus of attention could in principle be resolved (Posner et al., 1980) by instructing to fixate a specific point on the screen and shift the attention to other locations. Furthermore, e.g., auditory P300-based BCIs are possible (see section 8.1).

3.2.1 Mental Tasks

Different brain areas are predominantly involved in certain mental operations. Processing language-related material commonly results in strong activations at temporal sites on the left (see section 2.1). On the other hand, imagining a rotating cube yields right hemispheric activations. Reversely, such findings can be used to voluntary induce specific patterns to provide information for communication purposes. Keirn and Aunon (1990) instructed four subjects to perform one out of five randomly chosen tasks in a trial which should induce specific EEG patterns. For a duration of 10 seconds, the subjects should perform one of the following tasks which were repeated five times within the experiment:

Baseline task: Relax as much as possible.

Letter task: Mentally compose a letter to a friend without vocalizing.

Math task: Perform nontrivial multiplication problems like 49 × 78.

Visual counting task: Imagine sequentially written numbers on a blackboard.

Figure rotating task: Visualize and rotate a 3-dimensional block figure around an axis.

Trials of the different conditions were classified using *Multilayer Perceptrons* (MLPs) with backpropagation learning (Anderson and Sijercic, 1996). Feature vectors were Fourier transforms based on 6th order autoregressive coefficients from the 6 channels $C3, C4, P3, P4, O1$, and $O2$ (see Figure 2.4). Depending on the subject, classification accuracies between 38% and 70% were achieved, resulting in transfer rates between 0.74 bits/min and 5.05 bits/min. This means that by using this device 0.74 to 5.05 bits of information (Shannon and Weaver, 1949) can be transferred within one minute. Since the concept of information transfer rates is important for Brain-Computer Interfacing in general, it will be discussed in detail in section 3.3. Although this transfer rate might be too little and the tasks are probably too demanding for communication purposes, the finding that it is possible to identify these complex tasks by brain signals with an accuracy of up to 70% encourages to employ such techniques for, e.g., clinical purposes.

3.2.2 Steady-State Visual Evoked Potentials

Changes in visual stimuli result in *visual evoked potentials* (VEP), which can foremost be recorded at occipital sites. If stimulus changes occur below a frequency of 2Hz, the evoked potentials are denoted as VEPs. If a visual stimulus is presented repetitively at a rate of more than 6Hz, a periodic response called *steady-state visual evoked potential* (SSVEP) will result. A useful property of SSVEPs is that their frequency as measured in the EEG is the same as the frequency of the initiating stimulus (Regan, 1989). The amplitude of the SSVEP can be enhanced by directing attention to the location of the

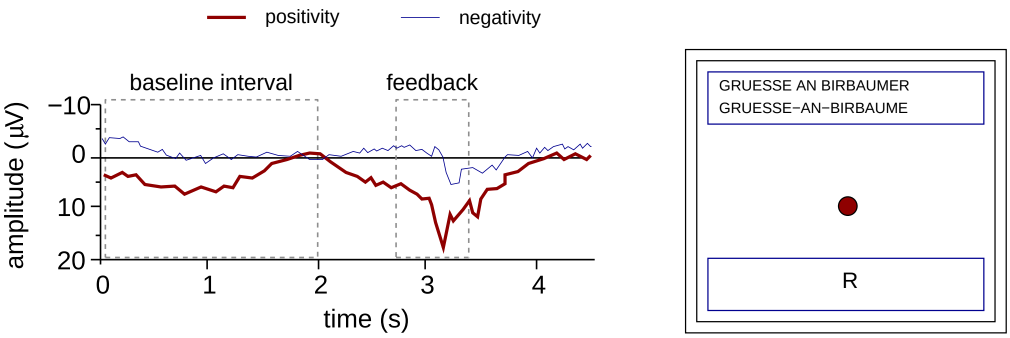


Figure 3.3: **Left:** After determining a baseline within the first 2 seconds of a trial, slow cortical potentials can either produce negative (thin blue line) or positive shifts (thick red line) to provide information. **Right:** A virtual keyboard controlled by slow cortical potentials. At the bottom, letters or a collection of letters occur and the user needs to move the ball from the center either towards the box on the bottom, or keep it away from it (Kübler et al., 1999).

flickering stimulus, which was experimentally proven for the frequency ranges of 8-12Hz (Morgan et al., 1996) and 20-28Hz (Müller et al., 1998).

These properties can be employed to construct a Brain-Computer Interface. Since the production of SSVEPs is an inherent mechanism of the brain, no training of the subjects to control their brain activity is necessary. Cheng et al. (2002) presented a 3×4 matrix (and an additional button) to their test subjects filled with buttons of numbers mimicking a telephone keypad. The buttons were flashing at different frequencies within the range of 6-14Hz. From the power spectrum of the data, which was recorded from the occipital sites $O1$ and $O2$ (see Figure 2.4), predominant frequencies were extracted. From the frequency in turn, the buttons on the screen could be inferred. Seven subjects participated in the experiment and their mean performances ranged from 3.05 bits/min to 48.93 bits/min with a maximum transfer rate of 55.69 bits/min. Another experiment performed by the same group utilized as many as 48 flickering light-emitting diodes (LEDs) of 6-16Hz, resulting in 68.00 bits/min maximum transfer rate (Gao et al., 2003).

An Air Force research laboratory employed SSVEPs as control signals for a flight simulator (Middendorf et al., 2000). By directing attention to one of two flickering stimuli, binary decisions (roll left/right) were provided. An accuracy of 80-95% has been achieved but no information about the time required for one decision was reported. Further on, the researchers successfully operated a *functional electrical stimulator* (FES) for knee angle commands. In their best sessions, three able bodied subjects achieved 95.8% of the required knee angles with latencies between 4.28s and 5.93s.

Recently, attempts were made to utilize this technique for game control (Lalor et al., 2004). Five subjects performed offline trials to find good signal processing parameters and classification techniques. Afterwards, the subjects played a game steering a fantasy figure balanced on a rope. By directing attention to one of two checkerboards at the screen, flickering at different frequencies, SSVEPs were induced, commanding the figure to either shift its weight to the left or to the right. The investigators suggested to use this technique for subjects with *attentional deficit hyperactivity disorders* (ADHD).

Since this approach highly depends on eye position, it is regarded as a *dependent* BCI. People need to move their eyes, and the eye position modulates the brain signals.

3.2.3 Slow Cortical Potentials

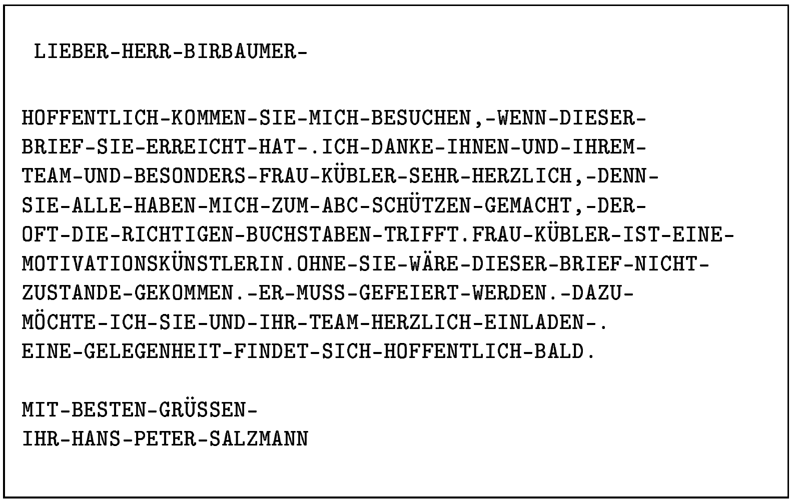
Slow cortical potentials (SCP) are potential shifts in the EEG signal over 0.5-10s (see section 2.5). The group around Birbaumer has shown that people can learn to control their SCPs (Birbaumer et al., 1990). At the beginning, this finding was used to enable people to gain some control over their epileptic seizures (Elbert et al., 1991), but it also turned out that this signal could be used to provide communication signals, which resulted in the so-called *Thought Translation Device* (TTD) (Birbaumer et al., 1999, 2003; Kübler et al., 1999, 2001). Recently, the TTD has further been successfully utilized for neurofeedback in children with attentional deficit hyperactivity disorders (Holtmann et al., 2004).

In a pioneering study, two patients suffering from advanced Amyotrophic Lateral Sclerosis which were not able to use muscle-driven interfaces and have been artificially respirated and fed for four years were trained to employ their SCPs to provide control signals for operating a spelling device. Satisfying skills in voluntary producing SCP changes were achieved after 327 and 288 sessions, respectively. The patients performed 6-12 sessions on a training day, each including 70 to 100 trials of 5-10 minutes. A trial consisted of a 2-second baseline and a response period of 2-4 seconds (see Figure 3.3, left). Within an operant conditioning scheme, control over the SCP was trained by providing visual feedback of the SCP amplitude to the patients. The training task for the patients was to move a ball on a video screen either to a box in the upper half or to a box in the lower half. The direction of the ball movement was calculated from the baseline-response difference in the SCP voltage. In the test phase, as depicted in Figure 3.3 (right), letters were selected by moving the ball towards a box in the bottom containing changing letters, or keep it away from it using the SCPs. With this device, the completely paralyzed patients regained the possibility to communicate, which was formerly prohibited due to their severe motor diseases. One of the patients wrote a letter to the leader of the group, which is depicted in Figure 3.4.

The TTD software was augmented with a word completion algorithm for a quick selection within a dictionary of 500 commonly used words (Hinterberger et al., 2004a). Another extension is a module for operating an internet browser (Mellinger et al., 2003) within which links of each website are marked by green and red frames. By either choosing the green or red set of links, subsequently only the links of the chosen color remain and will again be divided into red and green links. After a number of such binary decisions, each rejecting 50% of the remaining links, one link to follow can be chosen after a number of decisions. Although this device requires training and does not produce the highest information transfer rates, it is very remarkable since it has proven its real-world usability and is currently in use by several patients (Birbaumer et al., 2003) who became thereby able to express their thoughts and desires.

3.2.4 Sensorimotorcortex Activity

As it is introduced in section 2.1, the preparation and the execution of a movement is commonly accompanied by a decrease in μ (8-12Hz) and β (18-26Hz) waves primarily on the opposite (*contralateral*) side in motorcortex areas. Beside these *event-related desynchronizations* (EDS), also *event-related synchronizations* (ERS), an increase in



LIEBER-HERR-BIRBAUMER-

HOFFENTLICH-KOMMEN-SIE-MICH-BESUCHEN,-WENN-DIESER-
BRIEF-SIE-ERREICHT-HAT-.ICH-DANKE-IHNEN-UND-IHREM-
TEAM-UND-BESONDERS-FRAU-KÜBLER-SEHR-HERZLICH,-DENN-
SIE-ALLE-HABEN-MICH-ZUM-ABC-SCHÜTZEN-GEMACHT,-DER-
OFT-DIE-RICHTIGEN-BUCHSTABEN-TRIFFT.FRAU-KÜBLER-IST-EINE-
MOTIVATIONSKÜNSTLERIN.OHNE-SIE-WÄRE-DIESER-BRIEF-NICHT-
ZUSTANDE-GEKOMMEN.-ER-MUSS-GEFEIERT-WERDEN.-DAZU-
MÖCHTE-ICH-SIE-UND-IHR-TEAM-HERZLICH-EINLADEN-.
EINE-GELEGENHEIT-FINDET-SICH-HOFFENTLICH-BALD.

MIT-BESTEN-GRÜSSEN-
IHR-HANS-PETER-SALZMANN

Figure 3.4: Letter written by a locked-in patient using slow cortical potentials. The patient thanks the investigators for providing the spelling device and for being motivated by one of the researchers. He wishes to invite the investigators to celebrate this success. It took the patient about 16 hours to write the letter down (Birbaumer et al., 1999).

these rhythms exist, which occur *after* a movement and with relaxation. The spatiotemporal ERD and ERS patterns are similar for actual performing movements or just imagining these movements (Pfurtscheller and Neuper, 1997; Neuper and Pfurtscheller, 1999). Additionally, Kornhuber and Deecke (1965) reported a negative deflection at the contralateral side of a movement in central cortex regions, the so-called *Bereitschaftspotential* (readiness potential). Since this phenomenon also occurs when a movement is just imagined, it can be exploited for Brain-Computer Interfacing.

Taken together, by just imagining a movement, specific brain signals (ERD, EDS, Bereitschaftspotential) can voluntary be produced by a subject to provide information. A number of groups employed this finding for Brain-Computer Interfaces for *unidimensional* and *multidimensional* cursor control as well as for some applications as it is outlined in the following.

Unidimensional Control

Most work within this paradigm targeted on exploiting the μ waves over motorcortex areas like $C3$ and $C4$ (see Figure 2.4) to provide one-dimensional control signals. Those signals were employed to control different types of virtual keyboards: Within the approach of (Wolpaw et al., 2003), a cursor is moving from the left to the right and its vertical position can be influenced by the μ rhythm. Thus, the cursor can select a row at the right side containing letters or symbols by its vertical position (see Figure 3.5, left). Using this approach, subjects achieved 20-25 bits/min.

Pfurtscheller et al. (2003) recorded and analyzed α and β bands of two bipolar EEG channels[2] while performing two kinds of motor imagery. Employing ERD and ERS for μ rhythms, Pfurtscheller's group has built a virtual keyboard with Hidden-Markov-Models (HMM) as classifiers. In one experiment, performed with three able-bodied

[2] Electrodes were located 2.5cm anterior and posterior to $C3$ and $C4$, respectively (see Figure 2.4).

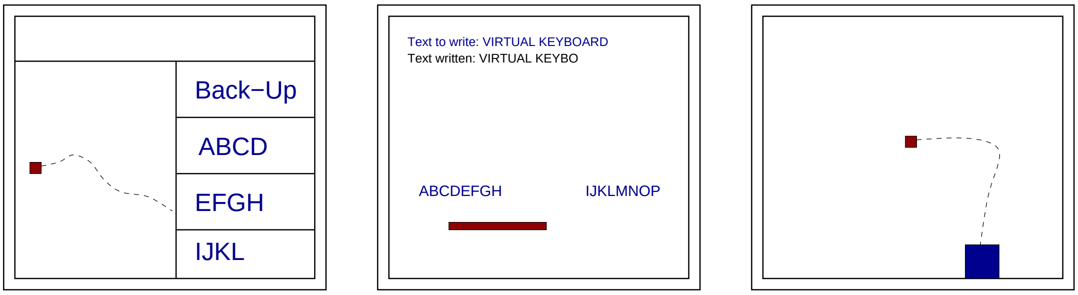


Figure 3.5: BCI applications employing sensorimotorcortex activity. **Left:** Movement imaginations result in up- and downwards shifts of a cursor moving from the left to the right. At the right side, a box containing symbols can be selected (Wolpaw et al., 2003). **Middle:** A bar can be shifted to the left or the right by movement imaginations, resulting in the selection of a set of letters (Pfurtscheller et al., 2003). **Right:** Two-dimensional movements are performed by independently controlling μ and β rhythms (Wolpaw and McFarland, 2004).

subjects, each subject employed a different imagination strategy. One subject imagined right versus left hand movements, another one right hand versus tongue, and the final subject left hand versus foot movements. In training sessions, data were acquired for HMM training (Obermaier et al., 2001). Afterwards, from the screen's center, a bar was extended to either the left or the right side by imagining the according movement (see Figure 3.5, middle). In the first step, 32 letters, including some special symbols, were divided into two subsets, each displayed on one side of the screen. By successive steps of isolation from the initial set, the correct letter can be chosen after 5 steps of binary decisions and two further steps of confirmation and correction. Using a word completion system with a dictionary of 145 common words, the number of selection steps could be reduced down to 4, resulting in an increase of the transfer rate from 0.67-1.02 letters/min up to 1.06-4.24 letters/min (Pfurtscheller et al., 2003).

The BCI system was evaluated in a field study on a population of 99 subjects, each performing short experiments of 20-30 minutes. After two initial training session, 93% of the subjects achieved above 60% classification accuracy (Guger et al., 2003).

Beside constructing virtual keyboards, some further applications have been designed for this paradigm. Pfurtscheller et al. (2003) developed a hand orthosis for a tetraplegic patient. The patient learned to control the orthosis almost perfectly by imagining either foot movements or right hand movements. After a training period of six days, he was able to perform about 6 opening/closing operations of the hand within a minute.

Multidimensional Control

As will be discussed in detail in section 3.3, increasing the number of choices has a strong impact on information transfer rates. Therefore, strong efforts have been conducted to provide control over more than just one dimension. Two different strategies can thereby be distinguished: Using *more than two different movement imaginations* and *independent control of μ and β rhythms.*

When performing **more than two different movement imaginations**, beside movements of the hands, also those of feet or the tongue are employed. Since the feet's position in the cerebral cortex lies within the rolandic fissure (see section 2.1), reversed polarity is observed at the scalp compared to hand and feet movements. In Scherer et al. (2004), three subjects were instructed to imagine left hand, right hand, and foot movements. The former virtual keyboard of Pfurtscheller et al. (2003) was extended to left, right and lower presentation of letters. Three Fisher's Linear Discriminant Analysis classifiers were trained to solve the classification problem, each trained to distinguish two classes (cf. section 5.3.3). Without fixed time constraints, the user could move the cursor within a graphical user interface in an asynchronous fashion to spell letters. The mean overall spelling rate was 1.99 letters/min.

Dornhege et al. (2004) employed imaginations of left hand, right hand, and foot movements as well. Their focus was on preprocessing and statistical data analysis for speed improvements. For this purpose, the group combined features, specialized on different aspects of the psychophysiological process like the Bereitschaftspotential, desynchronization dynamics and spatial patterns. They further broadened the concept of *Common Spatial Patterns* (CSP) (Koles and Soong, 1998) to multiple classes and into *Common Spatial Spectral Patterns* (CSSP), also considering spectral information (Lemm et al., 2005). In Blankertz et al. (2003) information transfer rates of up to 50.5 bits/min were reported. As an example for a gaming application, the group designed a *Brain-Pacman* game (Krepki et al., 2004).

Another way to achieve multidimensional control utilizes the finding that it is possible to gain **independent control of μ and β rhythms** over left and right sensorimotor cortices (Wolpaw and McFarland, 1994; Wolpaw et al., 2003). In a recent experiment, four subjects were trained to operate a 2-dimensional BCI in 2-4 sessions a week, each consisting of eight 3 minute lasting runs (Wolpaw and McFarland, 2004). The subjects performed 22-68 sessions, and the last three sessions were statistically analyzed. EEG amplitudes in the specific frequency bands were determined by autoregressive frequency analysis, and were the basis for calculating horizontal and vertical movements. Target stimuli appeared on one out of eight possible positions on the periphery of a computer screen, and a cursor was shown on the screen's center which could be controlled by EEG activity (see Figure 3.5, right). The subjects were instructed to hit the target with the cursor within 10 seconds. In the average, targets were reached within 1.9-3.9 seconds. When comparing the results with those obtained from invasive studies in monkeys (Taylor et al., 2002; Serruya et al., 2002; Carmena et al., 2003), time and accuracy as well as hit rates of this approach were within the range of the invasive investigations. Suspicions that non-invasive BCIs are not appropriate for efficient real-time control (Nicolelis, 2001; Fetz, 1999; Donoghue, 2002) could therefore be rejected.

3.2.5 P300 Evoked Potentials

BCIs utilizing the P300 component rely on the phenomenon that rare and significant stimuli reliably induce a P300 component, as already discussed in section 2.5. Farwell and Donchin (1988) employed this phenomenon for Brain-Computer Interfacing by presenting a stimulus matrix with flashing symbols. By directing the attention to a specific symbol, P300 components result when that specific symbol flashes. Thus, by identifying P300 components in the EEG, the symbol can reversely be inferred. One advantage of this approach is that it is independent of training the subjects since the P300 component naturally occurs in the human brain under the described circumstances. Furthermore, it is capable to achieve high information transfer rates (see section 6.6). A drawback of this method is that it depends on (visual) stimulations. Since this whole thesis is dedicated to this paradigm, the whole next chapter gives a review of P300-based BCIs.

3.3 Information Transfer Rates

Since Brain-Computer Interfaces are unsatisfying slow at present (e.g. the letter in Figure 3.4 took the subject 16 hours to write it down), one of the main goals in Brain-Computer Interfacing is to accelerate the devices. But in order to become able to compare different approaches, it is necessary to have a suitable measure for the speed of such a device.

Several parameters have to be considered when determining the speed of a BCI. First, the number of choices N that can be performed within one trial has a strong impact on the performance of a BCI. If a user intends to write a letter using 26 standard characters, he would need to perform $\log_2(26) = 4.7$ trials at binary choices (Shannon and Weaver, 1949). On the other hand, only one trial would be required if the number of choices would be $N = 26$. However, these conclusions can only be drawn when perfect accuracy in symbol prediction is guaranteed. According to Wolpaw et al. (2000), in the case of imperfect classification accuracy $p < 1$, the information $B(N, p)$ as measured in bits transferred by such a device within a trial can be calculated by

$$B(N, p) = \log_2 N + p \log_2 p + (1 - p) \log_2 \frac{1 - p}{N - 1}.$$

The impact of the classification accuracy p on the transferred bits $B(N, p)$ in one trial is depicted in Figure 3.6 (left) for specific choices of N. Some important relationships can directly be observed in this graph: First, at binary choice ($N = 2$), 10% decrease in accuracy results in 50% reduced information transfer. Second, by increasing the number of choices, strong performance benefits can be achieved and only about 80% accuracy is necessary with $N = 4$ to achieve the same performance with 100% accuracy at binary choices[3].

Another important parameter for speed assessments is the trial duration t, since it determines how many trials can be performed within a specific time. It is therefore common to calculate the amount of ***bits per minute*** transferred by a BCI as

$$B(N, p, t) = \frac{60}{t} \left(\log_2 N + p \log_2 p + (1 - p) \log_2 \frac{1 - p}{N - 1} \right).$$

In Figure 3.6 (right), the impact of the trial duration on the ***information transfer rate*** (ITR) $B(N, p, t)$ in bits per minute for a fixed accuracy of 75% is depicted. As stated above, for the selection of one letter from a set of 26 letters, 4.7 bits are required. Therefore, dividing the number of bits per minute by 4.7 reveals the number of letters

[3] But note that the chance level is decreasing with the number of choices. By e.g. increasing the numbers of choices from 2 to 4, the probability to choose the correct class by chance decreases from 50% to 25%.

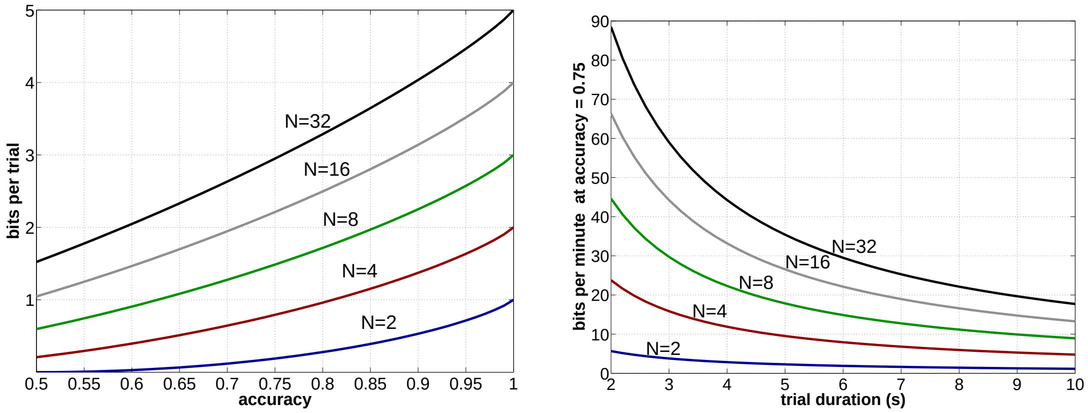


Figure 3.6: Information transfer rates. **Left:** Information in bits, transferred within one trial for different numbers of choices N and accuracies. **Right:** Information in bits per minute provided with specific numbers of choices for different trial durations at a constant accuracy of 0.75.

and, for example, about 6.4 letters could be produced with $N = 8$ and $t = 3$. An apparent effect exposed by the graph of Figure 3.6 (right) is that the trial duration highly affects the information transfer rate. This effect strongly increases with the number of choices.

Recently, a review comparing the mean performances of different BCI approaches and techniques was published by Serby et al. (2005) and is listed in Table 3.2. Note that most of the work reflects *theoretical* information transfer rates and e.g., ignore delays between trials. Unfortunately, it is not possible to calculate an information transfer rate in bits/min for the appealing work of Wolpaw and McFarland (2004), since their approach provides a continuous signal.

Thus, although the measure is quite fair and objective, it can not cover all systems and should not solely be employed for system comparisons. Alternative measures like the *Nykopp rate* (Kronegg et al., 2005) and *letters per minute* exist. At a constant information transfer rate, the latter measure can be enhanced by employing word completion algorithms. Further aspects, e.g., questions whether a system provides a continuous output signal, depends upon stimulations, is dependent or independent, or is in practical use, should also be taken into account.

In order to become able to compare the performance of different *algorithms* for BCIs, so-called *BCI Competitions* are conducted, in which BCI data sets are published which are to be classified by the competitors (Blankertz et al., 2004; Sajda et al., 2003). More details are described in section 6.6.

3.4 Summary

Brain-Machine Interfaces are devices between the brain and an artificial unit and can be designed as input or output devices. In input devices, signals from the artificial unit are transmitted to the brain in order to, e.g., restore lost sensory functions as in auditory prostheses. Furthermore, therapeutical stimulations of the brain can be performed to, e.g., alleviate symptoms of Parkinson's disease. In contrast, output devices utilize brain

Table 3.2: Comparison of different BCI approaches and algorithms of Serby et al. (2005) including earlier work of the author (Kaper and Ritter, 2004a).

Publication	Online System	Transfer Rate (bits/min)	Training Time	Number of Subjects
Donchin et al. (2000)	no	20.1	-	10 able
Donchin et al. (2000)	yes	9.23	-	5 (10) able
Babiloni et al. (2000)	no	2.65	-	5 able
Cincotti et al. (2003)	no	5.64	-	13 able
Levine et al. (2000)	no	3.46	-	17 patients
Birbaumer et al. (2000)	yes	2.35	Month	3 disable
Pfurtscheller et al. (2000)	yes	6.3	7 sessions	3 able
Pfurtscheller et al. (2003)	yes	9.48	Few weeks	4 patients
Wolpaw et al. (1991)	yes	10.88	2 months	4 (60) able
McFarland et al. (2003)	yes	8.49	Few months	8 (2 disable)
Kaper and Ritter (2004a)	no	47.26	-	8 able
Serby et al. (2005)	no	23.75	-	6 able

signals to steer an artificial device like a virtual keyboard or a robot arm.

Brain-Machine Interfaces can further be divided into invasive and non-invasive approaches. While the former ones rely on penetrating the body, the latter approaches need no incisions into the body. In recent years, several invasive approaches have been designed among which are the cochlea implant, visual prostheses, and therapeutical devices. It was even possible to steer the movements of rats using such an approach. Several invasive studies utilized brain signals from animals to steer robot arms and first studies in this area were conducted on humans.

Non-invasive BMIs, also denoted as Brain-Computer Interfaces, mostly utilize Electroencephalography to assess brain signals. They are embedded in specific paradigms, each focusing on certain brain processes. Current approaches employ EEG activations which arise when performing different cognitive operations (mental tasks), specific visual stimulations are provided to which the subject directs the gaze to (SSVEP), trained subjects learned to control their slow cortical potential (SCP), imaginations of motor behavior are performed (sensorimotor cortex activity), or the subject attends to rare and significant stimuli (P300).

The speed of an output device can be assessed by the *information transfer rate* which denotes the amount of information which is transferred within a specific time span. A widespread measure is to compute the number of bits transferred within a minute.

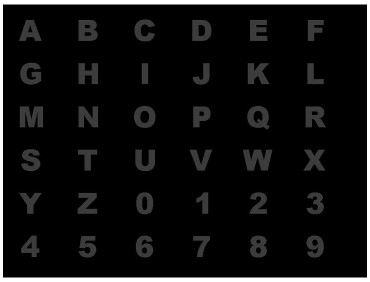
Employing the P300 component has the advantage of being independent of subject training and allows to achieve high information transfer rates as will be shown in the next chapter.

Chapter 4

Review of P300-Based Brain-Computer Interfacing

A P300 component is a positive deflection after about 300ms in the Event-Related Potential and can be elicited by rare stimuli to which a subject is directing the attention to (see section 2.5). This finding was employed by Farwell and Donchin (1988) to build a Brain-Computer Interface (BCI) in which symbols can be chosen from a matrix as depicted in Figure 4.1 to spell words. Later on, this so-called *P300 speller paradigm* was improved by Donchin et al. (2000) by using further classification techniques, and was also tested for disabled subjects. Since the P300 component is reliably elicited in the brain for the described stimuli, no training of subjects is needed. Furthermore, this approach allows for comparably high transfer rates because one out of (commonly) 36 symbols can be chosen within a choice, instead of, e.g., binary choices as in several other approaches (cf. section 3.2). Other researchers conducted further work with Brain-Computer Interfaces based on the P300 component. They performed experiments using paralyzed patients, achieved performance improvements, and extended the basic P300 speller paradigm to a more flexible interface. Additionally, further possibilities to exploit the P300 component in a BCI context were found.

A	G	M	S	Y	*
B	H	N	T	Z	*
C	I	O	U	*	TALK
D	J	P	V	FLN	SPAC
E	K	Q	W	*	BKSP
F	L	R	X	SPL	OUIT



A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	0	1	2	3
4	5	6	7	8	9

Figure 4.1: **Left:** Stimulus matrix with 6 rows and 6 columns including 26 letters from the alphabet and steering commands of Farwell and Donchin (1988). **Right:** Stimulus matrix as used in this thesis.

This chapter introduces into P300-based Brain-Computer Interfacing by giving a review of the work of Farwell and Donchin (1988) and Donchin et al. (2000) in section 4.1, followed by considerations about information transfer rates within this paradigm in section 4.2. Finally, in section 4.3, improvements of the basic P300 speller paradigm as well as related work utilizing the P300 component for Brain-Computer Interfaces are discussed.

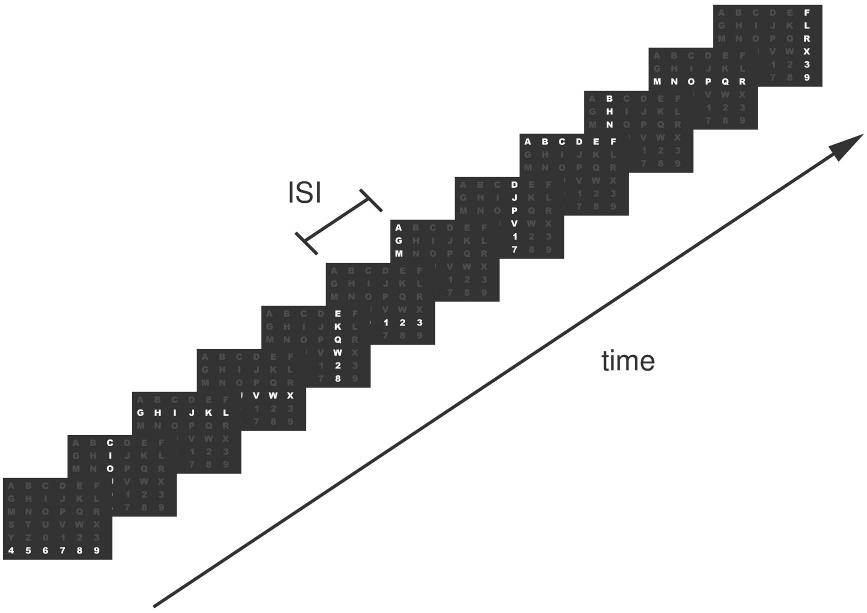


Figure 4.2: A complete sequence of 12 flashes, denoted as a *subtrial* in this thesis. Each row and each column is highlighted once in random order. The Interstimulus Interval (ISI) is the temporal distance between two highlightings.

4.1 The Basic P300 Speller Paradigm

The P300 speller paradigm relies on the P300 component, which can be elicited by rare stimuli which are relevant to the observing subject. It has its largest amplitudes at the location Pz and an onset latency of about 300ms varying with several parameters like age and body temperature (see section 2.4). The component is independent of modality, such that it occurs for visual as well as for auditory or tactile stimuli as introduced in section 2.5.

Farwell and Donchin (1988) utilized the P300 component in order to construct an EEG-based Brain-Computer Interface. They presented 36 symbols within a 6×6 matrix to their subjects (see Figure 4.1, left) which were instructed to choose one symbol from the matrix. Afterwards, the rows and columns of the matrix were flashing in random order and the subjects should mentally count how often their symbol is flashing (see Figure 4.2).

This setup provokes that a P300 is elicited when the row or the column containing the attended symbol (i.e., the *target row* or *target column*) is highlighted: Since only one out of six rows contains the symbol, and each row is flashing equally often, such an event is *rare* ($p = 0.17$; the same is true for columns). Furthermore, by instructing the subject to count the flashes of the specific symbol, the attention is directed to this event and it becomes relevant to the subject. Figure 4.3 (right) depicts the differences of the EEG patterns in the ERP as calculated from data belonging to the correct symbol (*target letter*), symbols within the same row/column as the target letter (*target row/column*), and symbols outside these regions (*standards*). The P300 amplitude in the target letter condition is enhanced compared to the other conditions. Thus, the row and column containing the target letter can reversely be identified by P300 occurrences. Exactly

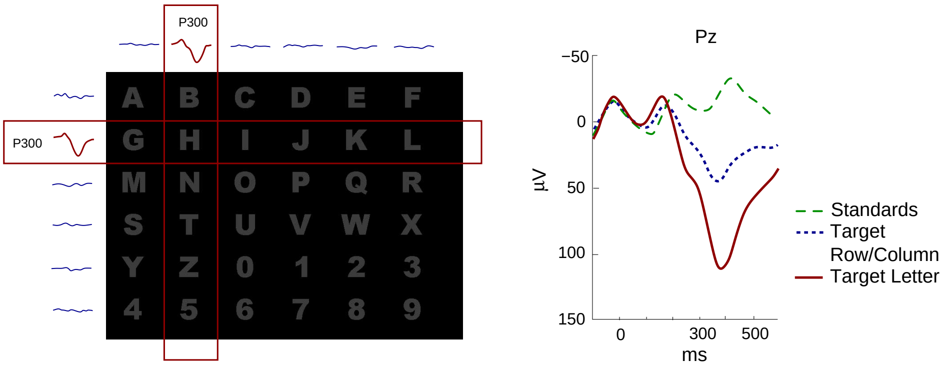


Figure 4.3: **Left:** From the six rows, the row with the most pronounciated P300 indicates the target row containing the target symbol. The same it true for the columns, such that the target letter can be identified by the intersection of the target row and the target column. Note that the depicted EEG signals are idealized and contain much more noise in real data. **Right:** ERP amplitude values from 10 subjects for ERP data according to the correct symbols (*target letter*), symbols from the correct row/column (*target row/column*), which are not the target letter, and other symbols (*standards*). Adapted from Donchin et al. (2000).

one row and one column should be accompanied by a large P300 amplitude after the flashing event. If a classifier succeeds in identifying these events, it can therefore infer the target letter: The intersection of the target row and the target column indicates the position of the target letter within the matrix (see Figure 4.3, left).

Unfortunately, EEG data expose a low signal-to-noise ratio, making it hard to identify P300 waves from single trials. Therefore, a number of repetitions is commonly needed to increase the signal-to-noise ratio by averaging data. Since such repetitions are time demanding, a main goal is to keep the number of repetitions low by developing well performing classification methods. Another way to enhance the speed of such a device is to decrease the so-called *Interstimulus Interval* (ISI). The ISI is the temporal distance between two flashes[1] (see Figure 4.2) and is therefore capable to decrease the duration of a trial and therewith increase the information transfer rate of the interface (see section 3.3). On the other hand, ISI reduction causes less pronounciated P300 components. Particularly below an ISI of about 600ms, P300 components resulting from a stimulus event might interfere with early components from succeeding events which can result in worse classification accuracies. Since ISI reduction results in a faster presentation speed but also in a lower P300 amplitude, an optimal ISI has to be found which allows for reasonable P300 classifications and high presentation speeds. Reversely, a good classifier enables the researcher to employ short ISIs, which in turn results in a faster device. Thus, in order to achieve a maximum speed for this device, an optimal trade-off between classification accuracy and presentation speed has to be found. Section 4.2 will discuss this topic in more detail.

[1] The correct term for the temporal distance of the onsets of two consecutive stimuli would be *stimulus onset asynchrony* (SOA). In contrast, ISI denotes the offset-onset distance of two stimuli. Nevertheless, the author will use ISI for the temporal distance of two events to stay in line with previous work (Farwell and Donchin, 1988).

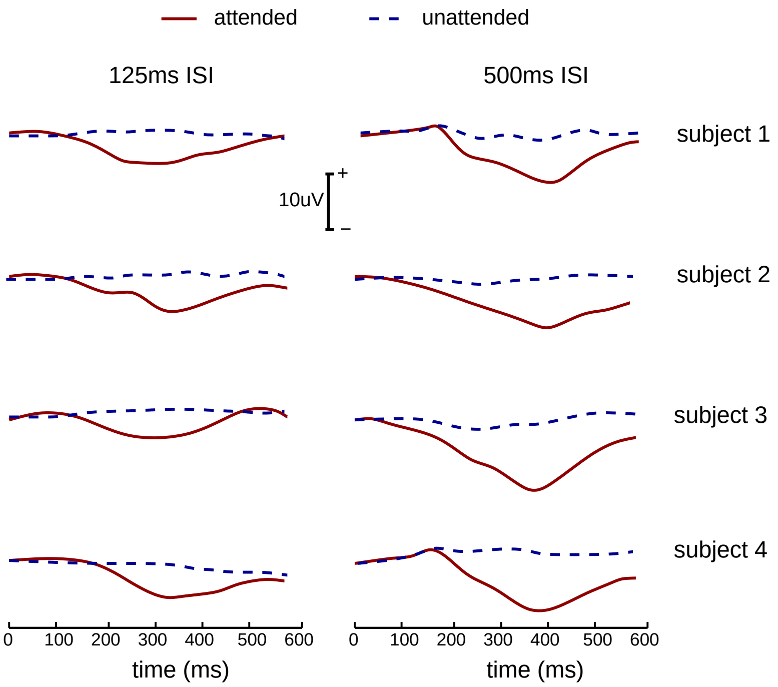


Figure 4.4: Average P300 amplitudes for 125ms and 500ms ISI for 4 different subjects. In both conditions, attended and unattended stimuli can clearly be distinguished by the P300 amplitude. On the other hand, in each subject, the P300 amplitude is lower for 125ms ISI compared to 500ms ISI. Adapted from Farwell and Donchin (1988).

Farwell and Donchin (1988) examined the impacts of the different ISIs 125ms and 500ms on the P300 amplitude. Figure 4.4 depicts their findings for P300 amplitudes from the Pz electrode of four subjects for both conditions. On average, 500ms ISI yielded base-to-peak amplitudes of 10.25μV for the target stimuli (1.65μV for unattended stimuli), while only 5.6μV for attended and 1.08μV for the unattended stimuli resulted with 125ms ISI. Taken together, large differences in the maximum amplitude between the 500ms and 125ms ISI condition were found. Employing the longer ISI yielded larger amplitudes which could presumably also better be detected by a classification algorithm.

4.1.1 Data Analysis

Before going into the details of data analysis, it is useful to introduce some definitions. The basic unit of the data analysis procedure is an *epoch*, i.e., an EEG time series following a highlighting. Farwell and Donchin (1988) employed 600ms time windows after stimulus presentation for the purpose of data analysis (see Figure 4.5). Twelve epochs, i.e., a complete series of flashes such that each row and each column was highlighted exactly one time, form a *subtrial*[2]. Within one *trial*, a certain number of subtrials is

[2] This distinction differs from the denotions of Farwell and Donchin (1988), where an epoch would be declared as a subtrial. But to have an expression for multiple highlightings of a complete sequence of 12 epochs in a trial, the author chose to denote this sequence a subtrial.

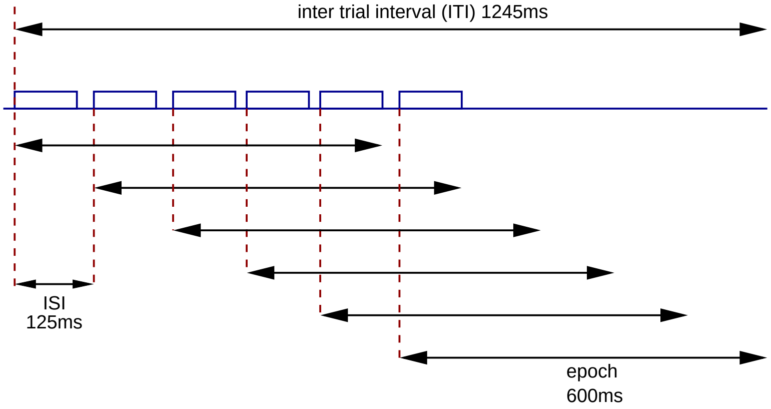


Figure 4.5: Time course of events in the 125ms ISI condition. A 600ms time series of EEG data (an *epoch*) was recorded, starting with the onset of a stimulus. Since the ISI is far below the epoch duration, an overlap of 475ms results for consecutive epochs. The inter-trial interval (ITI) is 1245ms and denotes the temporal distance between two trials, each containing 12 epochs.

performed.

For classification purposes, Farwell and Donchin (1988) employed the different techniques *area, peak picking, stepwise discriminant analysis* (SWDA), and *covariance* to detect P300 occurrences. *Area* and *peak picking* each employ a *P300 window*, i.e., a time window between onset and offset of the P300 deflection (see Figure 4.6). This window typically ranged between 220ms and 500ms. SWDA and covariance are data driven techniques and rely on disjoint training- and testsets of a collection of epochs. Parameters for these classifiers are adjusted on the training set and are then applied on the test set for classification purposes (see details about training- and testsets in section 5.3.1). The principles behind these classification techniques are as follows:

Area: Calculates the surface under the curve within the P300 window. It is reflected by the sum of the data points within this window.

Peak picking: Determines the difference between the highest positive peak within and the lowest negative point prior to the P300 window.

SWDA: Computes the distance of an epoch to the mean of a group containing P300 epochs as calculated from the training set. This score is obtained by applying a discriminant function to the data from each epoch.

Covariance: Assesses the covariance of an epoch with a template. The template is calculated as the average of epochs belonging to attended symbols in the training set.

4.1.2 Performance of the Classification Techniques

In a first study, Farwell and Donchin (1988) employed four subjects, each performing two sessions. While the feasibility of the technique was assessed and the subjects were familiarized with the system in the first session, the subjects produced data for assessing

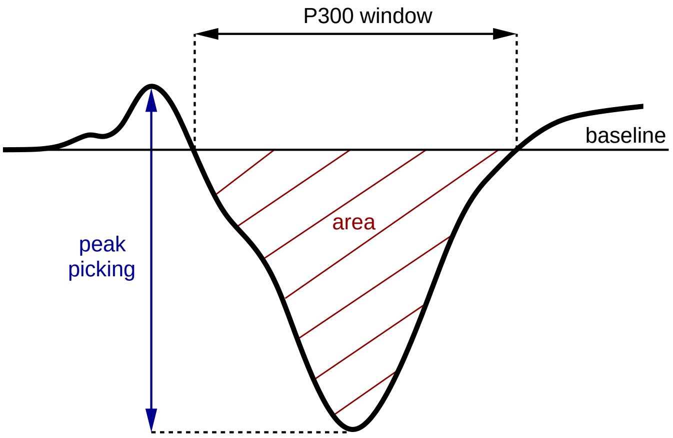


Figure 4.6: Classification techniques *area* and *peak picking*. While *area* calculates the surface under the curve within the P300 window, *peak picking* computes the peak-to-peak difference of the minimum before and the maximum within the P300 window. Note that the amplitudes are depicted with reversed polarity.

the operating characteristics of the system in the second study. EEG was recorded from the Pz electrode (see Figure 2.4), digitized at a rate of 50Hz and band pass filtered from 0.02Hz to 30Hz (cf. section 5.2.1). Data were analyzed in real-time in the first session, but only *after* the experiment in the assessment session, which is denoted as *offline* analysis in this context. The matrix from Figure 4.1 (left) was presented and highlighted with both 125ms and 500ms in the assessment session. The subjects were instructed to attend to a given letter and mentally count its highlightings. Five blocks of 30 trials (of one subtrial each) were performed for each ISI. The subjects were instructed to count the flashings of the letters “B”, “R”, “A”, “I”, and “N” in the different blocks. The four different classifiers from the previous section were applied to the data and their performance in predicting the correct letters was assessed.

Resulting durations to reach 80% and 95% classification accuracy for 125ms ISI and 500ms ISI, respectively, are summarized in Table 4.1. In the best case, 80% accuracy has been achieved after 11.1 seconds (subject 3, SWDA, 500ms ISI), and in the worst case after 114.8 seconds (subject 1, SWDA, 500ms ISI). Employing the best performing classifier for each subject, a range of 11.1 to 23.3 seconds would result for this criterion. The 95% accuracy level has been reached after 17.6 seconds in the best case (subject 3, SWDA, 500ms ISI), and in the worst case after only 202.8 seconds (subject 1, SWDA, 500ms ISI). According to the algorithm presented by Wolpaw et al. (2000), the best results correspond to information transfer rates (cf. section 3.3) of 18.50 bits/min and 15.77 bits/min for the 80% and 95% level, respectively. No general conclusions can be drawn about the classification method or the ISI: For the different subjects, different ISIs and different classification methods yielded the best results.

In a second study, Donchin et al. (2000) further analyzed this approach. Since one of the major challenges in BCI research is to provide paralyzed patients a device for communication purposes, next to 10 able-bodied subjects, they also recruited four disabled subjects, three with complete paraplegia and one with incomplete paraplegia to operate the BCI system. A matrix with white letters and a black background, similar to Figure 4.1 (right) was employed. Subjects were instructed to count the highlightings

Table 4.1: Required time in seconds to reach either 80% accuracy or 95% accuracy for different ISIs, classification techniques and subjects in Farwell and Donchin (1988). The best results for each subject are bold.

		80% Accuracy		95% Accuracy	
Method	**Subject**	**125ms ISI**	**500ms ISI**	**125ms ISI**	**500ms ISI**
Area	1	29.1	39.9	76.7	59.3
	2	49.0	56.6	-	-
	3	29.3	12.6	55.8	17.9
	4	45.5	44.9	82.2	52.9
Peak Picking	1	-	28.2	-	42.5
	2	-	**23.3**	-	**35.5**
	3	39.8	17.3	-	26.0
	4	38.8	**17.7**	70.4	**29.3**
SWDA	1	**15.7**	114.8	**21.6**	202.8
	2	33.4	56.9	57.5	-
	3	22.3	**11.1**	46.4	**17.6**
	4	54.4	26.7	-	49.5
Covariance	1	-	42.9	-	62.4
	2	-	-	-	-
	3	41.8	15.5	82.2	22.6
	4	36.7	28.6	64.0	52.0

Table 4.2: Average number of items per minute, achieved by disabled and able-bodied subjects considering a criterion of 80% accuracy and 95% accuracy, respectively (Donchin et al., 2000).

		80% Accuracy	95% Accuracy
Subject Group	**Preprocessing**	**Items/min**	**Items/min**
Able-Bodied	SWDA	6.3	3.4
	SWDA/DWT	7.8	4.3
Disabled	SWDA	4.8	2.8
	SWDA/DWT	5.9	3.2

of the letter “P”. Two classification strategies were employed for this study: SWDA as described in the previous section was calculated on epochs of 600ms, downsampled to 50Hz, yielding 30 data points for epoch classifications. Thereby, in this study, SWDA was applied on averages of the *cells* of the matrix, rather than to the rows and columns. Second, 640ms epochs were extracted and downsampled to 50Hz. Afterwards, *discrete wavelet transform* (DWT) using a Daubechies wavelet was applied and SWDA was applied to perform the classification task.

The classification results from ten able-bodied and the four disabled subjects were assessed and the average number of items per minute was calculated on the basis of 80% and 95% classification accuracy (see Table 4.2). For the best performing classification methods, 7.8 items/min and 4.3 items/min were achieved for the 80% and 90% accuracy level, respectively. These results correspond to 26.69 bits/min and 19.90 bits/min transfer rates. Disabled subjects achieved 5.9 items/min (20.19 bits/min) and 3.2 items/min (14.81 bits/min) for reaching the criteria.

In summary, Farwell and Donchin (1988) and Donchin et al. (2000) demonstrated that it is possible to employ the P300 component for predicting letters from a subject's EEG data, while the subject is focusing to letters within a flashing matrix. They

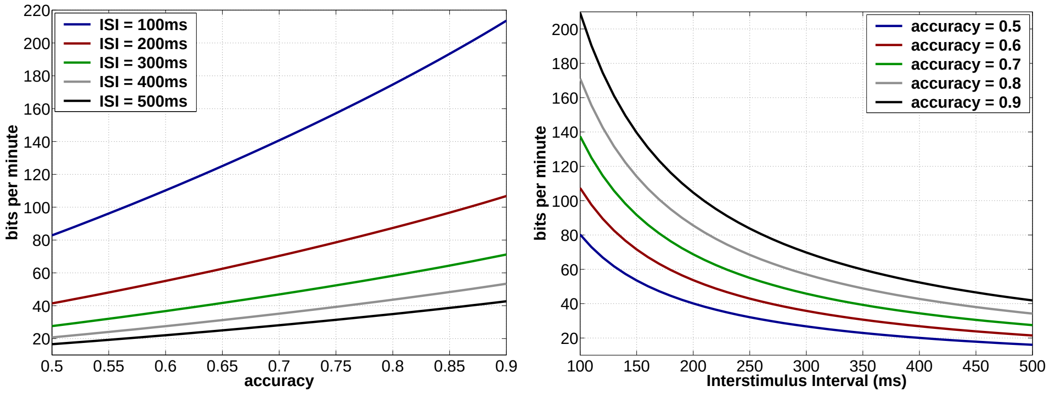


Figure 4.7: **Left:** Impact of different accuracies on the information transfer rate *bits per minute* for certain ISI durations. **Right:** Effect of the Interstimulus Interval on the information transfer rate for certain accuracies.

managed to achieve high information transfer rates without the necessity of training the subjects. However, the information transfer speed remains an important topic and is still below being satisfying. Thus, attempts to improve these rates are worth further considerations.

4.2 Information Transfer Rates in the P300 Speller Paradigm

Interdependencies between ISI, classification accuracies and the number of symbols within a matrix can be investigated by analyzing the information transfer rates as is detailed out in section 3.3. According to Wolpaw et al. (2002), such transfer rates can be calculated as

$$B(N,p,t) = \frac{60}{t}\left(\log_2 N + p\log_2 p + (1-p)\log_2 \frac{1-p}{N-1}\right), \tag{4.1}$$

with the number of choices N, the accuracy p and the time for a choice t. For the P300 speller paradigm with 6 rows and 6 columns, N is 36 and t can be rewritten as $12 \cdot t_{\text{ISI}}$ with t_{ISI} being the Interstimulus Interval, such that equation (4.1) turns into

$$B(36,p,t_{\text{ISI}}) = \frac{60}{(12\cdot t_{\text{ISI}})}\left(\log_2 36 + p\log_2 p + (1-p)\log_2 \frac{1-p}{35}\right). \tag{4.2}$$

Figure 4.7 depicts the relationships between accuracies and transfer rates for different ISIs. For a constant classification accuracy, a shorter ISI results in increased transfer rates. This gap in transfer rates increases with higher accuracies. With a long ISI, even high classification accuracies are unlikely to achieve competitive information transfer rates compared to low ISIs. For example, with an ISI of 500ms and an accuracy of 0.9, worse transfer rates would be achieved than with an ISI of 200ms and an accuracy of 0.6. Thus, ISI reduction is an important factor for speed improvements, and is likely to countervail lower classification accuracies caused by this reduction. Especially at low ISIs within the range of 100-200ms, small decreases in ISI result in strong enhances of

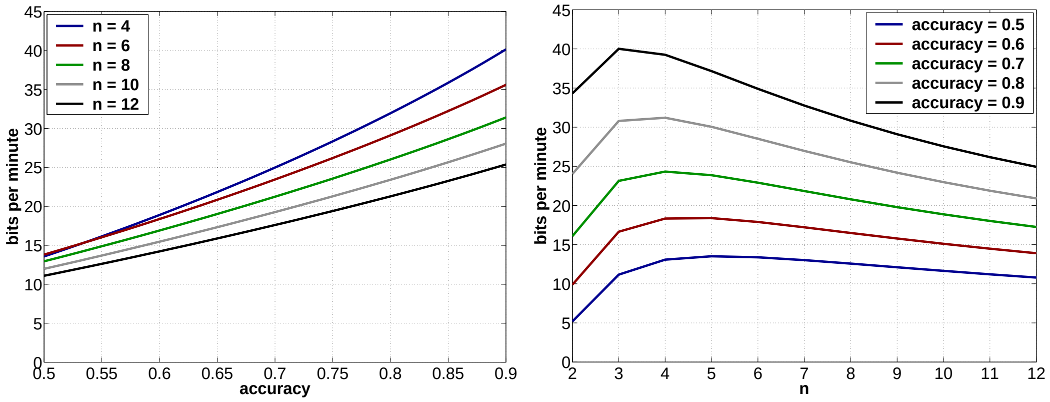


Figure 4.8: **Left:** Impact of different accuracies for certain n in a $n \times n$ stimulus matrix on the information transfer rate with a constant ISI of 500ms. **Right:** Impact of variations of n in an $n \times n$ stimulus matrix on the information transfer rate for certain accuracy levels with a constant ISI of 500ms.

information transfer rates (see Figure 4.7, right).

The basic P300 speller approach employs a 6×6 matrix but different expansions are possible: By increasing n in a $n \times n$ matrix, the number of symbols in the matrix raises quadratically. Therefore, more information could be transferred within one subtrial when using a larger number for n than just 6. For example, with $n = 8$, almost twice the symbols could be transferred than with $n = 6$, and the information transfer rate would improve from 5.17 bits/min to 6 bits/min for a single subtrial. On the other hand, with such a variation, the subtrial duration would also be enhanced: One trial lasts $t_{\text{trial}} = 2n \cdot t_{\text{ISI}}$ seconds, resulting in the following transfer rates:

$$B(n, p, t_{\text{ISI}}) = \frac{60}{2n \cdot t_{\text{ISI}}} \left(2 \log_2 n + p \log_2 p + (1-p) \log_2 \frac{1-p}{n^2 - 1} \right). \tag{4.3}$$

Allison and Pineda (2003) presented their subjects matrices of the three different expansions 4×4, 8×8, and 12×12, respectively. Digrams, i.e. a combination of two letters, served as symbols. The digrams in the matrix flashed for 100ms with a random delay of 450ms to 550ms between flashes.

An outcome of the study was that the matrix size did not affect N100 amplitude (see section 2.5) or latency. On the other hand, the P300 amplitude increased, and its latency decreased with the matrix size. Since enlarging the matrix results in a decreased probability for the target event, this finding is in line with findings from, e.g., Duncan-Johnson and Donchin (1977), stating that the P300 amplitude decreases with the probability of a target stimulus (cf. section 2.5).

Allison and Pineda (2003) did not perform classifications such that it remains unclear to which degree a classifier would benefit from amplitude increases for target letter stimuli in a larger matrix. Therefore, the question remains which of the antagonists has a stronger impact on the information transfer rate: While the duration of a trial would be prolongated with a larger matrix, a higher transfer rate could be achieved due to a higher amount of information and a higher target stimulus amplitude, presumably resulting in better classification accuracies.

4.3 Related Work

Beside Farwell and Donchin (1988) and Donchin et al. (2000), other researchers conducted experiments and developed further approaches to exploit the P300 component for Brain-Computer Interfacing.

Experiments with Paralyzed Subjects

It can a-priori not be taken for granted that the P300 speller paradigm works properly with paralyzed subjects, a group of subjects which could benefit to a large extent from a BCI. Therefore, Mellinger et al. (2004) employed seven patients with *Amyotrophic Lateral Sclerosis* (ALS) for experiments with the P300 speller device. It was also tested whether the device works well in the patient's home environment with several artifact sources among which are some typical for ALS patients. While the device worked well for one subject, yielding transfer rates of about 19.20 bits/min in offline analysis, results for the remaining six subjects stayed below 3 bits/min. Although this transfer speed is quite unsatisfying, it was shown that it is in principle possible to utilize the P300 speller paradigm for Brain-Computer Interfacing with ALS patients. The authors suspected that this performance could be enhanced with a different classification technique.

Adaptive P300 Speller Device

Serby et al. (2005) designed a P300 speller Brain-Computer Interface with a classification algorithm based on Independent Component Analysis and performed offline as well as online experiments with this setup. A special characteristic of this device is the circumstance that it adapts itself on the performance of the subject in the online mode. By continuously evaluating the classification results after each subtrial, it can be decided whether further stimulus presentations are necessary or if the presentation can terminate. Thereby, rows and columns are sequentially presented and evaluated in blocks, such that, e.g., 4 rows but 6 columns are presented within an trial to infer the correct symbol. With this approach, they achieved mean information transfer rates of 23.75 bits/min in the offline and 15.30 bits/min in the online version.

Single Display Speller Paradigm

An advantage of the flashing rows and columns in the P300 speller paradigm is that only 12 stimuli are needed to cover 36 symbols. On the other hand, 2 stimuli need to be classified correctly in order to infer the right symbol. If each symbol would be highlighted on its own, 36 flashes would be necessary for a complete subtrial, but on the other hand, only one stimulus would need to be classified accurately (unfortunately, 35 epochs would also need to be rejected correctly). However, Farwell and Donchin (1988) stated that *"successively choosing from among the 26 letters and communicating his choices via the P300 [...] would be unacceptably slow"*. Guan et al. (2004) investigated whether this assumption is true and turned away from flashing rows and columns. Instead, they highlighted just one symbol at once. Since the P300 amplitude raises due to the lowered probability of a relevant stimulus (1:36, instead of 1:6), it should be easier to identify a P300 from only a few trials. Furthermore, constraints on the experimental design which force a rectangular nature of the stimuli would not be necessary any more.

Six subjects attended in a study comparing the original setup of Farwell and Donchin (1988) (*FD speller*) with this *single display paradigm* (*SD speller*). The subjects were instructed to focus attention to specific symbols in the FD speller as well as in the

SD speller approach. The ISI for the former paradigm was set to 180ms, while it was reduced to 60ms in the latter approach. However, this setting results in comparable temporal relationships: A FD speller subtrial ($12 \cdot 180$ms), as well as a SD speller subtrial ($36 \cdot 60$ms) each required 2.16 seconds. A first outcome of the study was that the P300 amplitudes in the SD speller condition were enhanced in comparison to the FD speller condition (5.39μV vs. 4.04μV). Furthermore, 90% classification accuracy was reached after 15s in the FD speller, and after 8s in the SD speller condition. Thus, the SD speller paradigm yielded competitive transfer rates of up to 49.48 bits/min (cf. section 3.3). Further ISI reductions for the FD speller paradigm could have been possible, while smaller ISIs for the SD paradigm appear difficult, such that the speed of the FD speller paradigm could presumably have been improved. Nevertheless, the SD speller paradigm allows for more flexible interfaces with high transfer rates.

Employing the P300 for Interactions in Virtual Environments

Experiments on P300 elicitation commonly rely on static stimuli. In contrast, *virtual reality* (VR) expands the possibilities for researchers to dynamic environments by simultaneously preserving controllability.

Bayliss and Ballard (1999) recorded EEG activities while their subjects were controlling a VR driving simulator. The subjects were seated in a *go cart*, wearing a *head-mounted display* (HMD) and controlled a car in a virtual world. In this world, traffic lights were displaying yellow lights prior to green and red lights, making them more frequent than the red or green lights. The researchers investigated EEGs following yellow and red lights and found that in averaged ERPs, no P300 occurred for yellow lights while such a component could clearly be identified for red lights. Based on single trials, promising classification accuracies for identifying yellow and red lights in the range of 77% to 92% depending on the subject were found for the best performing classifier which was a robust Kalman filter. In another study, Bayliss (2003) employed the P300 to control items in a virtual apartment. By focusing attention to flashing items, they could provide commands for switching on/off a television, a stereo system, a lamp, saying “Hi”or saying “Bye”. Grand averages exposed clear goal responses in the ERP for the items, and 2.83 items could be selected in the VR environment per minute.

In summary, the researchers have demonstrated that wearing a HMD and simultaneously recording an EEG is possible. Furthermore, classification of P300 signals can be performed with reasonable classification accuracies in this context and can be used to provide steering signals in a dynamic environment.

Employing the P300 for the Detection of Deception

Farwell and Donchin (1991) suggested to employ the P300 component for the detection of deception (*lie detection*) in a *guilty knowledge test* (GKT) paradigm (Lykken, 1959). This paradigm compares answers of subjects to *relevant* and *neutral* questions. Subjects with knowledge of the crime should thereby expose different physiological reactions to relevant questions (Ben-Shakhar and Elaad, 2003). For P300-based detection of deception, the procedure would be as follows: First, within a series of *irrelevant* stimuli, some *target* stimuli occur, the subject is familiar with. The latter stimuli would induce a P300 component, because they are rare and relevant to the observer. A third kind of stimuli are *probes*, i.e., unique details of an event that are supposed to be only known by the suspect, if he was involved in the event. If the EEG patterns that result from the probes are similar to target stimuli, i.e., they expose a strong P300, the suspect has probably knowledge about the certain details.

According to Farwell, this so-called *brain fingerprinting* yield results with a confidence of 95%. Additionally, he augmented the P300 concept in this context to the *Memory and Encoding Related Multifaceted Electroencephalographic Response* (MERMER), which further includes subsequent electrical changes between 800ms and 1200ms. Employing MERMER, Farwell is convinced of gaining more than 99% accuracy (Farwell, 2006).

Due to the lack of sufficient independent peer-reviewed studies, several sceptical views about brain fingerprinting can be found (Wolpe et al., 2005; Rosenfeld, 2005). Nevertheless, these authors also claim that the P300 component can in principle be utilized for the detection of deception.

4.4 Summary

This chapter provided a review of Brain-Computer Interfaces based on the P300 component in the EEG signal. This component occurs for rare stimuli relevant to the observing subject. Farwell and Donchin (1988) exploited the P300 for constructing a BCI: They presented subjects a 6 × 6 matrix of 36 symbols and flashed its rows and columns. Subjects should select one symbol in the matrix and mentally count its flashing, making the rare (1:6) flashing at this position relevant, which resulted in a P300 component for such an event. Thus, by detecting the P300 component in the EEG signal recorded after a flashing, one row and one column could be determined, and their intersection in the matrix indicated the symbol the subject drawed attention to.

Farwell and Donchin (1988) employed different classification strategies and achieved up to 18.50 bits/min. In a later study, Donchin et al. (2000) improved the classification algorithms and conducted studies with able-bodied as well as with disabled subjects. Information transfer rates of maximal 26.69 bits/min could be achieved within this study.

The *information transfer rate* of a P300 speller BCI highly depends on the temporal distance of two events (Interstimulus Interval, ISI) and the classification accuracy. Thereby, decreasing the ISI can substantially improve the presentation speed. On the other hand, this modification results in overlaps between consecutive epochs and therewith in less pronounciated P300 signals, making good classifications less likely. Therefore, a trade-off between presentation speed and classification accuracy has to be found. Farwell and Donchin (1988) achieved reasonable results with 125ms ISI.

Since Farwell and Donchin (1988), several studies have been conducted employing the P300 speller BCI targeting to improve the performance and usability of the interface. Studies with paralyzed subjects, an adaptive speller device, and a modified stimulus presentation mode were presented. Furthermore, alternative utilizations of the P300 component in a BCI context have been developed as for interactions in virtual environments and for the detection of deception.

The main drawbacks of the P300 speller paradigm are that it depends on stimulations and does not provide a continuous signal. On the other hand, it allows for high information transfer rates and requires no training of the subjects.

Chapter 5

Data Analysis for the P300 Speller Brain-Computer Interface

A crucial part for achieving reasonable information transfer rates in Brain-Computer Interfacing is a suitable data analysis procedure. In order to relate specific brain activity to certain events like choosing a letter from a set of letters or operate a hand orthosis, the recorded data must be translated into such steering commands.

In general, the first step in BCI is to perform the **data acquisition** in order to receive the raw data. In the case of EEG signals, e.g. electrode positioning, impedance and shielding need to be considered to achieve a reasonable data quality. Then, **preprocessing** facilitates further analyses by e.g. reducing noise by filtering and computing adequate data representations, for example lower-dimensional representations, as feature vectors. Due to the low signal-to-noise ratio of EEG signals, this step is very important within this domain, especially when aiming to analyze single trials. In the case of the P300 speller paradigm, the subsequent **classification** step calculates from the feature vectors, representing brain activity patterns, whether a P300 occurred in a data sample or not (cf. section 4.1). Only after this binary classification procedure, a symbol within the stimulus matrix can be inferred by combining the binary classifications for the **symbol inference** step. Exactly one row and one column must be found which are most likely to contain the P300. From these certain rows and columns, a matrix entry can be identified. Figure 5.1 illustrates this whole data analysis procedure.

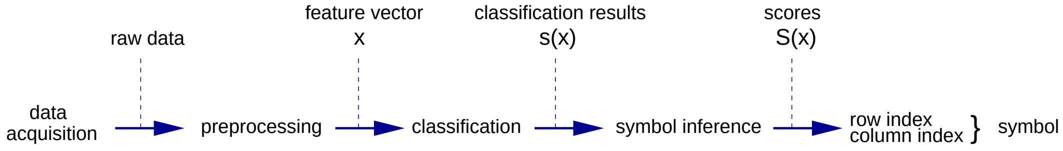


Figure 5.1: Data analysis chain for a P300-based BCI. After the raw data is acquired, it is preprocessed in order to facilitate further processing. Classifying the resulting feature vectors $\mathbf{x}$ yields the classification results $s(\mathbf{x})$ which indicate the row index i_r and column index i_c pointing on the symbol to be inferred.

In this chapter, each step of the data analysis procedure is described and methods as used in this thesis are discussed. After explaining important aspects of data acquisition in section 5.1, different preprocessing strategies like *Fourier transform* and *Principal Component Analysis* are discussed in section 5.2. Classification, as a crucial step, is discussed in more detail in section 5.3, including the theoretical backgrounds in terms of Statistical Learning Theory, classification strategies, and practical implications like cross-validation. Finally, classifications for symbol inference as it is performed in the context of the P300 speller paradigm is discussed in section 5.4.

5.1 Data Acquisition

In order to achieve a reasonable signal-to-noise ratio, several factors influencing the data quality have to be considered[1]. Foremost, it is important to have the subject grounded and to have low impedances between the electrodes. It is common to anticipate impedance values below 5kΩ, but within the context of Brain-Computer Interfacing and single trial analysis, even lower impedances are desirable. A key to achieve good impedances is to clean the skin from fat and oil with alcohol and abrade the skin using a skin preparation gel.

It is advisable to have the experimental room sound attenuated to prevent distractions of the subject. Furthermore, the room should be shielded against electromagnetic influences by using a Faraday cage. In order to receive good signals, the subject should be relaxed to prevent muscle activity, but not be drowsy, which would result in high α wave activity (see section 2.4) which might pollute the signal. As introduced in section 2.3, eyeblinks and movements should be prohibited by instruction and the experimental design, by e.g. introducing pauses for recreation. Eyeblinks can be identified by recording the *Electrooculogram* (EOG) which can further be used for artifact elimination techniques (see section 2.3).

Different electrode locations are approriate for different kinds of BCI experiments. While motor imagination primarily results in motor cortex activity, which can be assessed by electrodes over motor cortex areas at e.g. $C3$ and $C4$, a P300 is commonly most pronounciated at parietal locations like Pz (see Figure 2.4). On the other hand, applying electrodes on a wide range of locations yield more information and allow for more differentiated analyses like e.g. source localization (Lantz et al., 2003). Although it is common to employ the international 10-20 System by Jasper (1958) (cf. section 2.3), it can be useful to work with deviating locations, e.g., $C3'$ and $C4'$, which are 1cm frontal to $C3$ and $C4$, respectively, for investigating motor processes.

EEG amplifiers are *differential amplifiers*, which means that *pairs* of electrodes are amplified. While one electrode provides a reference signal, the potential of another electrode is measured with respect to this reference. In general, the three different kinds *common reference*, *average reference* and *bipolar derivations* can be distinguished (Ebe and Homma, 1994). In common reference, one or two electrodes, e.g. located at the ear lobes, provide the reference potential, and all other electrodes (commonly the scalp electrodes) are measured with respect to this potential. In average reference, the average of all electrodes provides the reference. In bipolar derivations, explicit pairs of electrodes are amplified. The *lateralized readiness potential* (LRP), measuring the potential difference of $C3$ and $C4$ is an example for the latter derivation (Coles, 1989).

Common materials for electrodes are silver/silver chloride (Ag/AgCl), silver (Ag), tin (Sn) and gold (Au). Accurate signals can be obtained with Ag/AgCl electrodes, particularly, when using sintered electrodes, while the other electrodes expose high pass filtering characteristics (Picton et al., 1995).

[1] More details can be obtained in section 2.3 and from Zschocke (1995).

Electrodes can be applicated as single electrodes, with an electrode cap, or with a *Geodesic Sensor Net* (GSN) (Tucker, 1993). Applying single electrodes requires to determine the locations using e.g., straps, while an electrode cap and the GSN incorporate the locations, and are therefore particularly useful for quick applications. On the other hand, the latter approaches use mechanical pressure to hold the electrodes in position which can be painful for the subjects, especially in long sessions.

5.2 Preprocessing

Preprocessing aims to produce appropriate feature vectors for the subsequent steps of analysis. In principle, two goals can be distinguished for this step. First, *noise reduction* aims to eliminate signals which do not correlate with the signal the researcher is intending to assess. For this purpose, **filters** are especially important within the EEG domain and are described in section 5.2.1. Second, another goal is to find an adequate *data representation* (Bishop, 1995). Data in the input space mostly incorporate high redundancies due to correlations within the data and unimportant information for the subsequent steps of analysis. Reducing the dimensionality of the data space without substantial loss of information is therefore usually possible. It is furthermore even desirable for two reasons. First, lower-dimensional data is easier to process due to decreased memory and computational demands[2]. In some cases, it is even inevitable to have low-dimensional data matrices in order to be able to perform e.g. matrix inversion[3], which is crucial for the upcoming Fisher's Linear Discriminant Analysis, an important algorithm for this thesis (see section 5.3.3). Second, it is in general advisable to follow the *intrinsic dimensionality* of the data. If the data lies entirely within a d-dimensional subspace of the input space, it is said to have the intrinsic dimensionality d (Fukunaga, 1982). For classification purposes, it is recommended to work with low dimensionalities due to the *curse of dimensionality* which says that the data representation becomes very sparse, when the data dimensionality raises, providing a poor data representation (Bishop, 1995). A powerful method to achieve a dimensionality-reduction is **Principal Component Analysis**, which is described in section 5.2.2.

Artifact elimination can also be a preprocessing step. In a simple approach, trials with eye movements, as assessed by EOG activity (see section 2.3) exceeding a specific threshold, e.g., 100μV are disbanded (Zschocke, 1995). Other techniques subtract weighted EOG activity from the scalp electrode signal (Gratton and Coles, 1989) or employ *Independent Component Analysis* for artifact elimination by discarding those independent components for subsequent analyses which are likely to reflect artifact activities (Jung et al., 2000). Although artifact elimination could reduce noise in the data, the present work does not incorporate explicit artifact reduction methods beside filtering for frequencies. Rather, the classifiers in this thesis should learn from the data to deal with artifacts by themselves. Nevertheless, artifact elimination could possibly further improve the classification results.

[2] For instance, matrix inversion with Gaussian Elimination leads to $\mathcal{O}(n^3)$ complexity for inverting a $n \times n$ matrix.

[3] Matrix inversion becomes improbable if the number of features (i.e., the data dimensionality) is larger than the number of data examples (Guyon et al., 2002). In such cases, the matrix easily becomes singular.

5.2.1 Filtering

Filtering in the context of EEG-based BCI denotes filtering for frequencies for the purpose of *signal separation* (Smith, 1999). Thus, filtering allows to selectively attenuate specific frequencies within the EEG signal and disband other frequencies. First, the two different kinds *low pass* and *high pass* filtering need to be considered. While the former one disbands frequencies above a specific *cut-off* frequency which results in a smoothing in the time domain, the latter one eliminates frequencies below such a cut-off frequency, and therefore works as a derivator which attenuates slowly varying components in the time domain. In particular for the examination of single trials, the importance of filtering has been identified: *"Filtering can also make possible the examination of ERPs on a single trial basis when the signal magnitude is sufficiently large with respect to the noise and/or the signal and noise spectra are sufficiently disparate. This has been found to be feasible when dealing with high amplitude, low frequency components such as the CNV, N200, P300 and slow wave."* (Picton et al., 1995).

In the context of BCI, it is useful to employ a **high pass filter** to prevent occurrences of low frequencies to avoid drifts caused by sweating artifacts and slow drifts of the electrode potentials (Zschocke, 1995). Such drifts would cause exceedings of the voltage ranges of the EEG amplifier (oversteering). On the other hand, signals which could be interesting, like potential drifts in the cortex surface (Caspers and Speckmann, 1974), might get lost using this technique. A **low pass filter** can be useful to filter out the 50Hz or 60Hz power supply frequency and muscular artifacts. Furthermore, most of the energy of low frequency ERPs like CNV, P300 and slow wave is concentrated below 20Hz (Picton et al., 1995). It is nevertheless desirable to sample data at a rate of at least twice the highest frequency of interest in the data. Sampling below this *Nyquist rate* might result in aliasing errors[4]. Since both high pass and low pass filters are *linear* filters, they can easily be combined to form a **band pass filter** by consecutive execution of the filters in arbitrary order.

Filtering can be achieved in an analog or digital fashion. While analog filtering is commonly performed in the EEG amplifier itself, digital filtering is computed in a subsequent data analysis step by a computer algorithm. Employing the *Fourier transform* (FT) is one way to perform digital filtering (Smith, 1999). The idea behind FT is that any periodic function can be decomposed into sinusoidal waves of different frequencies and phase relationships. Windows of time series data can then be transformed to the frequency domain. By eliminating certain frequencies in that domain and performing a backtransformation, a frequency filter can be realized. In a discrete formulation, the sequence $x_0, ..., x_{n-1}$ of n complex numbers is transformed by *discrete Fourier transform* (DFT) into a sequence of n complex numbers $f_0, ..., f_{n-1}$ by

$$f_j = \sum_{k=0}^{n-1} x_k \, e^{-\frac{2\pi i}{n} jk} \quad j = 0, ..., n-1 \tag{5.1}$$

where i is the imaginary unit $\sqrt{-1}$. While this transform requires computational costs of $\mathcal{O}(n^2)$, it can recursively be broken down to smaller DFTs in a *divide-and-conquer* strategy, resulting in the so-called *Fast Fourier transform* (FFT), which yields costs of

[4] Aliasing is an artifact that occurs when analog signals are digitized at an inadequate frequency, such that high frequency components can be improperly reflected as low frequency components.

$\mathcal{O}(n \log n)$ (Cooley and Tukey, 1965). FFT requires the data to be of a length of a power of 2, but this can easily be achieved by expanding the data vector to this size by padding the data with zeros (Press et al., 1992).

A filter can be constructed by employing FFT for transformations into the frequency domain. In the easiest case, the *ideal* or *brick wall* filter (Smith, 1999), coefficients reflecting frequencies to be disbanded are set to zero, and the data are transformed back to the time domain by

$$x_k = \sum_{j=0}^{n-1} f_j \, e^{-\frac{2\pi i}{n} jk} \quad k = 0, ..., n-1. \tag{5.2}$$

Note that this kind of filter can result in a damped *ringing* at the edges, i.e., frequencies outside the pass band do not vanish completely but have a residual impact on the signal due to the fact that the length of the window in the time domain is not infinite and has therefore no infinite frequency response. A common way to reduce ringing is to use cosine-smoothed *Hanning windows* instead of rectangular windows (Smith, 1999; Press et al., 1992).

5.2.2 Dimensionality Reduction by Principal Component Analysis

Principal Component Analysis (PCA) computes an alternative data representation by mapping the original data along uncorrelated dimensions that reflect the main variances of this data[5] (Bishop, 1995; Hyvärinen, 1999). Since the new basis system is hierarchically organized with respect to the degree of variance they capture, such that the first Principal Component also accounts for the most variance, PCA can easily be employed for dimensionality reduction by considering only the first l components for a data representation. For a PCA algorithm, the goal is to find a transformation matrix $\mathbf{W}$ mapping a vector $\mathbf{x}$ to an alternative representation

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{5.3}$$

Thereby, the direction of the first Principal Component $\mathbf{w}_1$, which is the first column of the matrix $\mathbf{W}$, explains the most variance in the original space and obeys

$$\mathbf{w}_1 = \arg \max_{||\mathbf{w}||=1} < (\mathbf{w}^T \mathbf{x})^2 > \tag{5.4}$$

where $< \cdot >$ is the expectation value. Any subsequent direction $\mathbf{w}_k$ ($k > 1$) can recursively be computed by subtracting the first $k-1$ Principal Components from $\mathbf{x}$ by

$$\hat{\mathbf{x}}_{k-1} = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x} \tag{5.5}$$

and calculating the direction of this component as indicated in (5.4):

$$\mathbf{w}_k = \arg \max_{||\mathbf{w}||=1} < (\mathbf{w}^T \hat{\mathbf{x}}_{k-1})^2 > . \tag{5.6}$$

[5] PCA is equivalent to the Karhunen-Loève transform and the Hotelling transform (Karhunen, 1947; Hotelling, 1933).

In practice, the simplest way to perform PCA is to employ *Singular Value Decomposition* (SVD) to calculate the matrix $\mathbf{W}$ (Ripley, 1999; Golub et al., 1999). Given a dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$, representing n data vectors $\mathbf{x}_i \in \mathbb{R}^p$ of zero mean each with n exceeding the dimensionality ($n \geq p$), SVD is calculated by

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{5.7}$$

with $\mathbf{\Lambda}$ being a diagonal matrix of decreasing non-negative entries. The matrix $\mathbf{U} \in \mathbb{R}^{n \times p}$ contains orthonormal columns, and $\mathbf{V} \in \mathbb{R}^{p \times p}$ is orthogonal. Then, the columns of

$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Lambda} \tag{5.8}$$

reflect the Principal Components and is therefore the desired matrix

$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}. \tag{5.9}$$

$\mathbf{\Lambda}$ contains the Eigenvalues λ_k on its diagonal. The Eigenvalues correspond to the variance they capture in the original space. Therefore, in order to achieve a data compression by using PCA which captures a certain amount of variance, the first l vectors can be employed which obey

$$\sum_{i=1}^{l} \frac{\lambda_i}{Tr(\mathbf{\Lambda})} \geq q \tag{5.10}$$

with q being the fraction of variance that should be captured, e.g. q=0.9 for capturing 90% variance. It is often sufficient to employ only a few number of Principal Components to capture the vast amount of variance, but the PCA components do a-priori not necessarily reflect useful information, since PCA *solely* relies on variance which must not be the most relevant criterion. Accordingly, although certain Principal Components might contribute only minor to the data variance, they can nevertheless contain valuable information. Additionally, PCA is only a *linear* transformation. Thus, if the data follows a non-linear distribution by e.g. being projected on a circle, the Principal Components will not reflect the intrinsic trend of the data (Marques de Sa, 2001). Depending on the data domain, Principal Components reflect certain aspects of the analyzed information. For instance, specific characteristics of faces, so-called *Eigenfaces* are reflected in these components when analyzing face databases (Turk and Pentland, 1991). Other studies found Gabor-like filters for the Principal Components when analyzing natural images (Hancock et al., 1992; Heidemann, 2006).

Some approaches try to overcome the restrictions of PCA. For example, Independent Component Analysis tries to relax the orthogonality constraints of PCA and instead of maximizing variance, it maximizes a measure of information, e.g. *entropy* (Hyvärinen, 1999).

5.3 Classification

Within the P300 speller paradigm, binary classification needs to be performed in order to determine whether a certain time series of EEG data (an epoch), contains a P300 or not. More formally, a classifier must categorize whether a time series $\mathbf{x}$ belongs to the class $\mathcal{P}^+$ (contains P300) or $\mathcal{P}^-$ (does not contain a P300). A strategy to perform this assignment can be derived in different ways. First, it is possible to employ *Model-Based* techniques. Those techniques rely on knowledge or assumptions about the underlying processes or signal characteristics. For instance, the classifiers *area* and *peak picking*, as discussed in detail in section 4.1.1, use knowledge about the characteristics of the P300 to perform the classification. An alternative approach are *Machine-Learning* classification strategies. Such techniques learn from given data to distinguish between examples from the different classes. The appealing fact about Machine-Learning techniques is that almost no prior knowledge about the data domain is necessary - the algorithms extract relevant information for the classification problem by themselves. A weakness of this approach is that although it is possible to analyze to some degree what the classifiers learnt (Golland, 2002), it is usually not possible to extract explicit human-readable rules, such that it is not clear in most cases, what the classifier really learnt. Thus, the usage of such methods is also accompanied with a loss of control, compared to rule-based systems, making their applications in e.g. critical medical contexts difficult. Another serious drawback of these classifiers is that they first need to be trained on a data set of training examples in order to be able to perform classification.

Statistical Learning Theory is a basis for Machine-Learning classification and is explained in section 5.3.1. More details about this topic can be obtained by e.g. Evgeniou et al. (2000), Hastie et al. (2001), and Vapnik (1995). This theory leads to the powerful technique *Support Vector Machines* which gained much attention in recent years due to its high generalization capabilities and good classification performance. This technique will be illustrated in more detail in section 5.3.2. For further information the reader is referred to Schoelkopf and Smola (2002), Cristianini and Shawe-Taylor (2000), and Burges (1998). Another, more simple Machine-Learning classifier is *Fisher's Linear Discriminant Analysis* and will be discussed in section 5.3.3 (Bishop, 1995). When evaluating classification performances, *cross-validation* is an important instrument to assess the generalization performance of a trained classifier as will be detailed out in section 5.3.4.

5.3.1 Statistical Learning Theory

The goal of Machine-Learning classification is to find a relationship between two (or more) variables derived from a set of sample data, the so-called *training data*. More formally, an algorithm seeks to find a function f from a set of functions $\mathcal{F}$ describing the relationships between $\mathbf{x} \in X \subseteq \mathbb{R}^d$ and $y \in Y \subseteq \mathbb{R}$. In real world data, an element of $\mathbf{x}$ does not correspond to one specific element of Y in a deterministic manner. In fact, the relationships between the elements of X and Y are commonly *probabilistic* and the underlying probability distribution is a-priori unknown. Thus, the objective of *Machine-Learning* is to find an estimator $f : X \rightarrow Y$ that predicts a value y from any vector $\mathbf{x}$. If $y \in \mathbb{R}$, this problem is called *regression*. In contrast, if y takes a limited number of discrete values, *classification* is performed with the special case of *binary classification*, when e.g. $y \in \{-1, 1\}$. Binary classification is the essential kind of classification for this thesis, targeting to distinguish between P300 and non-P300 examples, reflected by the classes $\mathcal{P}^+$ and $\mathcal{P}^-$, respectively.

One way to find a good estimator is to find a function $f \in \mathcal{F}$ which produces the least number of errors among the functions in $\mathcal{F}$. The first step in this strategy is to define errors in terms of a loss function $V(y, f(\mathbf{x}))$. Typical loss functions for regression (V_r) and classification (V_c) are

$$V_r(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2, \tag{5.11}$$

$$V_c(y, f(\mathbf{x})) = 1 - \delta_{y,f(\mathbf{x})}, \tag{5.12}$$

with $\delta_{y,f(\mathbf{x})}$ being Kronecker's delta[6]. Together with the probability distribution $P(\mathbf{x}, y)$, the average loss or *Expected Risk* can then be calculated as

$$R(f) = \int_{X,Y} V(y, f(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}dy \tag{5.13}$$

and the optimal function f_0 can be found by minimizing this risk:

$$f_0(\mathbf{x}) = \arg\min_{f \in \mathcal{F}} R(f). \tag{5.14}$$

Empirical Risk Minimization

Unfortunately, only a fraction of all possible examples of $\mathbf{x}$ and y can usually be accessed, such that the Expected Risk can seldom be calculated in practice. Instead, another mease, the *Empirical Risk* is calculated: Only a restricted set $\mathcal{D}_l$ of l samples constituting the training set is drawn from the unknown probability distribution $P(\mathbf{x}, y)$ in order to find f:

$$\mathcal{D}_l \equiv \{(\mathbf{x_i}, y_i) \in X \times Y\}_{i=1}^{l}. \tag{5.15}$$

Since the probability distribution $P(\mathbf{x}, y)$ is unknown and only the incomplete sample set $\mathcal{D}_l$ is available, it is improbable to find f_0. Instead, the Empirical Risk R_{emp} can be estimated on the data set $\mathcal{D}_l$ as an approximation of the Expected Risk:

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l} V(y_i, f(\mathbf{x_i})). \tag{5.16}$$

A target function $\hat{f}$ can then be found by *Empirical Risk Minimization* (ERM):

$$\hat{f}(\mathbf{x}) = \arg\min_{f \in \mathcal{F}} R_{emp}(f). \tag{5.17}$$

By the law of large numbers, the Empirical Risk converges towards the Expected Risk for a large number of samples (Evgeniou et al., 2000):

$$\lim_{l \to \infty} R_{emp}(f) = R(f). \tag{5.18}$$

Overfitting and Complexity

With a large set of functions $\mathcal{F}$ it is also likely to find a large number of functions, which perfectly learn the training data, especially in the case of classification or with only few samples. In fact, most of these functions would not reflect the trend of the

[6] Kronecker's delta is defined as $\delta_{i,j} = 1 \Leftrightarrow i = j$ and $\delta_{i,j} = 0 \Leftrightarrow i \neq j$.

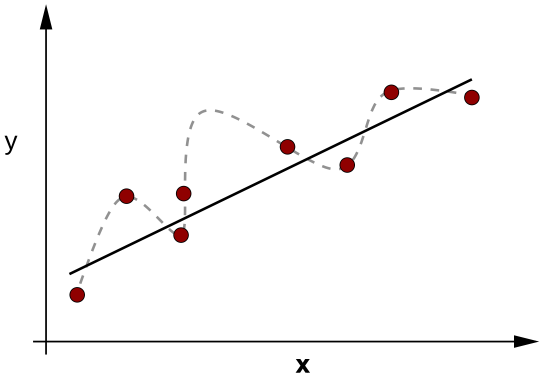


Figure 5.2: Two regression functions try to approximate the distribution of the data vectors as represented by the red circles. While the dashed gray function fits very close to the specific data vectors, the solid black function follows a more general trend of the data.

data appropriately and would therefore not work well for yet unseen data from outside the training data. This *overfitting* phenomenon is illustrated in Figure 5.2. In this example, for the sake of a very close approximation of the data, the regularity behind the data is not identified by the dashed function and a function of high complexity is postulated. It is unlikely that this function works well for unseen samples from the underlying probability distribution. In contrast, the solid line within the figure follows a more general trend of the data which appears to reflect the trend of the data more appropriately. One way to deal with overfitting is to follow the philosophy of Occam's razor[7] and the *principle of parsimony* and prefer the simpler theory (Domingos, 1999). Thus, although both functions in Figure 5.2 approximate the data distribution well, the less complex function, i.e., the solid black function, should be preferred. This principle can be formalized by e.g. penalizing the complexity of a function and/or restricting the set of functions. With only a small set of functions $\mathcal{F}$, uniform convergence of the Empirical Risk to the Expected Risk can be achieved (Evgeniou et al., 2000):

$$\lim_{l \to \infty} P\{\sup_{f \in F}(R(f) - R_{emp}(f)) > \epsilon\} = 0 \quad \forall \epsilon > 0. \tag{5.19}$$

One-sided uniform convergence is a necessary and sufficient condition for the ERM to be consistent. Thus, according to ERM, $\hat{f}_l$ can be found in this case. Thus, in general, a restricted set of functions is necessary. Choosing such a function class $\mathcal{F}$ is called *model selection* (Hastie et al., 2001) and the upcoming Vapnik-Chervonenkis dimensionality provides a measure for finding an appropriate function class of suitable complexity.

Vapnik-Chervonenkis Dimensionality

A precondition for becoming able to penalize complexity is a measure for complexity. The Vapnik-Chervonenkis (VC) dimension h provides such a measure as it reflects how many samples[8] can be separated by a function f from the set $\mathcal{F}$. For example, consider $\mathcal{F}$ being the set of hyperplanes in $\mathbb{R}^2$. Then, the maximum number h of binary labeled

[7] Commonly phrased as *"entities are not to be multiplied beyond necessity"*.

[8] In order to keep the considerations simple, only binary classifications will be considered in the following.

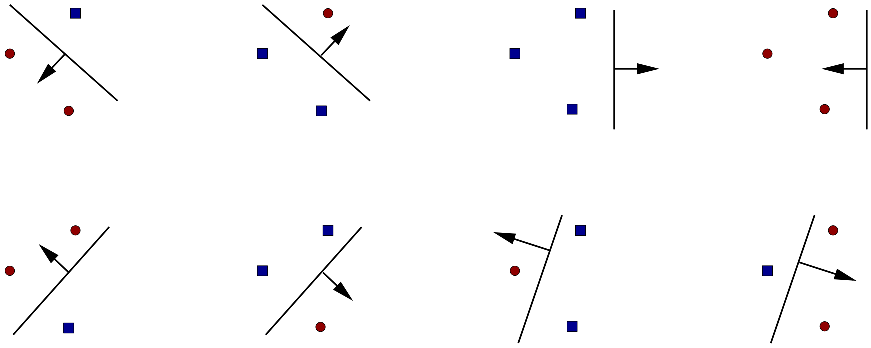


Figure 5.3: The Vapnik-Chervonenkis dimension h reflects the number of data points that a function class $\mathcal{F}$ can separate in all possible shatterings. In this example of hyperplanes in $\mathbb{R}^2$, h is 3, and the number of ways to divide the data into different classes is $2^h = 8$.

samples which can be separated is 3, resulting in $2^h = 8$ possible shatterings[9] (see Figure 5.3).

If $\mathcal{F}$ is a set of functions with VC dimension h, for any l examples $\{(\mathbf{x}_i, y_i)\}$ i.i.d. sampled from the distribution $P(\mathbf{x}, y)$, the following inequality holds with the probability $1 - \eta$ (Burges, 1998):

$$R(f) \leqslant R_{emp}(f) + \sqrt{\frac{h(\log(\frac{2l}{h} + 1)) - \log(\frac{\eta}{4})}{l}}. \tag{5.20}$$

Thus, the higher h, the more samples can be separated, and the higher is the complexity or *capacity* of a function set $\mathcal{F}$. For small values of h, the capacity term $\sqrt{\frac{h(\log(\frac{2l}{h}+1))-\log(\frac{\eta}{4})}{l}}$ in equation (5.20) is low, also resulting in a low upper bound on $R(f)$.

Taken together, in order to achieve good generalization to unseen samples, instead of solely minimizing the Empirical Risk, also the complexity of the hypothesis space $\mathcal{F}$ should be minimized (see also Figure 5.4). These requirements are formalized in the principle of *Structural Risk Minimization* (SRM).

Structural Risk Minimization

Structural Risk Minimization augments Empirical Risk Minimization by seeking to find the function f with the lowest Empirical Risk *and* the lowest complexity in terms of VC dimensionality as well. For this purpose, SRM defines a nested set of function sets $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_{n(l)}$ with $n(l)$ being a non-decreasing integer function of l where each hypothesis space $\mathcal{F}_i$ has a finite VC dimension which is larger or equal than that of all previous sets: $h_1 \leqslant h_2 \leqslant ... \leqslant h_{n(l)}$. SRM then yields the function class $\mathcal{F}_i$ such that the upper bound of the generalization error (5.20) is minimized. The abilities of a classifier to generalize to new, unseen data can be assessed by performing e.g. cross-validation as is explained in section 5.3.4.

[9] For instance, four samples would lead to the XOR-problem: There is no way to separate data points by a hyperplane in four corners of a square, when opposing corners belong to the same class.

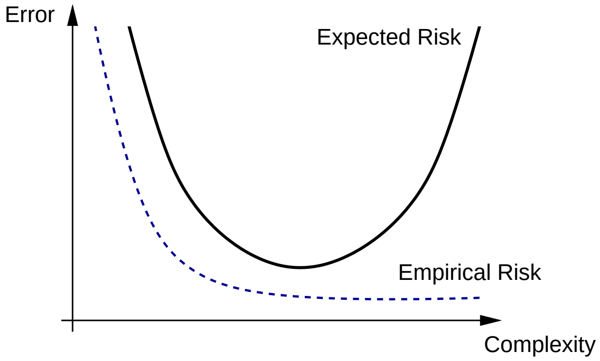


Figure 5.4: Solely minimizing the Empirical Risk by increasing the model complexity results in overfitting. In this case, the error on unseen data, as reflected by the Expected Risk, will raise after a certain degree of complexity. Thus, a trade-off between empirical error and complexity has to be found in order to achieve a good generalization to unseen data.

5.3.2 Support Vector Machines

Support Vector Machines (SVMs) employ the theory of SRM in order to create a powerful classifier with high generalization to unseen samples. SVMs were successfully applied in a large range of applications such as face detection, cancer detection, and handwritten digit recognition (Bennett and Campbell, 2000; Schoelkopf and Smola, 2002; Guyon, 2006).

Geometrically, SVMs employ hyperplanes which can be described as $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ in order to perform binary classifications. Projecting a sample on $\mathbf{w}$ reveals the class label by the function's sign:

$$s = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b). \tag{5.21}$$

While several hyperplanes could correctly divide the data space into two regions according to the specific classes, not every hyperplane is robust against variations of the data vectors. Therefore, generalization to new, unseen data can be low. If the dotted black hyperplane in Figure 5.5 separates the red circled data from the blue boxed data, small deviations from the actual positions of the data could easily result in misclassifications. On the other hand, employing the solid black hyperplane in the figure for separation purposes, a more robust classifier with respect to deviations to the actual data can be achieved. SVMs construct parallel hyperplanes to the separating hyperplane with the nearest samples from the different classes lying on the parallel hyperplanes. In order to find the optimal hyperplane with the best generalization capabilities, SVMs maximize the distance γ, the so-called *margin*, between the parallel hyperplanes (see Figure 5.5).

This fulfills the requirements of SRM since the capacity of the function class decreases when the margin increases: Let r_{Sphere} be the radius of the smallest sphere containing all data vectors and γ be the distance of the margin hyperplanes (the parallels to the optimal hyperplane) to each other as depicted in Figure 5.5. Then, the following inequality holds (Vapnik, 1995):

$$h \leq \frac{r_{\text{Sphere}}^2}{\gamma^2}. \tag{5.22}$$

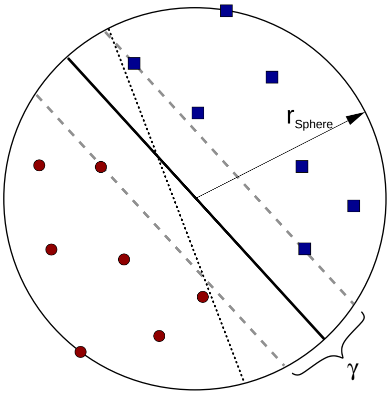


Figure 5.5: Data samples belonging to two different classes, separated by hyperplanes (dotted and solid black). The radius r_{Sphere} describes the smallest sphere containing all data. While the dotted black separating hyperplane would easily result in misclassifications when the data slightly vary from the actual positions, the solid black plane would be more robust against such deviations. The latter plane can be derived when constructing parallels to a separating hyperplane (gray, dashed lines) which touch the nearest samples to the hyperplane. That hyperplane which parallels expose the largest distance to each other is the optimal hyperplane with the best generalization capabilities (Burges, 1998).

According to SRM (and Occam), the function with a smaller value h should be preferred. The separating hyperplane can be described by the vectors on the margin hyperplanes. These *Support Vectors* are the nearest vectors to the hyperplane and are the only relevant vectors for the training problem. If the classifier would be trained again under exclusion of all other vectors, the Support Vectors would still be the same.

Linear Support Vector Machines

As already mentioned, a hyperplane separating data into two classes can be described by $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. If the data are linear separable, this plane can be found by maximizing the margin γ (see Figure 5.6) as follows: If $\mathbf{w}$ is normalized, such that

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \text{ for } y_i &= +1, \\ \mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \text{ for } y_i &= -1 \end{aligned}$$

holds, this expression can be summarized to

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \;\forall i. \tag{5.23}$$

A solution in two dimensions is illustrated in Figure 5.6. Data vectors on the parallels H_1 and H_2 of the optimal hyperplane are called *Support Vectors* (bordered items within the figure) and can be described by $\mathbf{x}_i \cdot \mathbf{w} + b = 1$ and $\mathbf{x}_i \cdot \mathbf{w} + b = -1$, resulting in

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 = 0 \;\; \forall\, (\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{SV}} \tag{5.24}$$

where $\mathcal{D}_{\text{SV}}$ is the space of the Support Vectors. The distance to the origin of a datapoint on H_1 is $\frac{1-b}{\|\mathbf{w}\|}$ and on H_2 is $\frac{-1-b}{\|\mathbf{w}\|}$, respectively, resulting in $\gamma = H_1 - H_2 = \frac{2}{\|\mathbf{w}\|}$. Another way of maximizing γ is therefore to minimize $\| \mathbf{w} \|$. Instead of using $\| \mathbf{w} \|$, it is more

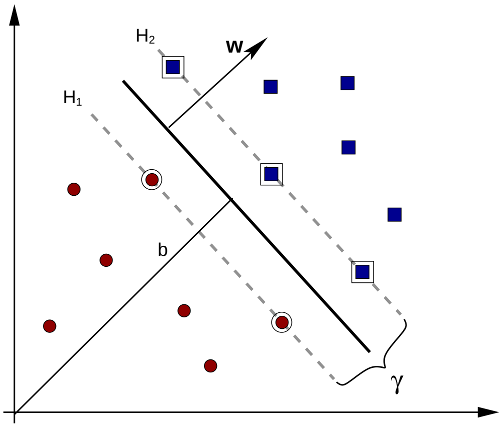


Figure 5.6: Support Vector Machines find the optimal hyperplane (solid line) to separate two classes by maximizing the margin γ. This hyperplane can be described by the vector $\mathbf{w}$ and the bias term b.

convenient to use the objective function

$$\frac{1}{2} \parallel \mathbf{w} \parallel^2 \tag{5.25}$$

for subsequent steps. In any case, equation (5.23) must still hold. In terms of optimization, a quadratic function must be minimized with respect to linear constraints, resulting in a convex function. A useful property of convex functions is that no local minima exist and any local solution is also the global solution. A convenient way to solve such convex functions is the use of Lagrange theory (Bishop, 1995). The *primal Lagrangian* of the problem is

$$L_P = \frac{1}{2} \parallel \mathbf{w} \parallel^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i. \tag{5.26}$$

This term can be minimized by setting the first differential to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \tag{5.27}$$

$$\Leftrightarrow \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \tag{5.28}$$

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{5.29}$$

Resubstituting these expressions in the primal Lagrangian (5.26) yields the *dual Lagrangian*

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j. \tag{5.30}$$

In this case, training a Support Vector Machine is equivalent to maximizing L_D by varying the Lagrange Multipliers α_i. The hyperplane $\mathbf{w}$ can then be computed using equation (5.28). The data vectors with $\alpha_i > 0$ build the Support Vectors, which are vectors on H_1 and H_2. For any other training vectors α_i will be zero. The optimal solution of the dual Lagrangian can be achieved using the *Karush-Kuhn-Tucker conditions*, which are necessary and sufficient for convex problems (Fletcher, 1987):

$$\begin{aligned} \frac{\partial}{\partial w_v} L_P = w_v - \sum_{i=1}^{l} \alpha_i y_i x_{iv} &= 0 & (5.31) \\ \frac{\partial}{\partial b} L_P = -\sum_{i=1}^{l} \alpha_i y_i &= 0 & (5.32) \\ y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 &\geq 0 & (5.33) \\ \alpha_i &\geq 0 & (5.34) \\ \alpha_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1) &= 0. & (5.35) \end{aligned}$$

Finding the solutions for these equations, and therewith computing values for b and all α_i is equivalent to solving the SVM problem. According to (5.24), Support Vectors fulfill $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 = 0$. As the *complementary condition* (5.35) reveals, α_i can then not be zero for these cases. On the other hand, α_i must be zero for those vectors with $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \neq 0$, i.e., vectors which are no Support Vectors. Thus, when calculating the weight vector $\mathbf{w}$ via (5.28), all vectors can be disregarded which are no Support Vectors since they have no influence on $\mathbf{w}$ due to their α-value of zero. Reversely, the Support Vectors contain the whole information about the optimal hyperplane. While $\mathbf{w}$ can be computed by (5.28) and solving the Karush-Kuhn-Tucker conditions, b needs to be calculated explicitly as the mean distance of the parallels H_1 and H_2 to the origin:

$$b = \frac{\max_{x_i : y_i = -1}(\mathbf{w} \cdot \mathbf{x}_i) + \min_{x_i : y_i = +1}(\mathbf{w} \cdot \mathbf{x}_i)}{2}. \qquad (5.36)$$

Classification by a trained machine is performed by determining on which side of the hyperplane a given test sample $\mathbf{x}_{\text{test}}$ lies. The class label is then calculated analogously to (5.21) by $\text{sgn}(\mathbf{w} \cdot \mathbf{x}_{\text{test}} + b)$.

Soft Margin Approach

So far, it was assumed that a linear separation of the data into two classes is possible. In real-world data, this case is very seldom. The assumptions about separability therefore need to be softened and therewith turned from a *hard margin* to a *soft margin* approach. Introducing *slack variables* allows for a solution of non-separable problems. As depicted in Figure 5.7, the condition (5.23) can be softened by introducing the slack variables ξ such that not all data vectors need to be on the “correct”side of the hyperplane:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq +1 - \xi_i \text{ for } y_i = +1, \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq -1 + \xi_i \text{ for } y_i = -1, \\ \xi_i &\geq 0 \ \forall i, \end{aligned}$$

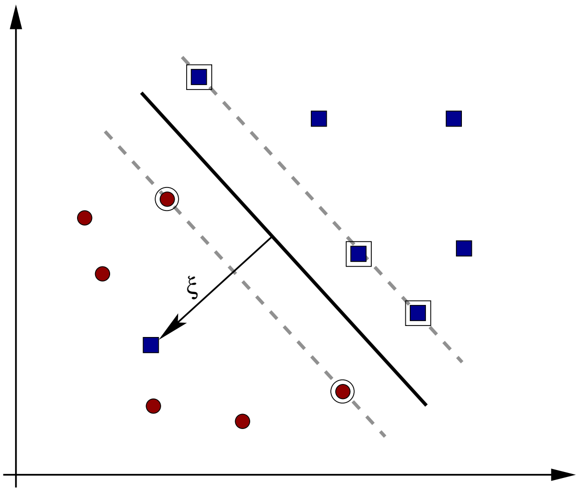


Figure 5.7: Within the soft margin approach, a certain number of violations to the separation is allowed, which is achieved by introducing slack variables ξ referring to data points located at the "wrong"side of the hyperplane.

which can again be summarized:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i \;\; \forall i.$$

Deviating data vectors should nevertheless be an exception. By introducing a penalty term $C \sum_i^l \xi_i^k$ in the cost function, costs are increased for slack variable occurrences:

$$\frac{\| \mathbf{w} \|^2}{2} + C \sum_i^l \xi_i^k.$$

C is a regularization parameter chosen by the user or a hyperparameter selection algorithm. A higher value for C corresponds to a stronger penalization of slack 0. Note that data lying within the margin on the correct side of the hyperplane are also penalized. For the power k, a positive integer, the problem is convex and for $k = 1$ and $k = 2$, the problem is even quadratic. Furthermore, if $k = 1$, neither ξ_i nor their Lagrange Multipliers are part of the dual problem. The expression

$$L_D = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{5.37}$$

must be maximized with respect to the constraints $0 \leq \alpha_i \leq C$ and $\sum_i^l \alpha_i y_i = 0$, resulting in

$$\mathbf{w} = \sum_{i=1}^{N_s} \alpha_i y_i \mathbf{x}_i, \tag{5.38}$$

where N_s is the number of Support Vectors. The only difference to the hard-margin approach lies in the upper bound C for the values α_i. The primal Lagrangian then

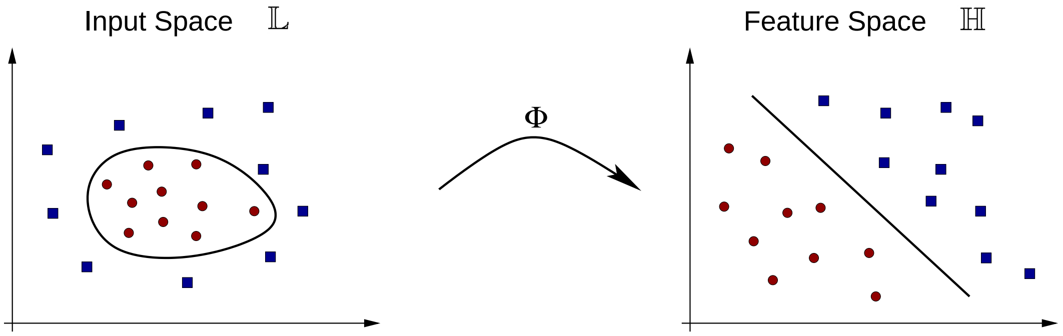


Figure 5.8: Data vectors which are not linear separable in the the input space $\mathbb{L}$ become linear separable after a transformation Φ into a feature space $\mathbb{H}$.

turns into

$$L_P = \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] - \sum_{i=1}^{l} \mu_i \xi_i .$$

The resulting Karush-Kuhn Tucker conditions are

$$\begin{aligned}
\frac{\partial}{\partial w_\nu} L_P = w_\nu - \sum_{i=1}^{l} \alpha_i y_i x_{i\nu} &= 0 && (5.39)\\
\frac{\partial}{\partial b} L_P = - \sum_{i}^{l} \alpha_i y_{i=1} &= 0 && (5.40)\\
\frac{\partial}{\partial \xi_i} L_P = C - \alpha_i - \mu_i &= 0 && (5.41)\\
y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i &\geq 0 && (5.42)\\
\xi_i &\geq 0 && (5.43)\\
\alpha_i &\geq 0 && (5.44)\\
\mu_i &\geq 0 && (5.45)\\
\alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] &= 0 && (5.46)\\
\mu_i \xi_i &= 0. && (5.47)
\end{aligned}$$

Again, b can be calculated using the complementarity conditions (5.46) and (5.47).

Non-linear Support Vector Machines

A trick to deal with linearly non-separable data is to transfer them into another, usually higher-dimensional dataspace $\mathbb{H}$ where the data become linear separable (see Figure 5.8). This transformation can be constructed by using the mapping $\Phi : \mathbb{L} \rightarrow \mathbb{H}$. Then, the training algorithm does not work on the scalar products in the data space $\mathbb{L}$ but in the feature space $\mathbb{H}$. For this purpose, the expression $\mathbf{x}_i \cdot \mathbf{x}_j$ in in the dual Lagrangian (5.37) is replaced by $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$.

But instead of explicitly calculating $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, a kernel function is capable to compute the mappings implicitly, resulting in much lower computational costs. A kernel function is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j), \tag{5.48}$$

and the dual Lagrangian (5.37) becomes

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \tag{5.49}$$

Thus, only the function $K(\mathbf{x}_i, \mathbf{x}_j)$ is required for the training algorithm and it is not even necessary to know exactly what lies behind Φ. All other considerations and computations from the previous sections still hold. Depending on the question under investigation, different kernels can be employed. A very common kernel, which works well for a large variety of applications is the *Gaussian* or *radial-basis function* (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2\right). \tag{5.50}$$

Here, the dimensionality of the feature space is infinite (Cristianini and Shawe-Taylor, 2000). It would therefore be hard to work with Φ explicitly. In the training algorithm, instead of using $\mathbf{x}_i \cdot \mathbf{x}_j$, the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ will be used resulting in a Support Vector Machine working within this infinite feature space. All considerations of the previous sections hold since the linear separation still happens - but only in another space. In the test phase, the sign of the following function reveals the class ($\mathbf{s}_i$ are the Support Vectors):

$$\begin{aligned} f(x) &= \sum_{i=1}^{N_s} \alpha_i y_i \Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{x}) + b & (5.51) \\ &= \sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b. & (5.52) \end{aligned}$$

At no point, Φ needs to be calculated explicitly. Rather, calculating the kernel function K is sufficient. It must nevertheless fulfill some constraints. Necessary preconditions are symmetry $K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z}, \mathbf{x})$ and the Cauchy-Schwarz Inequality

$$K(\mathbf{x}, \mathbf{z})^2 = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}))^2 \leq \| \Phi(\mathbf{x}) \|^2 \| \Phi(\mathbf{z}) \|^2 . \tag{5.53}$$

Sufficient is *Mercer's Condition*. It says that e.g. for any kernel that can be described as

$$K(\mathbf{x}, \mathbf{y}) = \sum_{p=0}^{\infty} c_p (\mathbf{x} \cdot \mathbf{y})^p$$

and converges uniformly, a feature space exists (c_p are positive real coefficients). More details can be obtained by (Schoelkopf and Smola, 2002). Common kernel functions are:

$$\begin{aligned} K(\mathbf{x},\mathbf{y}) &= \mathbf{x}\cdot\mathbf{y} && \text{(linear kernel)} && (5.54)\\ K(\mathbf{x},\mathbf{y}) &= (\mathbf{x}\cdot\mathbf{y})^d && \text{(polynomial kernel)} && (5.55)\\ K(\mathbf{x},\mathbf{y}) &= \exp\left(-\gamma \| \mathbf{x}-\mathbf{y} \|^2\right) && \text{(RBF kernel)} && (5.56)\\ K(\mathbf{x},\mathbf{y}) &= \tanh(\kappa\mathbf{x}\cdot\mathbf{y}-\delta) && \text{(sigmoidal neural net kernel)} && (5.57) \end{aligned}$$

While kernel functions allow to perform non-linear separation with originally linear classifiers whenever those can be described in a dual form, they have the drawback that it usually costs more efforts to find optimal hyperparameters. For instance, in linear SVMs, only the hyperparameter C needs to be chosen carefully. On the other hand, a SVM with RBF kernel further employs the hyperparameter γ. Finding the optimal combination of C and γ therefore becomes expensive when e.g., scanning C and γ with 10 values each, resulting in 100 trainings of the RBF SVMs, instead of just 10 trainings in the linear case (see also section 5.3.4).

5.3.3 Fisher's Linear Discriminant

A much simpler Machine-Learning classifier is *Fisher's Linear Discriminant Analysis* (FLDA) (Fisher, 1936). As in Linear Support Vector Machines, a linear hyperplane is derived from training data, revealing a weight vector $\mathbf{w}$ and a bias b (Bishop, 1995). Like in SVMs, projecting a test vector $\mathbf{x}_{\text{test}}$ on $\mathbf{w}$ yields

$$y = \mathbf{w}\cdot\mathbf{x}_{\text{test}} \tag{5.58}$$

and under consideration of a bias b, a class label can be calculated by the expression $y = \text{sgn}(\mathbf{w}\cdot\mathbf{x}_{\text{test}} + b)$. But in contrast to SVMs, $\mathbf{w}$ is derived in a different way (Duda et al., 2000): Given training data vectors $\mathbf{x}$ from two classes $\mathcal{C}_1$ and $\mathcal{C}_2$, the goal is here to find a projection which maximizes the class separation. To derive a measure for separation, a mean vector

$$\mathbf{m}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in\mathcal{C}_i}\mathbf{x} \tag{5.59}$$

for each class is required with N_i denoting the number of samples in the class $\mathcal{C}_i$. The projection of the mean vectors of class $\mathcal{C}_i$ onto $\mathbf{w}$ is then

$$\mu_i = \mathbf{w}^T\mathbf{m}_i \tag{5.60}$$

and a good separation of the classes should be expected when $\mathbf{w}$ lies in such a direction that the mean vectors of the classes expose a long distance in the projection

$$\mu_1 - \mu_2 = \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2). \tag{5.61}$$

Beside maximizing the distance in the projection, also the variance of the data within the classes should be considered. As Figure 5.9 illustrates, although a larger distance of the means would be obtained for projections onto $\mathbf{x}_1$, a clearer class separation would result for projections onto $\mathbf{x}_2$ due to the smaller variances along this direction. For this

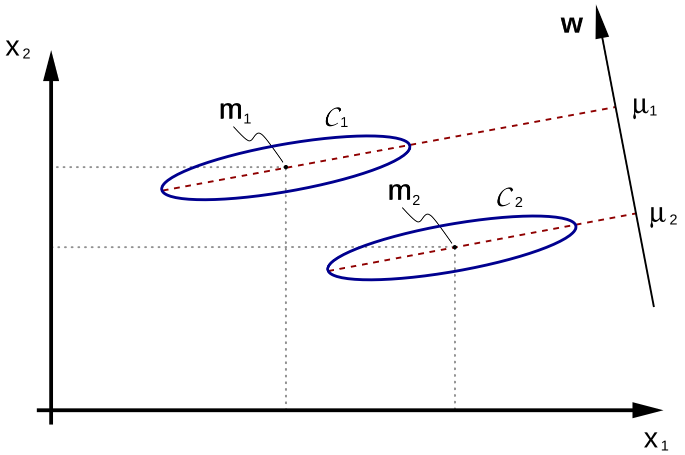


Figure 5.9: The blue circles represent data distributions of the classes $\mathcal{C}_1$ and $\mathcal{C}_2$. A data projection maximizing the distances between the means of the distributions is computed in Fisher's Linear Discriminant. Although the projection along $\mathbf{x}_1$ results in a larger distance of the means, the projection onto $\mathbf{x}_2$ allows for a better class separation due to a smaller variance along this direction. Thus, a score must also take into account the variances of the data distributions which results in a projection $\mathbf{w}$ as depicted within the figure. The Fisher criterion $J(\mathbf{w})$ provides such a measure.

purpose, the scatter

$$\sigma_i^2 = \sum_{n \in \mathcal{C}_i} (y_n - \mu_i)^2 \tag{5.62}$$

within the classes are employed by considering the total within-class covariance $\sigma_1^2 + \sigma_2^2$ in order to compute the *Fisher criterion*

$$J(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}. \tag{5.63}$$

Fisher's Linear Discriminant then employs that function $\mathbf{w}^T\mathbf{x}$ for which $J(\mathbf{w})$ is maximum. This optimal $\mathbf{w}$ can be determined by finding an expression of $J(\mathbf{w})$ which is an explicit function of $\mathbf{w}$. For this purpose, first the scatter matrix $\mathbf{S}_i$ needs to be defined:

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T. \tag{5.64}$$

Then, with the help of (5.58) and (5.60), the scatter can be reformulated as

$$\begin{aligned}
\sigma_i^2 &= \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^2 && (5.65)\\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{w} && (5.66)\\
&= \mathbf{w}^T\mathbf{S}_i\mathbf{w}. && (5.67)
\end{aligned}$$

The *within-class scatter matrix*

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{5.68}$$

is proportional to the sample covariance matrix for pooled d-dimensional data. With $\mathbf{S}_W$, the denominator of the Fisher criterion (5.64) can now be formulated as

$$\sigma_1^2 + \sigma_2^2 = \mathbf{w}^T \mathbf{S}_W \mathbf{w}. \tag{5.69}$$

Employing (5.60), its numerator can be calculated as

$$\begin{aligned}(\mu_1 - \mu_2)^2 &= (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2 && (5.70)\\ &= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} && (5.71)\\ &= \mathbf{w}^T\mathbf{S}_B\mathbf{w} && (5.72)\end{aligned}$$

with the *between-class scatter matrix*

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T. \tag{5.73}$$

Taken together, the Fisher criterion can be rewritten as

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w^T}\mathbf{S}_B\mathbf{w}}{\mathbf{w^T}\mathbf{S}_W\mathbf{w}} \tag{5.74}$$

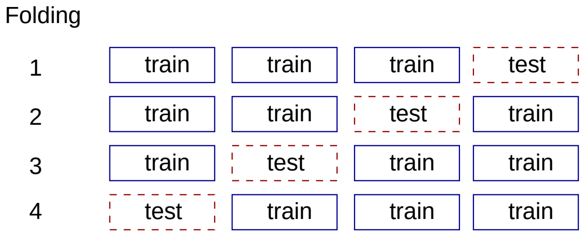
which is commonly known as the generalized *Rayleigh coefficient*. As shown in e.g. Bishop (1995), calculating $\frac{\partial \mathbf{J}(\mathbf{w})}{\partial \mathbf{w}} = 0$ reveals that $\mathbf{J}(\mathbf{w})$ is maximized when

$$(\mathbf{w^T}\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w} = (\mathbf{w^T}\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w} \tag{5.75}$$

which can, under consideration of (5.73), be simplified to the *canonical variate*

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{5.76}$$

Thus, a convenient solution for $\mathbf{w}$ can be computed by just inverting $\mathbf{S}_W$ and calculating the class means. $\mathbf{S}_W$ is symmetric and positive semidefinite and usually non-singular if the number of samples n exceeds the data dimension d (Duda et al., 2000). In order to be able to perform classifications, the bias b must be determined, which is a threshold along the one-dimensional subspace separating the projected data vectors. If the probability densities for the classes are both multivariate normal with equal covariance matrices, the Fisher discriminant is the Bayes optimal solution (Bishop, 1995). In this case, b is the point where the posterior probabilities are equal. On the other hand, if the data distribution is not normal, b can be found as a hyperparameter in a cross-validation scheme. Fisher's Linear Discriminants can easily be generalized for multiple classes (Bishop, 1995) and the kernel trick from the previous section can also be employed for this approach (Müller et al., 2001; Mika et al., 2001).



Folding				
1	train	train	train	test
2	train	train	test	train
3	train	test	train	train
4	test	train	train	train

Figure 5.10: Each of the 4 subsets of a dataset within a 4-fold cross-validation serves once as a test set. In each folding, three subsets are aggregated and constitute the training set for this folding while the remaining set serves as a test set. Thus, training and test sets are always disjoint. The datasets are systematically permutated and the average classification rate on the test sets in the different foldings gives an approximation about the classification performance on unseen data of the classifiers.

5.3.4 Cross-Validation

It is crucial to have disjoint sets for training and testing to avoid overfitting of a Machine-Learning classifier (see section 5.3.1). In order to assess the performance of a classifier, *cross-validation* can be performed. Cross-validation divides a dataset into k subsets of data for a *k-fold cross-validation* and takes $k-1$ subsets for training and the omitted subset for testing (Bishop, 1995). The sets are permutated and the average rate of the k test set classifications is taken as a measure for the classification rate as depicted in Figure 5.10. This can especially be important for finding optimal hyperparameters for a Machine-Learning classifier. By e.g. varying the hyperparameter C for a linear SVM in a systematical way and assessing the average classification performance, the "optimal"parameter can be found. When e.g. using RBF kernel SVMs, scanning value-combinations for the both hyperparameters γ and C is a common strategy (Keerthi and Lin, 2003).

5.4 Symbol Inference

The outcome of the whole data analysis procedure for the P300 speller BCI should be a symbol from the stimulus matrix, a subject directed attention to (cf. Figure 5.11). As introduced in section 4.1, when a subject directs attention towards one specific symbol within the stimulus matrix which rows and columns are flashing in random order, a P300 component is induced in the EEG time series (*epoch*) after a flashing of that specific symbol. Thus, the attended symbol can reversely be inferred from the basic distinction of P300 ($\mathcal{P}^+$) and non-P300 ($\mathcal{P}^-$) samples of EEG time series. It is necessary to determine exactly one row and one column by occurrences of P300 components to infer the symbol which would lie in the intersection of the row and the column. More formally, from an EEG time series $\mathbf{x}$, a symbol inference algorithm needs to identify that row i_r and that column i_c which is most likely to be associated with a P300: Among the rows and the columns, the winners i_r and i_c have to be chosen from the row indices $\mathcal{I}_{\text{rows}} = \{r_1, r_2, r_3, r_4, r_5, r_6\}$ and the column indices $\mathcal{I}_{\text{columns}} = \{c_1, c_2, c_3, c_4, c_5, c_6\}$, respectively. As already introduced in Kaper et al. (2004) and Meinicke et al. (2003), this can be accomplished by computing a certain *score* $S_i^\Psi(\mathbf{x})$ containing classification

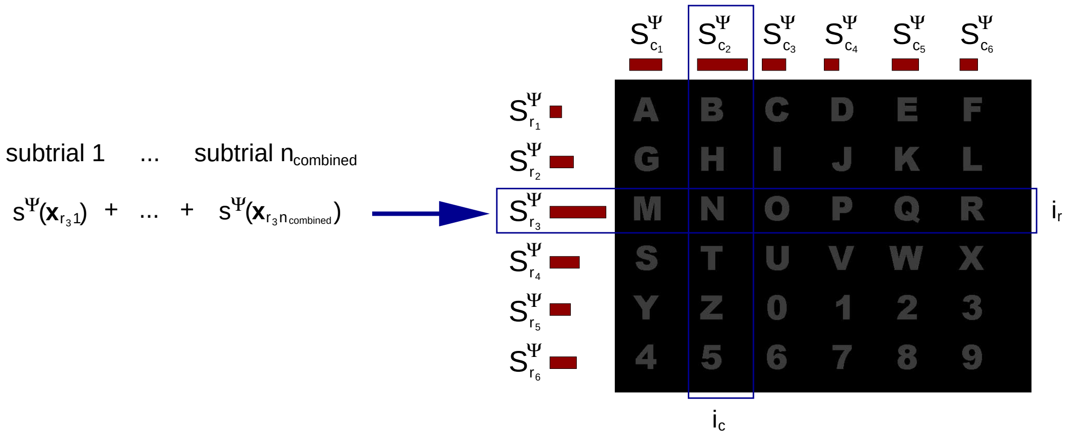


Figure 5.11: The score $S_i^{\Psi}(\mathbf{x})$, abbreviated as S_i^{Ψ}, is reflected by red bars in the figure. It is calculated as the sum of the subtrial scores $s^{\Psi}(\mathbf{x}_{ik})$ from the number of n_{combined} aggregated subtrials. Thereby, Ψ refers to a certain classifier and a subtrial refers to a sequence of flashing each row and each column once. Among the rows and the columns, S_i^{Ψ} with the highest value yields the index i_r and i_c, respectively, indicating the position of the symbol within the matrix.

results for the specific row or column i as derived from a classifier Ψ (e.g. a Support Vector Machine). The row and the column with the highest score among the rows and the columns is selected as the target row or column:

$$i_r = \arg\max_{i \in \mathcal{I}_{\text{rows}}} S_i^{\Psi}(\mathbf{x}) \quad \text{and} \quad i_c = \arg\max_{i \in \mathcal{I}_{\text{columns}}} S_i^{\Psi}(\mathbf{x}). \tag{5.77}$$

As section 4.1 reveals, it is usually not sufficient to employ just one subtrial[10] for classification. In contrast, a certain number of repetitions needs to be performed to reach reasonable classification accuracies. As it is depicted in Figure 5.11, the score $S_i^{\Psi}(\mathbf{x})$ must therefore also incorporate these repetitions, which can be achieved by accumulating the *epoch scores* $s^{\Psi}(\mathbf{x}_{ik})$ of the time series $\mathbf{x}_{ik}$ belonging to flashes of the specific rows and columns from n_{combined} subtrial repetitions:

$$S_i^{\Psi}(\mathbf{x}) = \sum_{k=1}^{n_{\text{combined}}} s^{\Psi}(\mathbf{x}_{ik}). \tag{5.78}$$

In this thesis, $s^{\Psi}(\mathbf{x}_{ik})$ is derived from a classifier $\Psi \in \{\text{Area, PP, FLDA, LSVM, RBF SVM}\}$, but others are possible. Among these classifiers, scores from the Model-Based approaches *area* and *peak picking* (PP) as discussed in section 4.1 can be calculated as

$$s^{\text{Area}}(\mathbf{x}_{ik}) = \sum_{t=t_1}^{t_2} x_{ik}(t), \tag{5.79}$$

$$s^{\text{PP}}(\mathbf{x}_{ik}) = \min_{t<t1} x_{ik}(t) - \max_{t>t1} x_{ik}(t), \tag{5.80}$$

$$\tag{5.81}$$

[10] Within a subtrial, each row and each column is flashed once (cf. section 4.1).

Table 5.1: Example for the calculation of accuracy rates in symbol inference for the symbol "S". From $n_{\text{subPerSymbol}} = 42$ subtrials belonging to one symbol, n_{combined} subtrials are aggregated, such that only a fraction of the symbols that can be inferred are remaining for $n_{\text{combined}} > 1$. Correct inferences are underlined and the accuracy p_{acc} is calculated as the ratio $n_{\text{correct}}/n_{\text{inferences}}$.

n_{combined}	**Inferred Symbols**	n_{correct}	$n_{\text{inferences}}$	p_{acc}
1	MKKSS4YYG79XS6USXCTSSTXUFNGGAVOSZW8VXBSSX8	9	42	0.21
2	SS4YYXYUSSSSCGS4UV3S3	8	21	0.38
3	GSY9USSSCANWSS	6	14	0.43
4	SS4SSSASWS	7	10	0.70
5	SYXSSASS	5	8	0.63
6	SSSSASS	6	7	0.86
7	SSSSSS	6	6	1.00

with t_1 being the onset of the P300 deflection, and t_2 being its offset. On the other hand, scores for the Machine-Learning approaches as considered in this thesis are derived as

$$s^{\text{FLDA}}(\mathbf{x}_{ik}) = \mathbf{w} \cdot \mathbf{x}_{ik} + b, \tag{5.82}$$

$$s^{\text{LSVM}}(\mathbf{x}_{ik}) = \mathbf{w} \cdot \mathbf{x}_{ik} + b, \tag{5.83}$$

$$s^{\text{RBF SVM}}(\mathbf{x}_{ik}) = \sum_{j=1}^{N_s} \alpha_j y_j K(\mathbf{s}_j, \mathbf{x}) + b, \tag{5.84}$$

among which FLDA and LSVM are linear approaches, whereas RBF SVM utilizes the kernel trick and incorporates a Gaussian kernel to deal with data which are not linear separable (cf. section 5.3.2). The approaches FLDA and LSVM each determine a separating hyperplane in the input space $\mathbb{L}$ as indicated by the vector $\mathbf{w}$, which can be calculated by (5.76) and (5.38), respectively. On the other hand, the RBF SVM score only relies on a hyperplane in the feature space $\mathbb{H}$, and not in the input space $\mathbb{L}$ (cf. Figure 5.8). Thus, it is necessary to calculate the score $s^{\text{RBF SVM}}(\mathbf{x}_{ik})$ from the Support Vectors $\mathbf{s}_j$ and the according Lagrange Multipliers α_j as well as their class labels y_j in the input space as also shown in (5.52). Finding the bias b is explained in section 5.3.2 for the Support Vector Machines and in section 5.3.3 for Fisher's Linear Discriminant. While it is common to employ the label of the predicted class in these binary classifiers, in this context, it is more appropriate to employ the real-valued scores as listed above to avoid ambiguities.

When performing offline analyses, i.e., data are collected and analyzed after the experiment (see chapter 6), a certain number $n_{\text{subPerSymbol}}$ of subtrials will result for each symbol. From these data belonging to a certain symbol, different numbers of subtrials n_{combined} can be aggregated in order to compute $S_i^{\Psi}(\mathbf{x})$. In the simplest case, $n_{\text{combined}} = 1$, and an accuracy rate for classification performance can be derived as follows: Let $n_{\text{subPerSymbol}}$ be the number of subtrials which were collected for a symbol, and n_{correct} the number of correctly inferred symbols (see top row of Table 5.1 for an example). With $n_{\text{combined}} = 1$, the accuracy p_{acc} can then be computed as

$$p_{\text{acc}} = n_{\text{correct}}/n_{\text{subPerSymbol}}. \tag{5.85}$$

In this case, the number of possible inferences for a certain symbol equals the number

$n_{\text{subPerSymbol}}$ of subtrials for this symbol. But this equation is only valid when just one subtrial is employed for a symbol inference. In case of employing data from more than one subtrial ($n_{\text{combined}} > 1$), the number of possible symbol inferences decreases to

$$n_{\text{inferences}} = \inf(n_{\text{subPerSymbol}}/n_{\text{combined}}) \tag{5.86}$$

when aggregating succeeding subtrials. Thus, as illustrated in Table 5.1, for $n_{\text{combined}} = 1$, the number of inferences is 42, while it decreases to 21 for $n_{\text{combined}} = 2$, to 14 for $n_{\text{combined}} = 3$, and so on. Since the number of possible symbol inferences decreases with the number of aggregated subtrials, classification accuracies are also determined on the basis of $n_{\text{inferences}}$:

$$p_{\text{acc}} = n_{\text{correct}}/n_{\text{inferences}}. \tag{5.87}$$

Note that the resolution for accuracies decreases with the possible number of inferences $n_{\text{inferences}}$. When aggregating data from e.g., 15 subtrials, only two symbols can be inferred in this example from the 42 symbols, resulting in possible accuracies of only 0%, 25%, 75% and 100%, respectively.

5.5 Summary

Analyzing EEG data to operate the P300 speller paradigm can be subdivided into the steps *data acquisition, preprocessing, classification*, and *symbol inference.*

In data acquisition, a high data quality can be obtained by low impedances, using a sound attenuated and electromagnetically shielded experimental chamber. Electrode materials, electrode locations, and amplifier design further influences the data quality.

Preprocessing aims to produce adequate data representations (feature vectors) for subsequent steps of analysis. Filtering by, e.g., band pass filtering is capable to eliminate influences from frequencies which are not under consideration. This is especially important for single trial analysis as performed in Brain-Computer Interfacing. Reducing the dimensionality of the feature vectors in order to achieve a data representation which follows the intrinsic dimensionality of the data can be achieved by Principal Component Analysis.

Classification in this context aims to relate the feature vectors with a label, reflecting either the class “P300” ($\mathcal{P}^+$) or “non-P300” ($\mathcal{P}^-$). For this purpose, Model-Based techniques or Machine-Learning techniques can be employed. While the former techniques depend on explicit knowledge of the problem structure, the latter learn from given data to perform such a classification. Statistical Learning Theory provides a framework for Machine-Learning problems. Its concept of Structural Risk Minimization is employed in the classification technique Support Vector Machines. They rely on finding a hyperplane which lies in the middle between data classes and can easily be extended to non-linear problems by employing the kernel trick. A much simpler classification technique is Fisher’s Linear Discriminant, which finds a separating hyperplane by maximizing the ratio of the between-class and within-class scatter, as encoded in the Fisher criterion, and can easily be computed.

Based on the results of these classifiers, symbol inference calculates scores for each row and each column from several repetitions of subtrials. These scores finally indicate the row and column indices of an entry in the stimulus matrix.

Chapter 6

Offline Experiments

Before designing Brain-Computer Interfaces which can actually be *operated* by a user, it is necessary to find and optimize appropriate preprocessing and classification algorithms and evaluate parameters under which the system works best. For this purpose, it is desirable to have reliable data at hand, which were produced under controlled conditions. Such data can be generated within so-called *offline* BCI experiments, in which data of a subject are recorded and analyzed *after* the experiment itself (see Figure 6.1). In contrast, in *online* BCI experiments, data are processed and analyzed during the experiment, such that the user can operate the system. Throughout the offline experiments presented within this chapter, the goal of online BCI experiments will be kept in mind and drive the experimental questions. The foundation stones of the work were set in (Hoppe and Kaper, 2003).

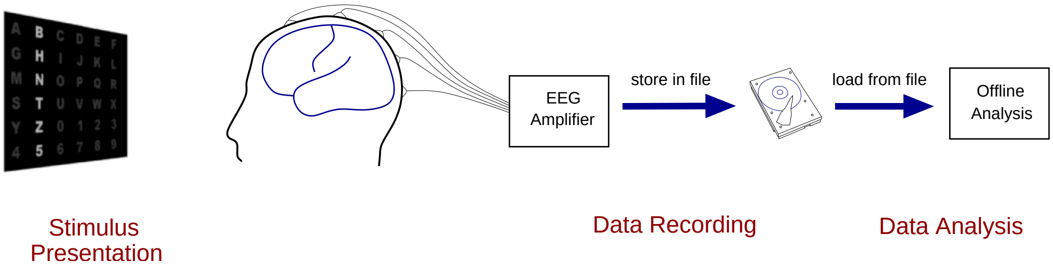


Figure 6.1: Scheme for *offline* BCI analysis within the P300 speller paradigm. Stimuli are presented which induce specific brain signals depending on the subject's attention. Electroencephalography measures electrical potentials from the scalp and stores the data in a file. These data are then analyzed after the experiment was performed.

In this chapter, offline experiments are performed with the goal to find optimal stimulus parameters, classifiers, and preprocessing strategies for driving online experiments. After explaining the general experimental method in section 6.1, preprocessing parameters are derived from a first experimental series and common Model-Based classification algorithms are compared with Machine-Learning classifiers (section 6.2) based on data from the electrode at the location Pz[1] (cf. Figure 6.2). Afterwards, classification performance is further improved by analyzing data from a larger set of electrodes in section 6.3. In order to accelerate the speed of the BCI device and to evaluate its generalization performance, the presentation speed is increased within section 6.4 and more subjects are considered. The capabilities to generalize to new subjects, i.e., whether it is possible

[1] This site was also the location for EEG recordings in the work of Farwell and Donchin (1988) and Donchin et al. (2000) as detailed out in chapter 4.

to train a classifier on data from certain subjects and use it for classifications of data from other subjects, are investigated in section 6.5. This would allow to apply a classifier without prior classifier training for the individual subject. Finally, the performance of an earlier version of the algorithm, as assessed within the BCI Competition 2003 (Blankertz et al., 2004), is described in section 6.6.

6.1 General Experimental Setup

Throughout the chapter, the general experimental setup will stay the same. Those aspects of *data acquisition, stimulus presentation*, and *data analysis* that will be constant are introduced in the following.

Data Acquisition

EEG electrodes were applicated to the positions $Fz, Cz, Pz, Oz, C3, C4, P3, P4, PO7$, and $PO8$[2], respectively using Ag/AgCl electrodes and a *Neuroscan Synamps 5083* amplifier (see Figure 6.2). The experimental chamber was sound attenuated and shielded from electromagnetic influences with a Faraday cage. Impedances of about 2kΩ were aimed for, which could not be achieved in every case, such that some electrodes exposed higher impedances. Nevertheless, impedance was always below 11kΩ. No correction for EOG artifacts was performed.

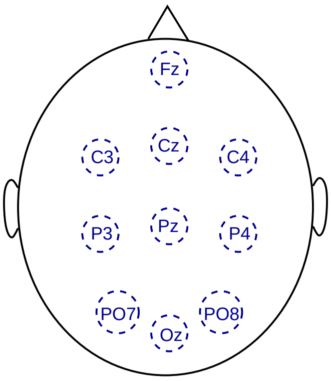


Figure 6.2: Electrode locations of the experiments conducted in this chapter.

Stimuli

After applying EEG electrodes, the subjects were instructed to mentally count the flashings of a symbol that was chosen by the presentation program, and given to the subject in advance of each trial (see Figure 6.3). In order to familiarize the subject with the program, some training trials were performed prior to each experiment, the data of which were not recorded. The expansions of the stimulus matrix for the subject were $14.5° \times 10°$.

The subjects performed a number of trials within an experiment, each subdivided into a constant number of subtrials. Thereby, a subtrial consists of a sequence of 12

[2] The locations $PO7$ and $PO8$ are 20% away from the Oz electrode towards the temporal lobe. These sites were chosen due to their proximity to the Pz site, which is known to expose the largest P300 amplitudes (see section 2.5).

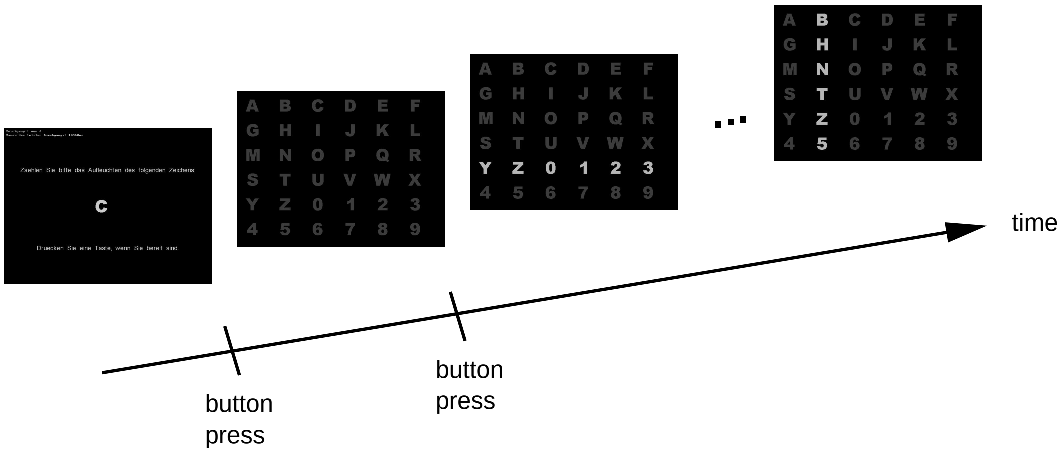

Durchgang 1 von 6
Dauer des letzten Durchgangs: 14560ms
Zaehlen Sie bitte das Aufleuchten des folgenden Zeichens:
C
Druecken Sie eine Taste, wenn Sie bereit sind.
A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 0 1 2 3
4 5 6 7 8 9
...
time
button press
button press

Figure 6.3: Sequence of a trial within an offline experiment. After the presentation program randomly chooses the symbol to concentrate on, the subject can proceed to a matrix presentation by pressing a button. A second button press initiates the flashing sequence.

events where each column and each row of the matrix was flashed exactly once in random order. The duration of the flashes, the so-called *Interstimulus Interval* (ISI, see section 4.1), was varied among the experiments discussed in this chapter. EEG data were recorded with respect to the certain experimental events, such that time series of a specific length following a stimulus presentation were recorded. These series are referred to as *epochs*. Note that depending on the ISI and the time window, overlaps of consecutive time series can occur.

Data Analysis

Data were recorded with a samplingrate of 200Hz and band pass filtered within the band 0.5Hz-30Hz at 96dB by the amplifier to correct for drifts and the power supply frequency of 50Hz (cf. sections 2.3 and 5.2.1). The stimulus presentation was driven by a program under `MS-DOS` since it is no multitasking operating system, which makes it easy to produce software with stable clocking properties. The proprietary Neuroscan software `Acquire 4.1.1` for driving the amplifier and recording the data was designed for Windows, such that `Windows 98` was used to record the data. Finally, data were analyzed using `Matlab` in a `Linux` environment because distributed calculations can be performed in the Neuroinformatic's `Linux` cluster. Since establishing a communication between these three operating systems is difficult and the Neuroscan amplifier does not offer direct access to the EEG data, designing an online BCI in this framework is a delicate task (cf. section 7.1).

After recording, the data were analyzed using `Matlab` and the Support Vector Machine algorithms from the `libsvm` toolbox of Chang and Lin (2001). For preprocessing, data were band pass filtered with parameters as determined in section 6.2.1. Afterwards, the amplitudes were scaled to the interval [-1,1]. For the different classifiers, it is necessary to scan a number of *hyperparameters*, i.e. parameters which need to be adjusted to control the behavior of the classifiers: For Linear SVMs, the Parameter C controlling the violations to the classification by slack variables (cf. section 5.3.2) needs to be chosen. RBF SVMs incorporate the additional hyperparameter γ for the

Table 6.1: In order to find reasonable hyperparameters for the different classifiers Linear SVM (LSVM), SVM with Gaussian kernel (RBF SVM), and Fisher's Linear Discriminant (FLDA), parameters were systematically scanned, and the corresponding classification performance was assessed in a 5-fold cross-validation. For this purpose, a start value was multiplied with a specific factor in each step. In case of RBF SVM, after finding first “optimal”parameter values C_{opt} and γ_{opt}, this procedure was repeated within close proximity of these values. In case of FLDA, the start value was added by 0.02 in each step.

Classifier	**Parameter**	**Start Value**	**Increase**	**Steps**
LSVM	C	1×10^{-5}	× 1.06	200
RBF SVM (1)	C	1×10^{-2}	× 2	25
	γ	1×10^{-7}	× 2	25
RBF SVM (2)	C	$C_{\text{opt}} \times \left(\frac{1}{1.1}\right)^5$	× 1.1	10
	γ	$\gamma_{\text{opt}} \times \left(\frac{1}{1.1}\right)^5$	× 1.1	10
FLDA	b	-2	+0.02	200

bandwidth of the Gaussian kernel (see section 5.3.2). For Fisher's Linear Discriminant, the bias b needs to be chosen (see section 5.3.3). Values for the hyperparameters were systematically varied with heuristic factors (Chang and Lin, 2001), and the resulting classification accuracy was then assessed on disjoint test sets in a cross-validation manner (cf. section 5.3.4). Parameters exposing the best average classification accuracies in the cross-validation are regarded as being optimal. For this purpose, a start value for each hyperparameter is multiplied by a factor in a certain number of steps for the different classifiers (Table 6.1). Note that in the case of FLDA, the bias b is varied in a linear fashion: From a start value, a constant value is added in each step. In the case of RBF SVMs, each *pair* of hyperparameters (C, γ) was used for training and testing. Unfortunately, it would be computationally too expensive to look for optimal hyperparameters for both C and γ with 200 steps each, as it is performed in FLDA and LSVM as well, since it would result in 40000 evaluations. Therefore, two levels of parameter selection were employed: After finding a first “optimal”hyperparameter combination of C and γ within 25 steps for each hyperparameter, a finer search was conducted of 10 steps each within close proximity of this parameter combination (see Table 6.1, rows 4 and 5).

6.2 Improving Classification Accuracies on Single Electrode Data

The goal of this first investigation is to derive appropriate preprocessing parameters and to compare preprocessing and classification strategies on data solely acquired from the Pz electrode. P300 components are most pronounciated at this site and Farwell and Donchin (1988) also utilized this site for their analyses. To gain first insights into the data, especially about which features might be useful, one can look at the Event-Related Potential (ERP, see section 2.5) and the power spectra of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples, i.e., samples belonging to the correct row or column of the given symbol, or not. In other words, the ***oddball event*** (see section 2.5) “correct row”or “correct column”produces a $\mathcal{P}^+$ epoch in the EEG time series.

Two male subjects (1A: 24 years, 1B: 26 years) participated in this experiment, each performing 140 trials. A trial was divided into 3 subtrials (cf. section 4.1), resulting in 420 subtrials, and therewith in 840 $\mathcal{P}^+$ and 4200 $\mathcal{P}^-$ samples. In the data of subject 1A, an invalid trial occurred due to drift corrections performed by the amplifier, resulting in only 834 $\mathcal{P}^+$ and 4170 $\mathcal{P}^-$ samples. Although data were recorded from the whole set of electrodes as described above, only data from the Pz electrode were employed for analyses within this experiment. After recording, data were analyzed offline. Ten out of the 36 symbols were randomly chosen by the presentation program ($n_{\text{symbol}} = 10$), and each symbol was repeated 14 times, resulting in $n_{\text{subPerSymbol}} = 3 \cdot 14 = 42$ subtrials per symbol. The Interstimulus Interval was set to 500ms (485ms highlighting + 15ms delay, see also section 4.1). In advance of each trial, the subject was given a symbol as randomly chosen from the presentation program. The matrix was then presented to enable the subject to direct attention to the specified symbol. By pressing a button, the subject could start the flashing sequence (see Figure 6.3). Within a trial, the flashing sequence consisted of a series of 3 subtrials, each lasting 6 seconds and containing 12 flashes, such that each row and each column was highlighted once within a subtrial, and a trial lasted 18 seconds. The sequence of the highlightings was random within each subtrial. Thus, in a subtrial, there are two “positive”samples ($\mathcal{P}^+$), one belonging to the row, and one belonging to the column with the specified symbol. All other 10 samples are “negative”($\mathcal{P}^-$) and should therefore not contain a P300.

6.2.1 Deriving Preprocessing Parameters

While most P300-based BCI approaches employ only heuristic preprocessing parameters, for example 600ms time windows and a 0.02-35Hz band pass filter in Farwell and Donchin (1988), in this section, the time window and the frequency band as well are to be chosen in a more systematical way.

Methods

With the start of each event (i.e., a flashing row/column), an epoch of 2000ms was extracted and Event-Related Potentials as introduced in section 2.5 were calculated by averaging events which should contain a P300 ($\mathcal{P}^+$) and those which should not ($\mathcal{P}^-$), separately. Differences between $\mathcal{P}^+$ and $\mathcal{P}^-$ conditions can be identified by subtracting according ERPs. In a second step, the power spectra of these differences were calculated for each subject to identify relevant frequency ranges.

Results

The ERPs of the experiment are shown in Figure 6.4 (note that in EEG research, negative amplitudes are commonly drawn upwards). While the first row depicts the ERP of $\mathcal{P}^+$ samples, the middle row contains the ERP of $\mathcal{P}^-$ samples, and the third row exposes their differences. Within the ERP of the $\mathcal{P}^+$ samples, negative peaks with a temporal distance of about 500ms, starting with the first peak at about 200ms can be identified for subject 1A. A similar rhythmic structure with more sustained signals exists for subject 1B, where negative deflections also occur each 500ms. While a strong positive deflection within the time frame of 300ms to 600ms is present for subject 1A, subject 1B lacks this component. For $\mathcal{P}^-$ samples, the regularity of the negative peaks remains for both subjects, but no P300 component can be identified for subject 1A.

In the difference $\mathcal{P}^+ - \mathcal{P}^-$, both subjects expose a strong positive deflection between

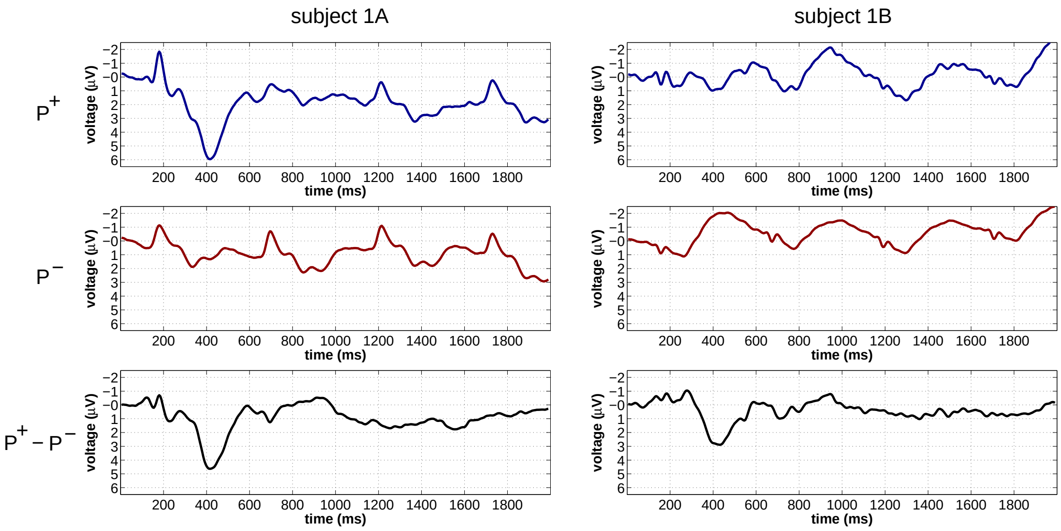


Figure 6.4: ERPs for subject 1A (left) and subject 1B (right). While the first row contains data from $\mathcal{P}^+$ samples, the middle row depicts data from $\mathcal{P}^-$ samples and the bottom row exposes their differences $\mathcal{P}^+ - \mathcal{P}^-$.

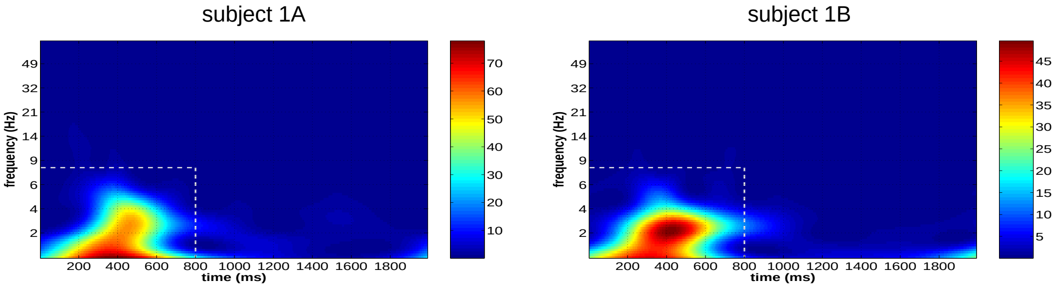


Figure 6.5: Power spectra of the ERP differences for subject 1A (left) and subject 1B (right). The dashed white rectangle indicates the time-frequency frame for subsequent analyses.

300ms and 600ms without any further remarkable components. Power spectra of the $\mathcal{P}^+ - \mathcal{P}^-$ difference also reveal similar components for both subjects. From the beginning to about 800ms, components at a frequency up to about 8 Hz can be identified.

Discussion

The regular negative deflections with a temporal distance of 500ms for subject 1A and subject 1B can be interpreted as N200 components, which are mainly related to the *sensory* processing of a stimulus (see section 2.5). They are likely to reflect the sensory processing of subsequent flashes which occur with a frequency of 500ms (t_{ISI}=500ms). However, in the average of the $\mathcal{P}^+$ samples only the first flash corresponds to the oddball event, and should therefore elicit a P300 as also explained in section 2.5. Thus, the positive deflection in subject 1A's data between 300ms and 600ms can easily be interpreted as a P300 and correlates with the semantic meaning of the stimulus, while this component does barely occur for the other flashes within this 2000ms time window. When computing the differences between the ERPs, the N200 components of their ERPs

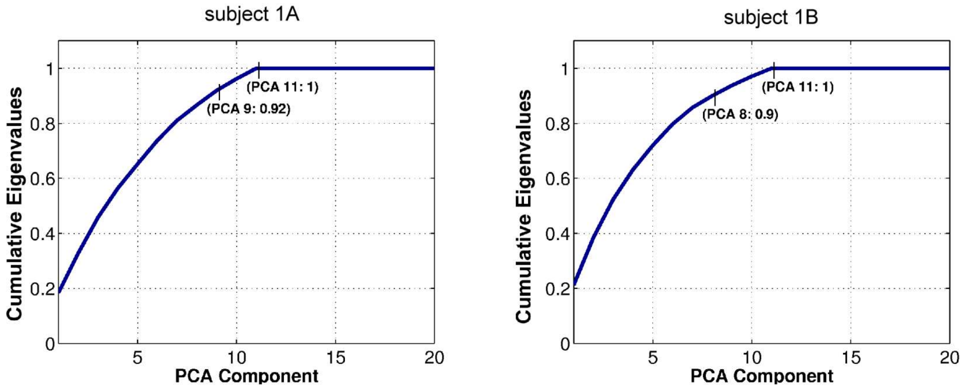


Figure 6.6: Cumulative Eigenvalues as a fraction of the sum of all Eigenvalues, reflecting the degree of captured variance for subject 1A (left) and subject 1B (right).

eliminate each other. Thus, the N200 components are not likely to carry information which could contribute to the classification task. Only the P300 as well as the N200 component prior to the P300 carry this information. This pattern also occurs for subject 1B, although a P300 can not clearly be identified in the ERP of $\mathcal{P}^+$ samples. However, when comparing ERPs of $\mathcal{P}^+$ and $\mathcal{P}^-$, the negative deflection at about 300ms to 600ms in the ERP of $\mathcal{P}^-$ is missing in the ERP of $\mathcal{P}^+$. Thus, due to the comparably slow negative deflections in subject 1B, a superposition of negative and positive components seem to happen, resulting in amplitude values of about 0μV.

The power spectra of the ERP differences reveal components within the time frame of 0ms-800ms with frequencies of not more than 8Hz. Thus, in the remaining sections, feature extraction will rely on a 800ms time frame, corresponding to 160-dimensional data vectors at a samplingrate of 200Hz. Furthermore, a low pass filter of 8Hz will be utilized (cf. section 5.2.1). With this kind of filter, α waves (frequencies between 8 and 13 Hz, cf. section 2.1), which can be a source for artifacts, are a-priori excluded.

6.2.2 Principal Component Analysis

Performing Principal Component Analysis (PCA) reveals information about the variance of a dataset and allows to reduce the dimensionality of the data by projecting them onto a certain number of Principal Components (cf. section 5.2.2). From the Eigenvalues of the resulting 160 Principal Components, the variance captured by each Principal Component can be calculated. The fraction

$$q_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{160} \lambda_j} \tag{6.1}$$

of the specific Eigenvalues λ_i from the sum of all Eigenvalues reflects the degree of variance captured in the components up to this point.

Methods

For performing PCA, data were preprocessed as suggested by the previous section by extracting epochs of 800ms and performing band pass filtering (0.5-8Hz). Since the

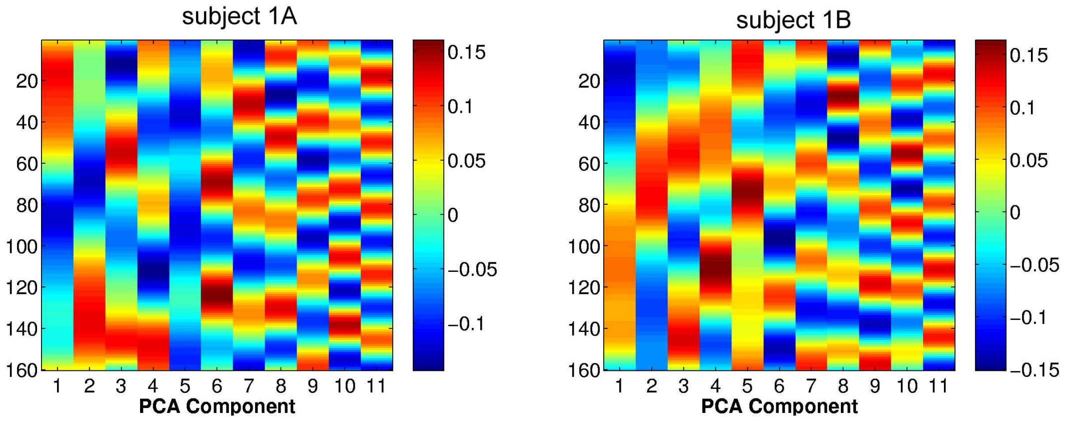


Figure 6.7: PCA matrices for the first 11 Principal Components for subject 1A (left) and subject 1B (right). The columns of the matrices are the Principal Components, which amplitudes are encoded in colors. Apparently, mainly certain frequencies are reflected by the different Principal Components.

number of the $\mathcal{P}^-$ samples exceeds the number of $\mathcal{P}^+$ samples by the factor 5 within this paradigm, a subset of $\mathcal{P}^-$ samples was randomly chosen to construct a balanced set with equal numbers of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples. Then, a balanced set of 834 $\mathcal{P}^+$ and 834 $\mathcal{P}^-$ data samples was extracted. PCA was applied on the data and their Eigenvalues and Principal Components were examined.

Results

The cumulative Eigenvalues for the different Principal Components, reflected as fractions of the sums of all Eigenvalues is depicted in Figure 6.6. A variance of more than 99.9% is captured for both subjects within the first 11 Principal Components. For subject 1A, the first 9 Principal Components reflect 90% variance, while this amount of information is captured in 8 Principal Components in subject 1B.

PCA matrices containing the first 11 Principal Components are shown for both subjects in Figure 6.7, in which the columns contain the different Principal Components. Apparently, predominantly certain *frequencies* are encoded within the different components. While the first component reflects a frequency of about 1Hz (note that 160 data points correspond to 800ms in the time domain), the frequency rises with subsequent components up to about 6Hz in Principal Component 11.

Discussion

According to the Eigenvalues, a high data compression without loss of information from 160 to 11 dimensions is possible for both subjects by using PCA, maintaining 90% of variance in either the first 9 (subject 1A) or 8 (subject 1B) Principal Components. In both subjects, the different Principal Components reflect different frequencies and phase shifts. Thus, rather than decomposing the original data in components which resemble a P300 or other typical EEG components, oscillatory bases were found by PCA. This outcome is related to the findings of Hancock et al. (1992) and Heidemann (2006) who found two-dimensional Gabor-patches of different frequencies when analyzing collections of natural scenes with Principal Component Analysis.

6.2.3 Symbol Inferences using Model-Based Classifiers

In the following, the classification performance of the Model-Based classifiers *area* and *peak picking* as (among others) employed by Farwell and Donchin (1988) and previously discussed in section 4.1.1 will be assessed.

Methods

Classification in order to infer a symbol was performed as discussed in more detail in section 5.4: From an EEG time series $\mathbf{x}_{ik}$, reflecting an epoch belonging to the flashing of a row or column i within the subtrial k, the score $s^{\Psi}(\mathbf{x}_{ik})$ was calculated to infer a symbol. For this section, one of the Model-Based classifiers *area* and *peak picking*, i.e., $\Psi \in \{Area, PP\}$, was chosen. The parameters for both classifiers were the boundaries t_1 and t_2 of the P300 window, which need to be calculated from the ERP from a number of trials, as can be seen in equations (6.2) and (6.3). As can also be observed in Figure 4.6, *area* calculates the surface under the curve within the P300 window, and *peak picking* measures the difference between the maximum within and the minimum prior to the P300 window:

$$\begin{aligned} s^{\text{Area}}(\mathbf{x}_{ik}) &= \sum_{t=t_1}^{t_2} x_{ik}(t), && (6.2) \\ s^{\text{PP}}(\mathbf{x}_{ik}) &= \min_{t<t1} x_{ik}(t) - \max_{t>t1} x_{ik}(t). && (6.3) \end{aligned}$$

It is usually necessary to employ more than just one subtrial for a symbol inference such that the score

$$S_i^{\Psi}(\mathbf{x}) = \sum_{k=1}^{n_{\text{combined}}} s^{\Psi}(\mathbf{x}_{ik}), \tag{6.4}$$

aggregating the number of n_{combined} subtrials is computed to infer the correct symbol by choosing that row of index $i_r \in \mathcal{I}_{\text{rows}} = \{r_1, r_2, r_3, r_4, r_5, r_6\}$ and that column with index $i_c \in \mathcal{I}_{\text{columns}} = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ which is assigned with the maximum score $S_i^{\Psi}(\mathbf{x})$ among the rows and columns, respectively (cf. section 5.4):

$$i_r = \arg \max_{i \in \mathcal{I}_{\text{rows}}} S_i^{\Psi}(\mathbf{x}) \quad \text{and} \quad i_c = \arg \max_{i \in \mathcal{I}_{\text{columns}}} S_i^{\Psi}(\mathbf{x}). \tag{6.5}$$

The number of aggregated subtrials n_{combined} was systematically varied and the number of correct symbol predictions, divided by the number of all possible predictions $n_{\text{inferences}}$ for the specific n_{combined} has yielded the accuracy

$$p_{\text{acc}} = n_{\text{correct}} / n_{\text{inferences}}. \tag{6.6}$$

Table 6.2 gives an example of performed symbol predictions utilizing different numbers n_{combined} of aggregated subtrials. The predictions were performed with the *area* classifier for all collected subtrials belonging to the symbol “S”on data from subject 1A. Consult section 5.4 for more details about symbol inference.

As explained in sections 3.3 and 4.2, in order to assess the speed of the BCI device and to be able to compare the results with other approaches, the more general measure *information transfer rate* in bits/min, taking also the time for spelling a symbol into

Table 6.2: Calculation of accuracy rates in symbol inference for the symbol "S". From $n_{\text{subPerSymbol}} = 42$ subtrials belonging to one symbol, n_{combined} subtrials are aggregated, and only $n_{\text{inferences}}$ can be inferred. The number of correctly inferred symbols n_{correct} divided by $n_{\text{inferences}}$ yields the accuracy p_{acc}. Correctly inferred symbols are underlined.

$n_{\textbf{combined}}$	**Inferred Symbols**	$n_{\textbf{correct}}$	$n_{\textbf{inferences}}$	$p_{\textbf{acc}}$
1	$\texttt{MKK}\underline{\texttt{SS}}\texttt{4YYG79X}\underline{\texttt{S}}\texttt{6U}\underline{\texttt{S}}\texttt{XCT}\underline{\texttt{SS}}\texttt{TXUFNGGAVO}\underline{\texttt{S}}\texttt{ZW8VXB}\underline{\texttt{SS}}\texttt{X8}$	9	42	0.214
2	$\underline{\texttt{SS}}\texttt{4YYXYU}\underline{\texttt{SSSS}}\texttt{CG}\underline{\texttt{S}}\texttt{4UV3}\underline{\texttt{S}}\texttt{3}$	8	21	0.381
3	$\texttt{G}\underline{\texttt{S}}\texttt{Y9U}\underline{\texttt{SSS}}\texttt{CANW}\underline{\texttt{SS}}$	6	14	0.429
4	$\underline{\texttt{SS}}\texttt{4}\underline{\texttt{SSS}}\texttt{A}\underline{\texttt{S}}\texttt{W}\underline{\texttt{S}}$	7	10	0.70

account, can be computed as

$$B(36, p_{\text{acc}}, t_{\text{ISI}}, n_{\text{combined}}) = \frac{60 \cdot \left(\log_2 36 + p_{\text{acc}} \log_2 p_{\text{acc}} + (1 - p_{\text{acc}}) \log_2 \frac{1-p_{\text{acc}}}{35}\right)}{n_{\text{combined}} \cdot 12 \cdot t_{\text{ISI}}} \tag{6.7}$$

for n_{combined} subtrials and an Interstimulus Interval of t_{ISI}. For the current investigation, the data were split into two halfs, each containing half of the symbols. On the first half, the parameters t_1 and t_2 for the classifiers were determined (see below). Afterwards, the symbols for the second half were computed. This procedure was then repeated on switched sets, i.e., the parameters were determined on the second set and then applied for classifying the first set.

Results

As depicted in Figure 6.8, for **subject 1A**, the mean classification accuracy reached 80% ($p_{\text{acc}} = 0.800$) with 12 subtrials and the *area* classifier, while a limit of 90% accuracy was achieved after 14 subtrials ($p_{\text{acc}} = 0.933$). Several authors regard reaching 80% or 90% classification accuracy as a limit for a practical usage of a BCI (Farwell and Donchin, 1988; Serby et al., 2005). Thus, in the following, the times for exceeding these criteria will be crucial measures for the performance of the BCIs. In this experiment, each subtrial lasted 6 seconds. Therefore, durations of 72s and 84s, respectively, would be required to satisfy the criteria. Peak picking reached the 80% criterion after just 8 subtrials or 48s ($p_{\text{acc}} = 0.840$) and 90% accuracy with 13 subtrials or within 78s ($p_{\text{acc}} = 0.900$). The maximum information transfer rate (cf. section 3.3) in subject 1A was 4.10 bits/min using the area classifier and 4.86 bits/min under the *peak picking* classification method. When reaching 80% accuracy first, information transfer rates of 2.99 bits/min and 3.51 bits/min, respectively, were achieved for the *area* and *peak picking* classifiers.

In contrast, *area* and *peak picking* as well did both not exceed 30% accuracy for **subject 1B**, and information transfer rates also stayed below 1 bit/min in any case.

Discussion

For both classifiers, strong interindividual differences were found. Although *area* and *peak picking* both reached a criterion of 80% accuracy after a certain number of repetitions for subject 1A, they failed to produce reasonable results for subject 1B, such that it would not be practicable to use the methods for subject 1B. While in this experiment, for subject 1A, in the average subtrials of a length of 48s for the *area* classifier

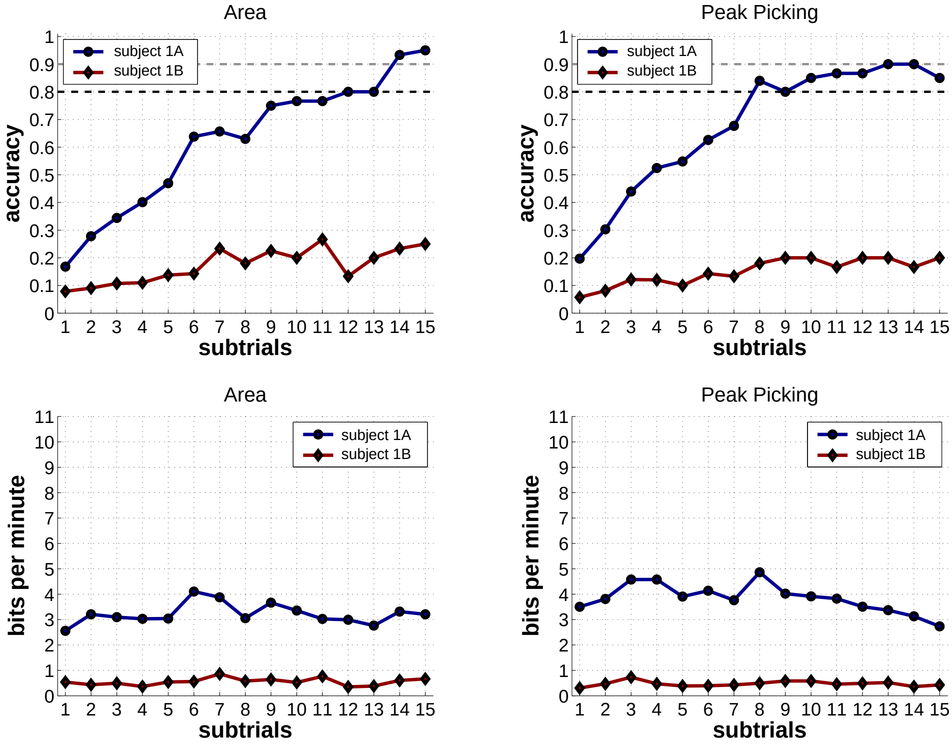


Figure 6.8: Classification accuracies (top row) and information transfer rates in bits/min (bottom row) for subject 1A (blue line) and subject 1B (red line). Results were computed using the Model-Based classifiers *area* (left) and *peak picking* (right) employing data from the Pz electrode. While the classifiers work comparably well for subject 1A, they only expose a weak performance for subject B. A subtrial denotes a sequence of flashing all 6 rows and 6 columns once and lasted 6 seconds in this experiment.

and 72s for the *peak picking* classifier were needed to reach 80% accuracy, Farwell and Donchin (1988) achieved better results with these techniques, ranging from 12.6 s to 56.6 s (area) and 17.3 s to 28.2 s (peak picking) to reach the 80% accuracy criterion (see section 4.1.1).

6.2.4 Binary Classification using Machine-Learning Classifiers

After the Model-Based classifiers yielded only unsatisfying results, the performance of Machine-Learning classifiers will be investigated.

In a first step, the three classification techniques Fisher's Linear Discriminant (FLDA), Linear Support Vector Machine (LSVM) and Support Vector Machine with Gaussian kernel (RBF SVM) will be compared (see section 5.3). In contrast to *area* and *peak picking*, which both rely on *averaged* trials, these approaches are used to classify *single* trials. Only in a second step, classifications of a number of single trials will be aggregated in order to infer the symbols (see section 5.4). Relying on single trials makes it possible to first compare the performance of the classifiers for *binary* classifications distinguishing between $\mathcal{P}^+$ and $\mathcal{P}^-$ samples in order to receive preliminary information about how well they will work for symbol inferences. This procedure is computationally much cheaper and therefore allows for a larger number of comparisons.

Beside investigating effects of the different classifiers on classification accuracies, it will be assessed how well the classifiers work with dimensionality reduced data using PCA. As detailed out in section 5.2.2, due to the *curse of dimensionality* and computational costs, it is advisable to find a data representation which follows the intrinsic dimensionality of the data. Furthermore, with a lower dimensionality, singularity of the data matrix becomes more improbable, making it possible to apply FLDA using simple matrix inversions (see section 5.3.3). While it is not possible to employ PCA for the Model-Based methods *area* and *peak picking*, which both depend upon the specific temporal structure of the data, Machine-Learning techniques can easily be adapted to a given data structure and do not take into account the temporal relationships here. PCA dimensionality reduced feature vectors were used for classification, such that data projections onto 2, 8, 9, and 11 Principal Components will be examined. The previous PCA analyses revealed that 8 Principal Components reflect 90% variance of the data from subject 1A, and 9 Principal Components 90% variance of subject 1B's data. For both subjects, 11 Principal Components capture 99.9% variance (see section 6.2.2).

Methods

Classification performance was assessed using a balanced set of 834 $\mathcal{P}^+$ and 834 $\mathcal{P}^-$ samples of the data. This set was further divided into two balanced halfs of 417 $\mathcal{P}^+$ and 417 $\mathcal{P}^-$ randomly selected samples. One half served as the training set, and the other half as the test set (see section 5.3.1). The Machine-Learning classifiers were trained on the training set, and their performance in correctly classifying the unseen data of the test set was evaluated afterwards. Training consisted of two steps: In a first step, the hyperparameters of the classifiers were varied as described in section 6.1 and summarized in Table 6.1, and the classification performance was assessed in a 5-fold crossvalidation scheme (see section 5.3.4): The training data were divided into 5 sets, and 4 sets were used for training while the remaining set served as a test set; this procedure was then repeated for each possible combination of these sets, and the mean classification performance on the 5 test sets was computed. In a second step, those hyperparameters which yielded the best mean classification performances were selected and the whole training data were used for adapting the Machine-Learning classifiers with this parameter configuration. Then, the test data were classified and the classification accuracy, i.e., the number of correctly inferred labels divided by the number of all samples, was calculated.

Results

Table 6.3 exposes the classification results as achieved for the two subjects by employing the different numbers of Principal Components and different classifiers. The classification results obtained with feature vectors of 8, 9, 11, and 160 dimensions ranged from 0.693 to 0.707 for subject 1A, and from 0.646 to 0.668 for subject 1B. Employing 2-dimensional feature vectors yielded accuracies of only 0.621 to 0.636 for subject 1A, and 0.585 to 0.612 for subject 1B. In an overall comparison, the RBF SVM classifier with 11 or 160 dimensions exposed the best classification performance.

Discussion

Three important outcomes can be observed in the results. First, the 160-dimensional data vector can be reduced to 11 (and even less) dimensions without loss in classifi-

Table 6.3: Classification accuracies in binary classification using the Machine-Learning classifiers FLDA, LSVM, and RBF SVM for the subjects 1A and 1B. Beside classifying data from the *Pz* electrode with the original 160 dimensions, also feature vectors with reduced dimensionality were analyzed. By employing PCA, such feature vectors of 2, 8, 9, and 11 dimensions were calculated. Thereby, 8 and 9 dimensions reflect 90% of captured variance for subject 1A and subject 1B, respectively. For both subjects, 11 dimensions capture 99.9% variance of the data. With 160 dimensions, the data matrix became singular and FLDA could not be performed without further adaptations.

		Principal Components				**Original Space**
Subject	**Classifier**	2	8	9	11	160
1A	FLDA	0.621	0.700	0.698	0.689	-
	Linear SVM	0.624	0.693	0.707	0.687	0.687
	RBF SVM	0.636	0.698	0.698	0.695	0.696
1B	FLDA	0.595	0.648	0.646	0.646	-
	Linear SVM	0.612	0.649	0.651	0.651	0.651
	RBF SVM	0.585	0.651	0.663	0.668	0.668

cation accuracies. This finding indicates that PCA has qualified as a good technique for dimensionality reduction in this context. Second, when employing 11-dimensional feature vectors, it is possible to use the computational less expensive FLDA, and this classifier exposes almost the same performance as the state-of-the-art Support Vector Machines. Furthermore, RBF SVM did not necessarily perform better than the other methods, such that the regularity behind the data was well reflected by linear techniques. Third, in contrast to the Model-Based classifiers, only small differences in classification accuracy between the subjects can be observed. Apparently, the ability of the Machine-Learning classifiers to adapt to signals reduced the interindividual differences. This first rough investigation of classification performances helps to design further examinations but does nevertheless not allow to draw far reaching assertions. Thus, to allow for more conclusions and to provide a fair comparison with the Model-Based techniques, computing symbol inferences is necessary.

6.2.5 Symbol Inferences using Machine-Learning Classifiers

Binary classification revealed that the strong differences in classification performance between the subjects as observed for the Model-Based classifiers almost vanished when using Machine-Learning classifiers. Furthermore, by reducing the dimensionality of the data to only 11 dimensions by using Principal Component Analysis, no loss in performance occurred. In this section, symbols the subjects directed attention to are to be inferred by the Machine-Learning classifiers.

Methods

The data were divided into two halfs, each containing 50% of the symbols. Training of the Machine-Learning classifiers was performed as in binary classification: The first half served as a training set, within which a 5-fold crossvalidation was performed on a balanced set in order to find the suitable hyperparameters. Then, the classifiers were trained on the whole data from this set, and the data of the second half were classified to compute the correct symbols as described in 5.4 for different numbers of subtrial combinations. Afterwards, this whole procedure was also performed on switched sets, i.e., the first half served as a test set and the second half as the training set.

As for binary classification, the trained classifier as well as the PCA matrices were computed on the training sets. The whole dimensionality of 160 was employed for LSVM and RBF SVM classifications. Using data projections onto the first 11 Principal Components, beside LSVM and RBF SVM also FLDA was used for data analysis, which became possible without further modifications for this dimensionality (cf. section 5.3.3).

Results

As it is depicted in Figure 6.9, for **subject 1A**, employing the whole **dimensionality of 160**, a classification accuracy of 80% was exceeded when employing 8 subtrials[3] and the Linear SVM ($p_{\text{acc}} = 0.815$), and after 9 subtrials using the RBF SVM ($p_{\text{acc}} = 0.825$). However, the classifier's performance temporarily sank below 80% for 10 subtrials (LSVM: $p_{\text{acc}} = 0.767$, RBF SVM: $p_{\text{acc}} = 0.792$). Both classifiers exceeded 90% classification accuracy after 12 subtrials. Maximum information transfer rates were 5.08 bits/min and 5.10 bits/min for LSVM and RBF SVM, after 4 and 5 subtrials, respectively. Under the precondition of reaching 80% accuracy, transfer rates of 4.63 bits/min and 4.12 bits/min were obtained for LSVM and RBF SVM and for exceeding 90% accuracy, 3.65 bits/min were achieved with both classifiers. For **subject 1B**, 80% accuracy was achieved with 12 subtrials for both classifiers (LSVM: $p_{\text{acc}} = 0.800$, RBF SVM: $p_{\text{acc}} = 0.830$). Again, LSVM experienced a short fallback below 80% at 14 subtrials ($p_{\text{acc}} = 0.767$). The classifiers failed to reach 90% accuracy for subject 1B within 15 subtrials. Best Information transfer rates were 3.91 bits/min and 4.19 bits/min after 4 and 3 subtrials for the classifiers LSVM and RBF SVM.

When projecting the data onto **11 Principal Components**, it became possible to directly employ FLDA. For **subject 1A**, the different classifiers FLDA, LSVM, and RBF SVM each reached 80% accuracy after 8 subtrials, but RBF SVM exposed superior performance with $p_{\text{acc}} = 0.875$ in contrast to $p_{\text{acc}} = 0.815$ for FLDA and $p_{\text{acc}} = 0.855$ for LSVM (see Figure 6.10). An accuracy of 90% was reached after 12 subtrials for each classifier ($p_{\text{acc}} = 0.90$ each). The best information transfer rates were 5.24 bits/min (FLDA, 4 subtrials), 5.33 bits/min (LSVM, 6 subtrials), and 6.33 bits/min (RBF SVM, 4 subtrials). Considering a criterion of mandatory 80% accuracy, 4.63 bits/min, 5.07 bits/min and 5.23 bits/min were achieved with the classifiers FLDA, LSVM, and RBF SVM, respectively. For **subject 1B**, the best information transfer rate was 4.49 bits/min (LSVM, 4 subtrials). All classifiers reached the 80% criterion after 13 subtrials (FLDA: $p_{\text{acc}} = 0.867$, LSVM: $p_{\text{acc}} = 0.800$, RBF SVM: $p_{\text{acc}} = 0.800$) with information transfer rates of 3.17 bits/min, 2.97 bits/min and 3.04 bits/min, respectively, for the three classifiers. They failed to reach 90% accuracy within 15 subtrials. In contrast to analyses with the 160-dimensional data space, no fallback occurred in these cases.

Discussion

The Machine-Learning classifiers outperformed the Model-Based classifiers in two ways. First, the classification accuracies of the Machine-Learning approaches were superior compared to the Model-Based approaches. Second, the lack in performance for subject 1B with Model-Based methods disappeared when using Machine-Learning techniques.

The ability of Machine-Learning techniques to adapt to the data, rather than obeying

[3] In this experiment, one subtrial lasted 6 seconds.

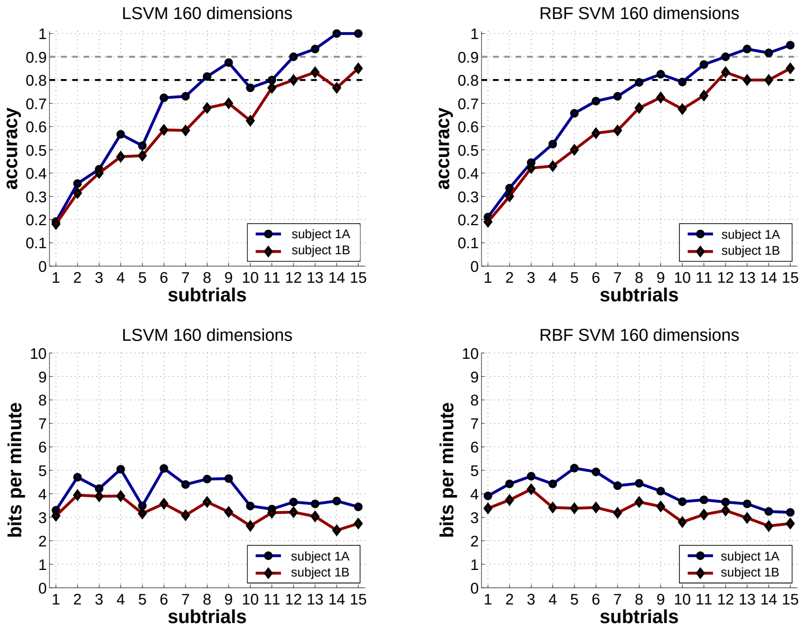


Figure 6.9: Classification accuracies (top row) and information transfer rates (bottom row) for subject 1A (blue graph) and subject 1B (red graph) using the classifiers Linear SVM (left), and RBF SVM (right) on data of 160-dimensional feature vectors representing data recorded from the Pz electrode.

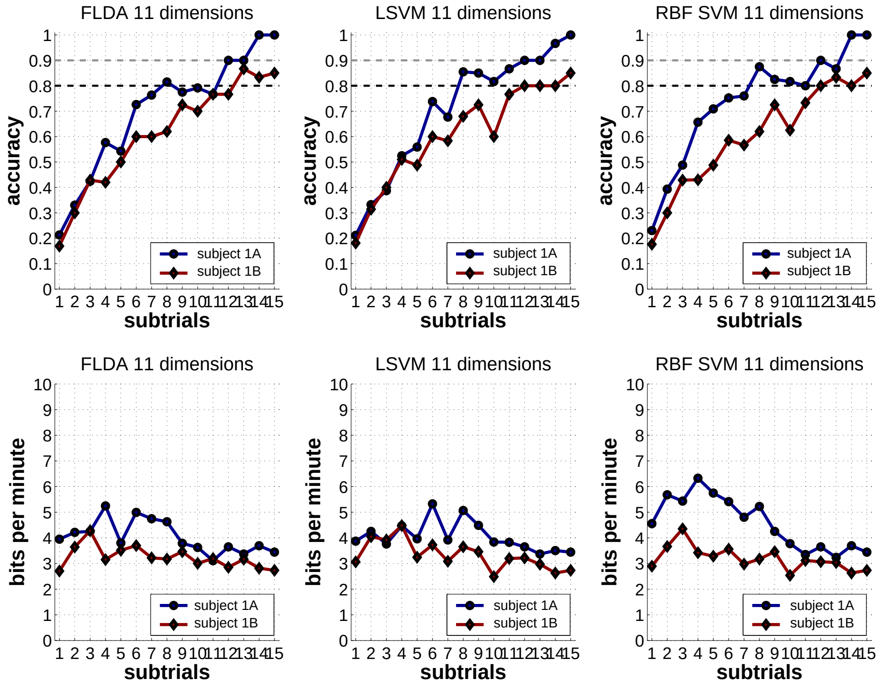


Figure 6.10: Classification accuracies (top row) and information transfer rates (bottom row) for subject 1A (blue graph) and subject 1B (red graph) using the classifiers Fisher's Linear Discriminant (left), Linear SVM (middle), and RBF SVM (right) on data from the Pz electrode projected onto 11 Principal Components.

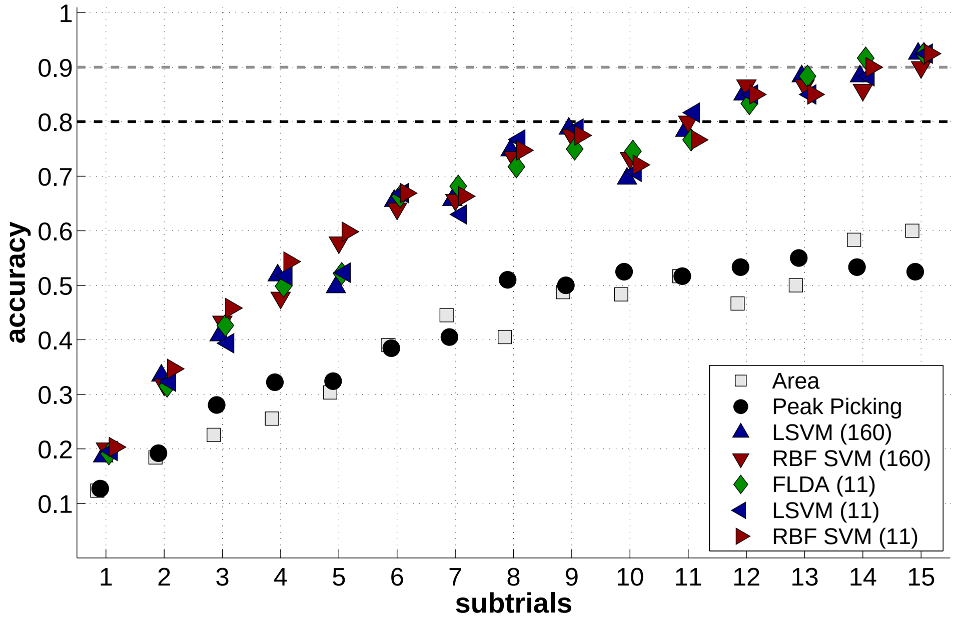


Figure 6.11: Mean classification accuracies among both subjects for symbol inferences of the different classifiers employing either 160-dimensional or PCA-reduced 11-dimensional feature vectors representing data from the Pz electrode. In this experiment, a subtrial (a sequence of 6 row and 6 column flashes) lasted 6 seconds.

certain assumptions about the data, led to much better results than achieved with the Model-Based techniques *area* and *peak picking*. As already revealed by Figure 6.4, the data from the two subjects differ substantially, and the data from subject 1B expose ERP waves which do not fit exactly into common assumptions about a P300, as they are employed in the classification techniques *area* and *peak picking*. Particularly for such subjects, the classification accuracies can substantially be improved by using Machine-Learning techniques.

Performing Principal Component Analysis to reduce the dimensionality of the feature vectors by projecting the data onto 11 dimensions, capturing more than 99.9% of the variance, yielded almost the same performances as employing the original data with 160 dimensions (see Figure 6.11). Thus, as in binary classification, PCA qualified as a good choice for dimensionality reduction in this context. Using the dimensionality reduced feature vectors, classification with Fisher's Linear Discriminant Analysis became possible. Although this technique is much simpler (and computationally less expensive) than Support Vector Machines, comparable classification results were achieved with this techniques.

With the increases in classification accuracy using Machine-Learning classifiers, the information transfer rates were also improving. But with a maximum transfer rate of 6.33 bits/min they stayed nevertheless unsatisfying.

6.2.6 Conclusion

In a first step, preprocessing parameters were derived from the ERPs and the power spectra, suggesting epoch lengths of 800ms and band pass filtering of 0.5-8Hz. These parameters were subsequently used for preprocessing.

When calculating Principal Components, it turned out that a dimensionality reduction from 160 to 11 dimensions was possible by retaining more than 99.9% of the data variance. The according Principal Components reflect different frequency and phase information from 1Hz to 6Hz.

Ambiguous results were found for the Model-Based classifiers *area* and *peak picking.* While they performed comparably well for subject 1A, they failed to produce viable results for subject 1B. In contrast, applying the Machine-Learning classifiers Linear and RBF SVM has yielded reasonable classification performances for both subjects. In the mean, the Machine-Learning classifiers outperformed the Model-Based classifiers. Thereby, the non-linear classifier RBF SVM produced almost the same results as the linear classifiers, making it likely that the classification problem is already well represented by linear techniques.

Binary classifications with Machine-Learning classifiers of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples as well as symbol inference computations have shown that projecting the data onto the first 11 Principal Components revealed almost the same classification rates compared to using the original 160-dimensional data. Therefore, PCA qualified as an appropriate technique for dimensionality reduction in this context. Since RBF SVM as a non-linear method yielded comparable results as the linear methods, non-linear techniques for dimensionality reduction (like a kernel PCA) need not to be considered.

With dimensionality reduced feature vectors it became possible to directly perform classifications with FLDA, since the within-class scatter matrices did no longer became singular (cf. section 5.3.3). The classification performance for FLDA was almost the same as for the Support Vector Machines. Furthermore, the dimensionality of the feature vectors (11 or 160) did not affect the classification results. A summary of the classification accuracies obtained with the different classification strategies is depicted in Figure 6.11.

Although the Machine-Learning classifiers enhanced the classification performance, the classification accuracies remained low, and in the mean, about 8 subtrials, i.e., 92 seconds would be necessary to spell a symbol with an accuracy of 80%, which is unsatisfying slow when intending to use this device for communication purposes in an online version. Thus, the following section aims to increase the classification accuracies by employing data from more electrodes.

6.3 Improving Classification Accuracies using Data from Multiple Electrodes

As a first possibility to further increase the performance of the Machine-Learning classifiers and therewith the information transfer rates, taking data from more than just one electrode for classifications can be considered. The prior analyses were based on data from the Pz electrode because the P300 is known to be most prominent at this site, and the original work of Farwell and Donchin (1988) and Donchin et al. (2000)

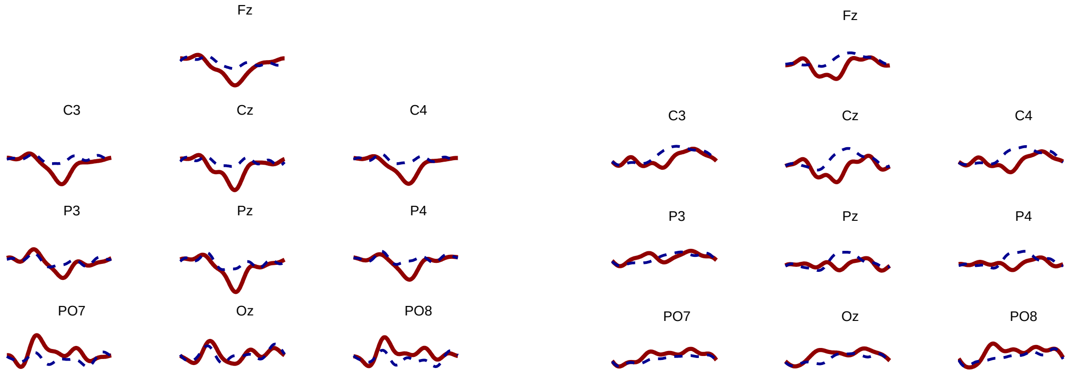


Figure 6.12: Scalp distributions of ERPs from $\mathcal{P}^+$ (red solid line) and $\mathcal{P}^-$ (blue dashed line) events for subject 1A (left) and subject 1B (right).

relied on data from this electrode. However, when examining data from other scalp positions than Pz, differences between $\mathcal{P}^+$ and $\mathcal{P}^-$ epochs can also be identified in the ERP. Most of these differences are similar to those observed at Pz, but some of them, especially at the sites $PO7, PO8$ and Oz also expose another structure of differences (see Figure 6.12). This finding encourages to employ data from more than just one electrode location. If oddball events (cf. section 6.2), i.e., flashings of the row or column containing the symbol to attend to, also induce different activations compared to non-oddball events at further sites than Pz, this might enhance the signal-to-noise ratio and result in better classification performances when considering more electrode sites.

This section aims to examine this suspicion based on data from the experiment of the previous section. The experiments is the same, but instead of using only data from the Pz electrode, also data from the further locations are to be considered. As a first step, Principal Components will be analyzed. It will be investigated to which degree they are still capable to decrease the dimensionality of the feature vectors with the extended electrode space. In contrast to the previous section, it would be necessary to extend the model assumptions of the Model-Based classifiers *area* and *peak picking* to further locations than Pz. Since the Model-Based techniques already lacked performance in the previous section, they will not be considered in the following. Instead, Machine-Learning techniques can easily be adapted to a new data structure, and binary classification as well as symbol inference will solely be performed with Machine-Learning classifiers in the remaining sections.

6.3.1 Principal Component Analysis

Methods

Principal Component Analysis was performed in the same way as in the previous section, but the feature vectors were 1600-dimensional, reflecting the 10 scalp sites $Fz, Cz, Pz, Oz, C3, C4, P3, P4, PO7$, and $PO8$ as well as the 160 sampling points for time series of 800ms data. Again, PCA was calculated on a balanced set of 834 $\mathcal{P}^+$ and 834 $\mathcal{P}^-$ samples. Since the new data vector incorporates data from 10 electrodes simultaneously, a Principal Component reflects temporal *and* spatial patterns as well. Figure 6.13 (left) gives an example of such a spatio-temporal Principal Component. The

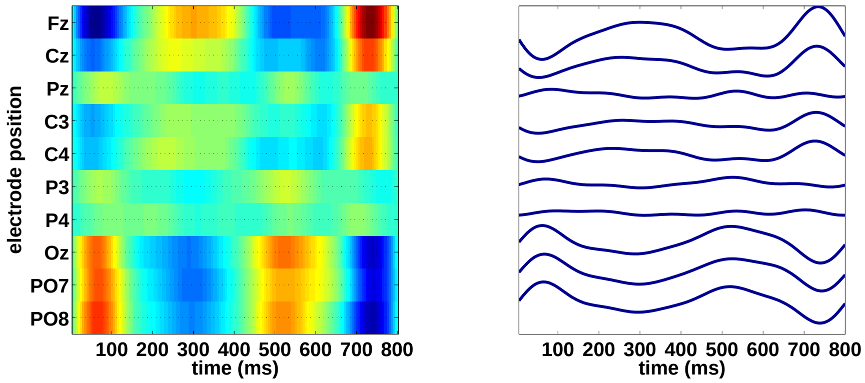


Figure 6.13: Example of a Principal Component taken from subject 1A (Principal Component 14, cf. Figure 6.15). **Left:** Electrode positions are vertically aligned, while time is drawn on the x-axis. Color shading indicates intensities, comparable to amplitude values. Clear differences between the last three electrode sites $PO7, PO8, Oz$ and the other sites can be observed in their temporal structure. **Right:** Intensity plots of the component reflecting the electrode sites which were color encoded in the left figure.

1600-dimensional data vector is split into 10 rows, each reflecting an electrode position, and into 160 columns, containing the 800ms points in time. Thereby, color shading indicates intensities as depicted in 6.13 (right).

Results

The distribution of the cumulative Eigenvalues for both subjects is depicted in Figure 6.14. For subject 1A, 99.9% of the data variance are captured within the first 84 Principal Components, while 99% and 90% are incorporated in 56 and 17 components, respectively. For subject 1B, 90%, 99% and 99.9% are captured within the first 26, 71, and 95 components, respectively.

The first 28 Principal Components, including the first 17 and 26 components which reflect 90% variance in subject 1A and 1B, are depicted in Figures 6.15 and 6.16, respectively.

Discussion

Strong dimensionality reductions from 1600 to less than 100 dimensions by retaining 99.9% of the data variance were achieved for both subjects by Principal Component Analysis. The Principal Components reflect spatio-temporal activation patterns. A first investigation of the components revealed that predominantly different frequencies were reflected within the different components. While these frequency patterns were distributed over all electrodes in almost the same way up to PCA 7 in subject 1A and PCA 11 in subject 1B, differentiations between electrodes occurred for higher Principal Components as can be observed in Figures 6.15 and 6.16. Particularly the occipital and parieto-occipital sites $Oz, PO7$ and $PO8$ often form a cluster of different patterns compared to the other sites (e.g., PCA 14 for subject 1A and PCA 12 for subject 1B). This finding corresponds to the different activations at these sites already found in the ERP (cf. Figure 6.12).

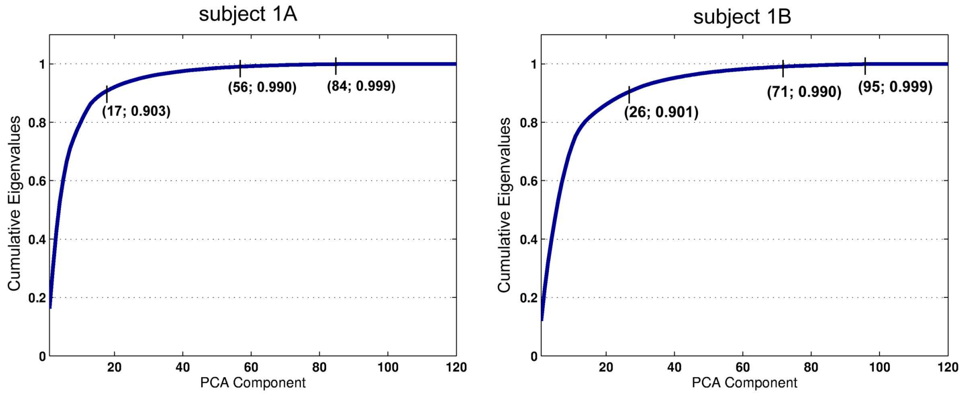


Figure 6.14: Cumulative Eigenvalues of Principal Components from the data of subject 1A (left) and subject 1B (right) using 1600-dimensional feature vectors representing time series of 10 electrodes.

6.3.2 Binary Classification with Principal Components as Feature Vectors

In order to estimate the classification accuracies with data projections onto different numbers of Principal Components, the computationally rather inexpensive binary classification is performed.

Methods

Classification accuracies were assessed as in section 6.2.4: The 834 $\mathcal{P}^+$ and 834 $\mathcal{P}^-$ samples data were further divided into two balanced halfs of randomly chosen 417 $\mathcal{P}^+$ and 417 $\mathcal{P}^-$ samples each. A 5-fold cross-validation was then performed on one half to find suitable hyperparameters, which were then employed for training the data on this half. Afterwards, classification of the other half was performed.

Results

Table 6.4 contains the classification results as obtained for the different numbers of Principal Components and different classifiers for subject 1A, while Table 6.5 shows this information for subject 1B.

For subject 1A, classification accuracies were increasing with the number of Principal Components for each classifier, but only little differences were found between 84- and 1600-dimensional feature vectors. With 84 dimensions in the feature vector, FLDA yielded the same accuracy rate as LSVM.

For subject 1B, the general trend that employing more Principal Components results in better accuracies was also present with the exception that employing 95 Principal Components produced slight better classification results than using the full space with 1600 dimensions. Furthermore, with FLDA and 71 Principal Components, better results could be obtained than with 95 Principal Components.

Compared to section 6.2.4, the best classification results could be improved from 0.707 (9 Principal Components, LSVM) and 0.668 (11 Principal Components, RBF SVM) for subject 1A and 1B, respectively, to 0.886 (1600 dimension, RBF SVM) and 0.845 (71 Principal Components, FLDA).

The number of Principal Components affected the classification accuracies such that better accuracies were achieved when using more Principal Components. However,

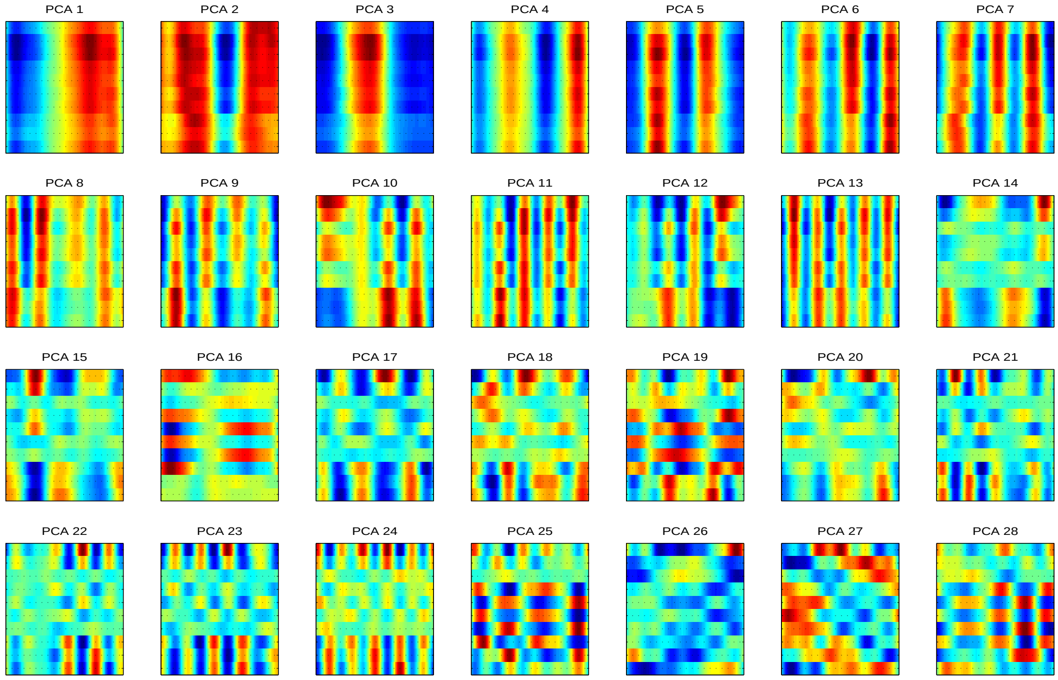


Figure 6.15: Subject 1A: First 28 Principal Components. While the first components mainly reflect certain frequencies for all electrode sites, varying patterns can be observed for different sites with higher components (>PCA 13). Figure 6.13 gives an example about the structure of the activation patterns.

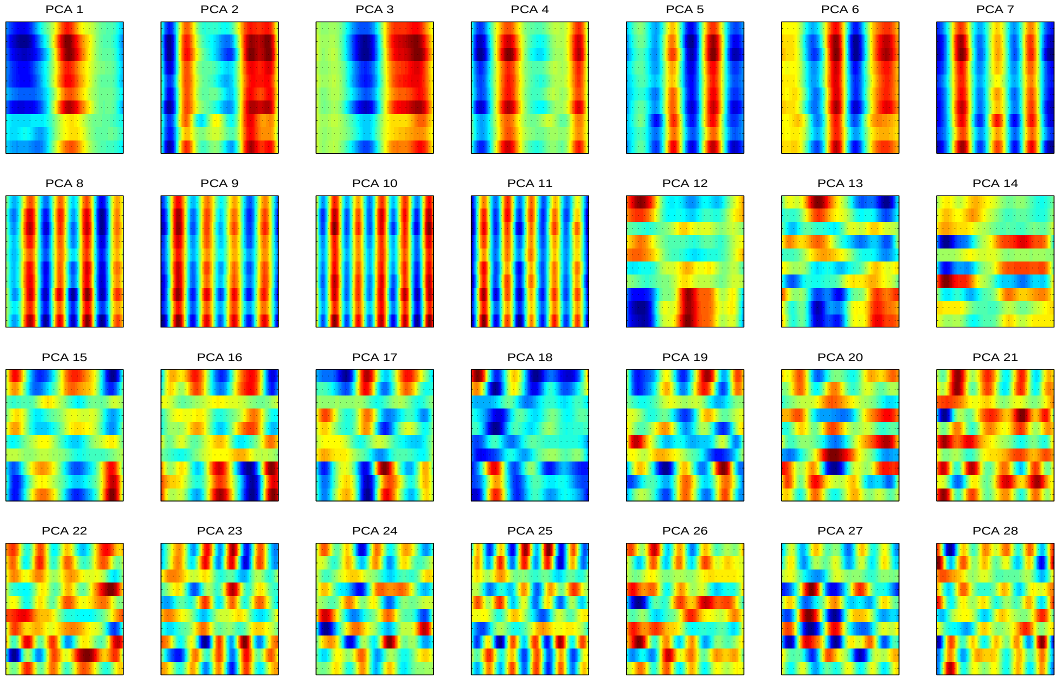


Figure 6.16: Subject 1B: First 28 Principal Components. Again, first components reflect same frequency patterns for the different electrode sites, while they spread in higher components (>PCA 11).

Table 6.4: Classification accuracies in binary classification based on data from 10 electrodes using the Machine-Learning classifiers FLDA, LSVM, and RBF SVM for subject 1A. The full data space of 1600 dimensions as well as dimensionality reduced feature vectors were employed for classification. In this subject, projections on the first 17, 56, and 84 Principal Components reflect 90%, 99% and 99.9% of the data variance, respectively. With 1600 dimensions, the data matrix became singular and FLDA could not be performed without further adaptations.

	Principal Components					**Full Space**
Classifier	2	10	17	56	84	1600
FLDA	0.645	0.753	0.825	0.873	0.879	-
Linear SVM	0.655	0.764	0.812	0.878	0.879	0.884
RBF SVM	0.627	0.736	0.814	0.877	0.883	0.886

Table 6.5: Classification accuracies in binary classification for subject 1B. In this subject, projections on the first 26, 71, and 95 Principal Components reflect 90%, 99% and 99.9% of the data variance, respectively.

	Principal Components					**Full Space**
Classifier	2	10	26	71	95	1600
FLDA	0.651	0.671	0.800	0.845	0.839	-
Linear SVM	0.638	0.669	0.806	0.831	0.844	0.830
RBF SVM	0.649	0.667	0.808	0.830	0.844	0.841

only slight differences in classification accuracy were obtained for the feature vectors employing the full 1600-dimensional space or with 99.9% PCA-reduced features. Again, all Machine-Learning classifiers yielded results within the same accuracy range.

Discussion

Employing data from all 10 electrodes has yielded much better classification accuracies compared to using data from the Pz electrode alone (cf. section 6.2.4). The performance of the classifiers was in the same range. The RBF SVM did not necessarily produce better results, indicating that the classification problem is already well reflected by linear techniques. Classification accuracies were barely worsen when using 99.9% of the information represented in the Principal Components compared to using the full space of 1600 dimensions. Thus, dimensionality reduction as performed by PCA has found an adequate representation of the data structure. For subject 1B, classification accuracies even rised when employing the first 95 Principal Components. This phenomenon might be caused by reduction of noise contained in the dropped 0.1% of the data variance. Nevertheless, for less than 95 Principal Components, the general trend was that employing fewer Principal Components also resulted in worse classification accuracies.

6.3.3 Symbol Inferences

As performed in section 6.2.3, the Machine-Learning classifiers are employed to infer symbols from the data in the following.

Table 6.6: Calculation of accuracy rates in symbol inference for the symbol “S”in subject 1A using the LSVM classifier and the 1600-dimensional feature vector. From $n_{\text{subPerSymbol}} = 42$ subtrials, which is the number of subtrials belonging to one symbol, n_{combined} subtrials are aggregated, such that $n_{\text{inferences}}$ inferences can be computed for this number of aggregations. Correct inferences are underlined and symbols were inferred on the same data source as the results exposed in Table 6.2.3.

$n_{\textbf{combined}}$	**Inferred Symbols**	$n_{\textbf{correct}}$	$n_{\textbf{inferences}}$	$p_{\textbf{acc}}$
1	SGSSTSYYSSSSSSM7SSSSSSSGSAXSSSMASGSUSYWSSS	27	42	0.643
2	SSSYSSSMSSSSSSSSSSSSS	19	21	0.905
3	SSSSSSSSSSSSSS	14	14	1.000
4	SSSSSSSSSS	10	10	1.000

Methods

Data were split into two halfs, each half containing 50% of the symbols. One half served as the training set, the other half as the test set and vice versa. Training was performed as in binary classification: Balanced sets were constructed on the training set and a 5-fold cross-validation was performed to find suitable hyperparameters for the classifiers which were then trained on the whole training set and applied to the test set in order to infer the symbols. Symbol inference was then performed as discussed in detail more in sections 6.2.5, 6.2.3 and 5.4.

This procedure was performed using the original 1600-dimensional feature vectors for LSVM and RBF SVM as well. On the other hand, PCA dimensionality reduced feature vectors capturing 99.9% variance, resulting in 84 dimensions for subject 1A, and 95 dimensions for subject 1B, were a further basis for classifications with FLDA, LSVM, and RBF SVM.

Results

For a direct comparison, Table 6.6 contains symbol inferences from the same data source as Table 6.2 with the difference that the former results were calculated using 10 electrodes and the LSVM classifier, while the latter ones were calculated using the *area* classifier on the Pz electrode alone. Constant perfect accuracy was reached after 13 subtrials using the *area* classifier on the Pz electrode, and after just 3 subtrials with 10 electrodes and the LSVM classifier.

Figure 6.17 depicts classification accuracies and information transfer rates for subject 1A and subject 1B as calculated with RBF SVM and LSVM on **1600-dimensional data** from the 10 electrodes. For **subject 1A**, both classifiers reached 80% as well as 90% classification accuracy with data from 2 subtrials (LSVM: $p_{\text{acc}} = 0.927$, RBF SVM: $p_{\text{acc}} = 0.918$). A maximum information transfer rate of 31.54 bits/min (1 subtrial) was achieved using LSVM, and of 29.82 bits/min with the RBF SVM. However, with the constraint to reach at least 80% and 90% accuracy, transfer rates decrease to 22.25 bits/min and 21.78 bits/min, respectively.

For **subject 1B**, LSVM achieved an accuracy of $p_{\text{acc}} = 0.814$ after 2 subtrials, and exceeded 90% accuracy after 3 subtrials ($p_{\text{acc}} = 0.936$). RBF SVM exceeded 80% (and also 90%) accuracy after 3 subtrials ($p_{\text{acc}} = 0.943$). The best information transfer rates for subject 1B were 22.56 bits/min and 22.28 bits/min for LSVM and RBF SVM, respectively. Reaching the 80% criterion first yielded corresponding transfer rates of

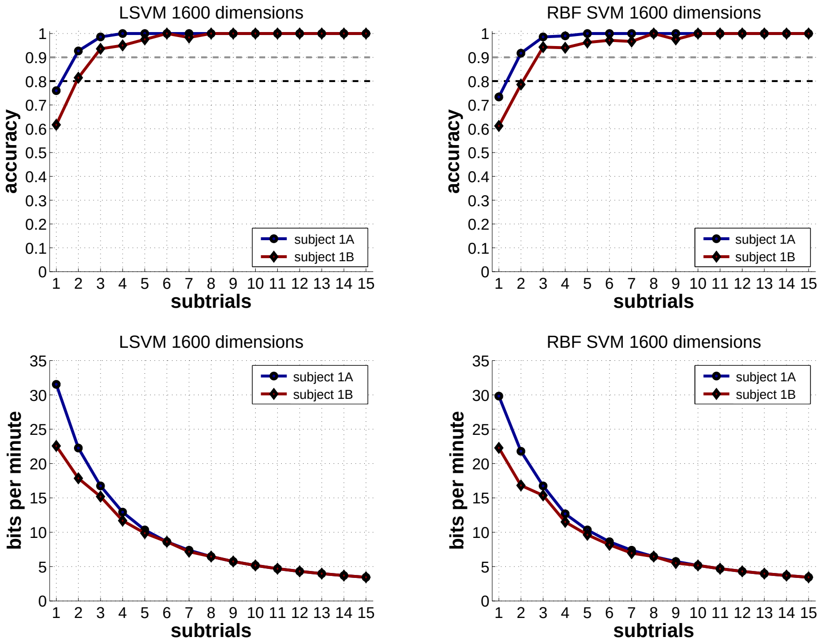


Figure 6.17: Classification accuracies (top row) and information transfer rates (bottom row) for subject 1A (blue graph) and subject 1B (red graph) using the classifiers Linear SVM (left) and RBF SVM (right) on data of 1600 dimensions.

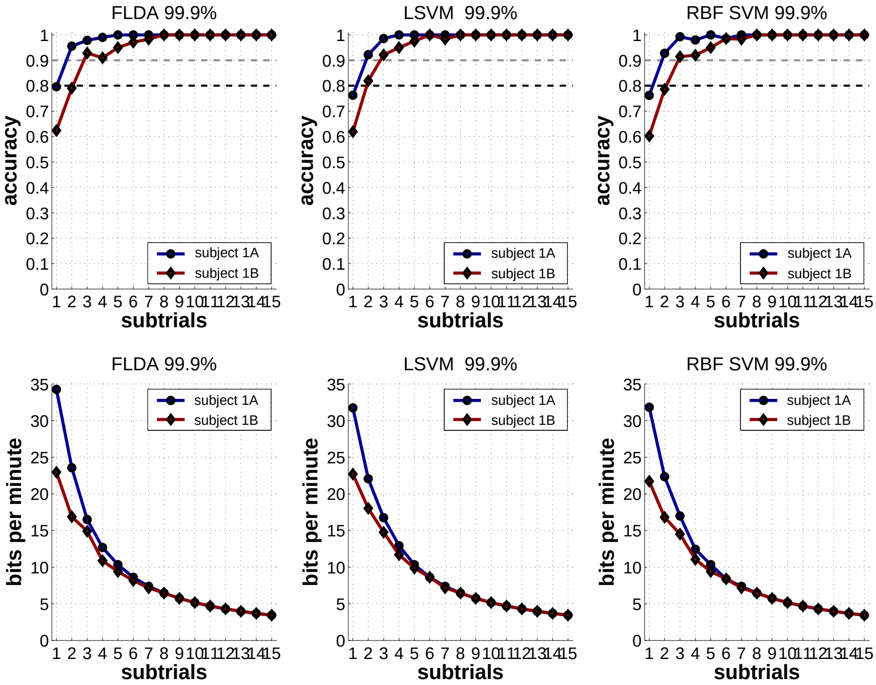


Figure 6.18: Classification accuracies (top row) and information transfer rates (bottom row) for subject 1A (blue graph) and subject 1B (red graph) using the classifiers FLDA (left), Linear SVM (middle), and RBF SVM (right) on data projected onto Principal Components capturing 99.9% variance (i.e., 84 for subject 1A and 95 for subject 1B).

17.86 bits/min and 15.37 bits/min for LSVM and RBF SVM, respectively.

For **dimensionality reduced features capturing 99.9% variance**, Figure 6.18 reveals the classification results and transfer rates for the different classifiers FLDA, LSVM, and RBF SVM under this condition. For **subject 1A**, the first exceedings of the 80% (and 90%) limit were reached after 2 subtrials for each classifier (FLDA: $p_{\text{acc}} = 0.956$, LSVM: $p_{\text{acc}} = 0.922$, RBF SVM: $p_{\text{acc}} = 0.928$). Best information transfer rates were consistently achieved for one subtrial (FLDA: 34.25 bits/min, LSVM: 31.74 bits/min, RBF SVM: 31.81 bits/min), but decreased when considering a minimum limit of 80% accuracy (FLDA: 23.57 bits/min, LSVM: 22.06 bits/min, RBF SVM: 22.34 bits/min).

When analyzing the dimensionality reduced data of **subject 1B**, 80% accuracy was reached after 3 subtrials with FLDA ($p_{\text{acc}} = 0.929$) and RBF SVM ($p_{\text{acc}} = 0.914$) as well, while LSVM only required 2 subtrials ($p_{\text{acc}} = 0.819$). Again, the best information transfer rates were observed for one subtrial (FLDA: 22.95 bits/min, LSVM: 22.70 bits/min, RBF SVM: 21.72 bits/min), but decreased when considering the 80% criterion (FLDA: 14.90 bits/min, LSVM: 18.02 bits/min, RBF SVM: 14.51 bits/min).

Discussion

Strong increases in classification performance were achieved by augmenting the electrode space from the Pz electrode to an ensemble of 10 electrodes distributed over the scalp. While it was necessary to employ at least 8 subtrials to reach 80% classification accuracy, only two subtrials for subject 1A, and three subtrials for subject 1B were required. This increase in classification accuracy also resulted in improved information transfer rates. The overall best transfer rate was 34.25 bits/min, and with a minimum classification accuracy of 80%, transfer rates between 14.90 bits/min and 22.25 bits/min were achieved.

Similar to the findings in the previous section, by projecting the data onto Principal Components capturing 99.9% of the data variance, no substantial reductions in classification performance, regardless of classification technique, were observed. No combination of classification and dimensionality reduction technique was superior to the other (see also Figure 6.19).

6.3.4 Conclusion

Similar to section 6.2, without loss of classification accuracy, strong dimensionality reductions could be performed with PCA. Only 84 (subject 1A) and 95 (subject 1B) Principal Components were necessary to employ for representing more than 99.9% of the variance of the original 1600-dimensional data space.

Binary classification revealed that strong classification improvements can be obtained when employing the whole set of 10 electrodes: While using one electrode yielded accuracies of maximal 0.707 (subject 1A) and 0.668 (subject 1B), with 10 electrodes, maximal accuracies of 0.886 (subject 1A) and 0.845 (subject 1B) were reached.

This trend was also confirmed when calculating accuracies for symbol inferences and corresponding information transfer rates, such that compared to section 6.2 the best information transfer rates could be improved from 6.33 bits/min and 4.49 bits/min up to 34.25 bits/min and 22.95 bits/min for subject 1A and subject 1B, respectively. As Figure 6.19 summarizes, in the average, all three classification techniques yielded

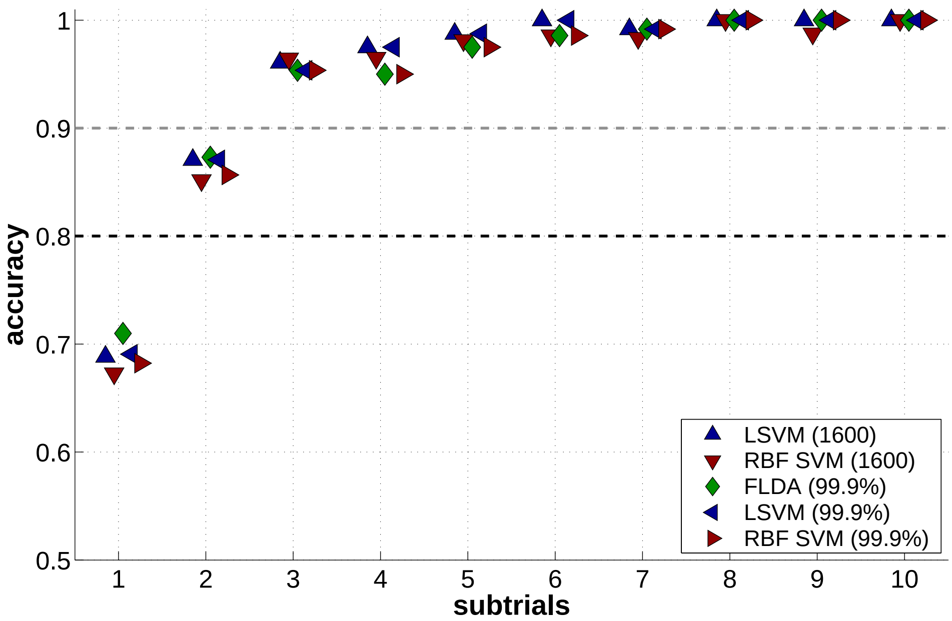


Figure 6.19: Classification accuracies for the different classifiers using the whole dimensionality of 1600 or 99.9% captured variance in Principal Components.

classification results within the same range, regardless of employed dimensionality.

Since e.g., Farwell and Donchin (1988) could improve their transfer rates by employing a shorter Interstimulus Interval, it can be suspected here that decreasing the ISI might also result in even better information transfer rates than obtained so far. Furthermore, analyzing data from just two subjects might not be sufficient for general assertions, such that the following section will also extend the investigations to more subjects.

6.4 Improving Information Transfer Rates by ISI Reduction

Although the classification accuracies could be improved in the previous sections by employing Machine-Learning classifiers and multiple electrodes, it would nevertheless be desirable to have higher information transfer rates. As discussed in section 4.2 and depicted in Figure 4.7, decreasing the Interstimulus Interval is capable to improve the information transfer rate. Little decreases in classification accuracies as they can be expected for small ISIs due to less pronounciated P300 components (see Figure 4.4) can easily be compensated by the increases in presentation speed. Thus, the ISI will be set to 140ms (125ms highlighting + 15ms delay) in the following to achieve better information transfer rates. Furthermore, in order to generalize the findings, data from eight subjects will be collected and analyzed.

Eight volunteers (denoted as 2A-2H, age 20-34) participated in the experiment. Each subject was instructed to count the flashings of one symbol in the matrix which was randomly chosen by the presentation program and presented to the subject in advance of each trial. A trial consisted of 4 to 6 subtrials. The subjects performed 450 to 720 subtrials each[4], but randomly selected subsets of 450 subtrials from each subject were

[4] The subjects performed two further blocks of different experiments, resulting in different numbers

Table 6.7: Number of Principal Components required to capture specific amounts of data variance for the subjects 2A to 2H.

	Captured Variance			
Subject	90%	95%	99%	99.9%
2A	22	33	64	90
2B	16	25	54	84
2C	24	35	66	91
2D	22	33	63	88
2E	16	27	55	83
2F	21	32	62	91
2G	25	37	65	91
2H	21	31	60	86
Mean	**20.88**	**31.63**	**61.13**	**88.00**

employed for the analyses. The Interstimulus Interval was set to 140ms which results in an overlap of the epochs because their length was 800ms. First investigations with the RBF SVM were already described in (Kaper and Ritter, 2004b) and (Kaper and Ritter, 2004a).

6.4.1 Principal Component Analysis

It is valuable to identify Eigenvalue characteristics of the Principal Components to become able to perform efficient dimensionality reductions. Thereby, it would be interesting which differences in the Eigenvalue distributions between the subjects exist.

Methods

From each subject, 900 $\mathcal{P}^+$ and 900 $\mathcal{P}^-$ samples were randomly chosen to perform the Principal Component Analysis for each subject separately.

Results

As exposed by Table 6.7, in the mean, at least 90%, 95%, 99%, and 99.9% data variance were captured in the first 21, 32, 62, and 88 Principal Components, respectively. Thereby, 99.9% of the data's variance of each subject was captured within the first 83 (subject 2E) to 91 (subjects 2C, 2F, 2G) first Principal Components. As much as 90% was captured even in the first 16 (subjects 2B and 2E) to 25 (subject 2G) Principal Components. Figure 6.22 gives an impression of the first Principal Components of the subjects[5].

Discussion

Very similar Eigenvalue distributions can be observed for the different subjects such that 99.9% of the variance is captured within the first 83 to 91 Principal Components. Therefore, employing 100 Principal Components appears to be sufficient to capture at least 99.9% variance of the EEG data for a subject. Since employing feature vectors capturing 99.9% variance produced competitive results in the previous sections and allowed to use FLDA, only feature vectors relying on projections onto the first 100 Principal Components will be considered in the following.

of subtrials.

[5] Note that the Principal Components are not assorted to their Eigenvalues in that figure. Details are explained in section 6.5.1

6.4.2 Binary Classification

Binary classification of the data on a balanced set yields a first estimation of classification performance for symbol inferences. In contrast to the previous sections, only dimensionality reduced feature vectors are considered.

Methods

As explained above, feature vectors of 100 dimensions were constructed for *each* subject as they contain at least 99.9% of the data's variance. The data from the different subjects were each divided into balanced halfs, a training and a test set, of randomly chosen $\mathcal{P}^+$ and $\mathcal{P}^-$ samples, and a 5-fold cross-validation was performed on the training set to find suitable hyperparameters for the classifiers. The classifiers were then trained with these hyperparameters on the training set and applied on the test set to classify the test data.

Results

Binary classification of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples yielded results as listed in Table 6.8. Mean accuracies of 0.831 ± 0.059, 0.828 ± 0.057 and 0.827 ± 0.059 were achieved using FLDA, LSVM, and RBF SVM, respectively. In the single cases, classification results ranged between 0.749 (subject 2B, RBF SVM) and 0.920 (subject 2F, FLDA).

Discussion

Although the ISI was strongly reduced, resulting in overlapping epochs and therewith in less pronounciated P300 components, encouraging classification accuracies between 0.749 and 0.920 were achieved. Again, the three different classifiers yielded comparable results. Thus, ISI reduction did not worse the signal quality in such a way that classification accuracies became unacceptable. In contrast, they stayed pleasantly high. Interindividual differences occurred, such that 3 of the 8 subjects stayed below 80% classification accuracy, and 5 exceeded 80% accuracy. Therewith, similar results to those obtained with 500ms ISI (cf. section 6.3.2) were achieved although the ISI was reduced.

6.4.3 Symbol Inferences

After achieving encouraging binary classification results even for the small Interstimulus Interval, computing symbol inferences also appears to be promising.

Methods

As in previous sections, symbol inference was calculated by dividing the dataset into two halfs, each containing 50% of the symbols employed in the experiment. First, cross-validation was performed on one set, revealing appropriate hyperparameters for the classifiers, which were then trained on this whole half, and classifications of symbols were performed on the other half. Afterwards, the halfs were interchanged and the procedure was repeated. In any case, training and test sets stem from the same subject and the same experiment.

Results

As depicted in Figure 6.20 (top row), each classifier reached 80% classification accuracy with 3 subtrials and 90% accuracy with 5 subtrials in the mean. For the best performing subjects, 80% accuracy was reached after 2 subtrials using FLDA (subject 2F,

Table 6.8: Binary classification accuracies performed by FLDA, LSVM, and RBF SVM for subjects 2A to 2H based on data projections onto the first 100 Principal Components.

	Subject								
Method	2A	2B	2C	2D	2E	2F	2G	2H	**Mean**
FLDA	0.882	0.761	0.808	0.842	0.881	0.920	0.769	0.783	0.831 ± 0.059
LSVM	0.871	0.768	0.808	0.833	0.886	0.913	0.770	0.778	0.828 ± 0.057
RBF SVM	0.876	0.749	0.807	0.839	0.882	0.912	0.776	0.774	0.827 ± 0.059

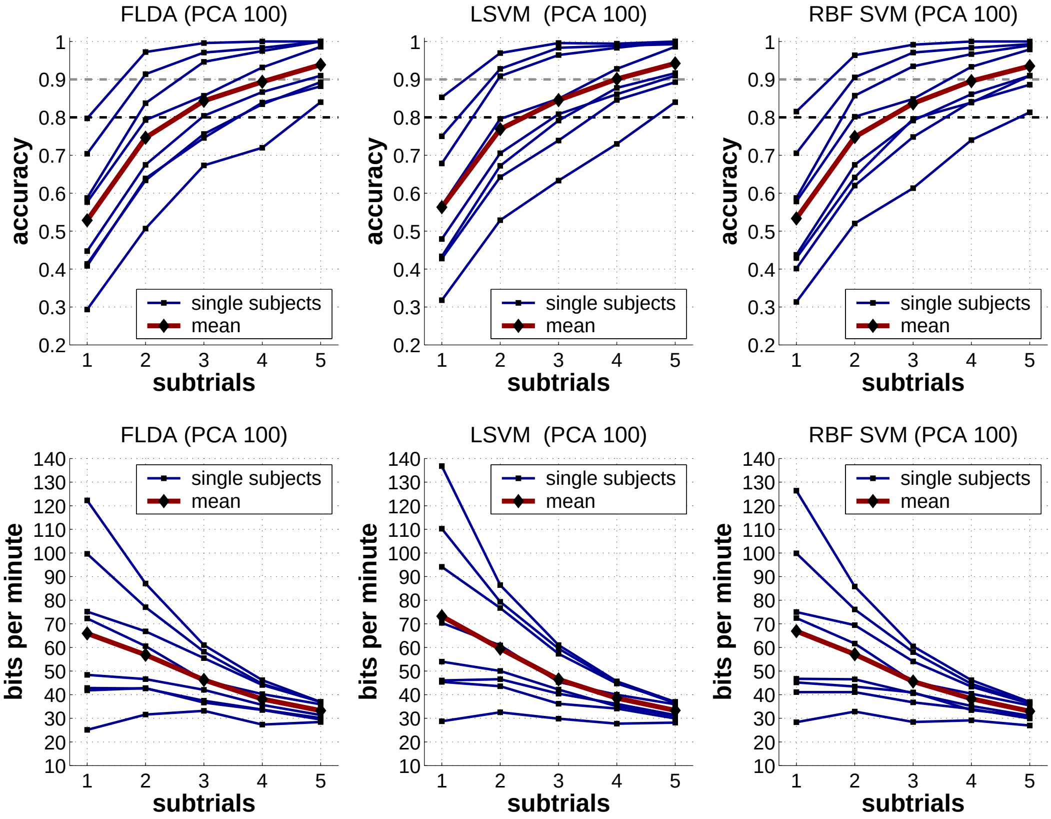


Figure 6.20: Classification accuracies (top row) and information transfer rates (bottom row) for the eight subjects as yielded by the classifiers FLDA, LSVM, and RBF SVM using 100 Principal Components for dimensionality reduction. For certain numbers of aggregated subtrials for symbol inference, thick red lines reflect the mean performances among subjects, while their single performances are drawn thin blue.

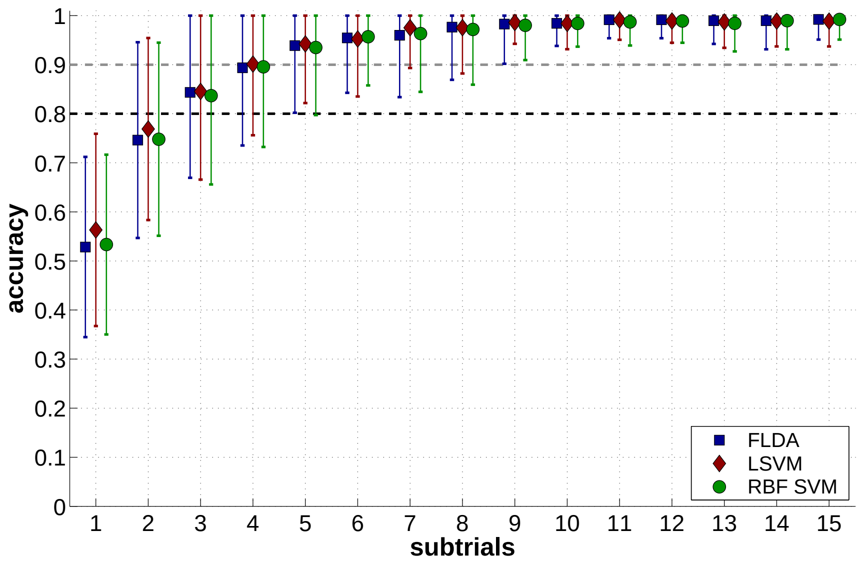


Figure 6.21: Mean classification accuracies and standard deviations as obtained for the classifiers FLDA (blue squares), LSVM (red diamonds), and RBF SVM (green circles).

$p_{\text{acc}} = 0.972$), and after only one subtrial using the SVM classifiers (subject 2F, RBF SVM: $p_{\text{acc}} = 0.815$, LSVM: $p_{\text{acc}} = 0.853$). In contrast, the worst performing subject reached 80% accuracy after 5 subtrials, regardless of classification technique.

Best mean information transfer rates were 65.94 bits/min (FLDA), 73.22 bits/min (LSVM), and 66.89 bits/min (RBF SVM), as also depicted in Figure 6.20 (bottom row). Under the precondition of reaching a limit of 80% accuracy first, information transfer rates of 46.26 bits/min (FLDA), 46.48 bits/min (LSVM), and 45.59 bits/min (RBF SVM) resulted. The best information transfer rate for a single subject was 136.77 bits/min (LSVM), and the worst rate for reaching 80% accuracy was 26.94 bits/min (RBF SVM).

Discussion

As in previous examinations, the performances of the three classification techniques were within the same range and no remarkable performance differences could be observed for reaching 80% accuracy. Comparable classification performances for the certain numbers of subtrials as described in the previous section employing an ISI of 500ms were also achieved for 150ms ISI and the larger population of eight subjects. Due to the lower ISI in the current experiment, the amount of information that can be transferred within in a specific time increased: In the mean, transfer rates of 73.22 bits/min could be achieved and of 46.48 bits/min when considering the limit of at least 80% classification accuracy. In single cases, transfer rates of up to 136.77 bits/min were achieved, which corresponds to spelling about 26 symbols in a minute. Strong differences in classification performance between the subjects were found: While some subjects achieved more than 100 bits/min, others produced less than 30 bits/min.

The calculation of the transfer rates does not consider delays between trials such that they remain quite theoretical. Thus, assuming a delay of 2 seconds between trials would yield 62.44 bits/min instead of 136.77 bits/min for the best case and 33.28 bits/min instead of 46.48 bits/min in the mean. Nevertheless, these transfer rates are competitive compared to other approaches as section 3.3 indicates. The highest reported information transfer rates of alternative approaches were 68.00 bits/min (Gao et al., 2003) and 50.50 bits/min (Blankertz et al., 2003).

6.4.4 Conclusion

Principal Component Analysis on data from the eight subjects revealed that employing the first 100 Principal Components is sufficient to capture more than 99.9% of the variance of the data in these cases. Thus, subsequently, feature vectors of projections onto the first 100 Principal Components were used for classification. This was further motivated by the findings in the previous sections that data reductions to capture 99.9% variance also produced competitive results and enabled to use FLDA.

Despite of ISI reduction, binary classification revealed that the EEG signals could nevertheless reasonably be distinguished as classification accuracies between 0.749 and 0.920 indicate. Therefore, as initially suspected, reducing the Interstimulus Interval yielded higher information transfer rates and led to competitive theoretical information transfer rates of up to 136.77 bits/min and 46.48 bits/min in the mean. Nevertheless, quite strong interindividual differences in classification performance were found and the upcoming section will investigate generalizations among the subjects.

6.5 Generalization Capabilities

Generalization capabilities of preprocessing and classification strategies are to be investigated in this section in two ways. First, it will be examined whether it is possible to calculate a PCA matrix on data from a set of subjects and apply it for the purpose of dimensionality reduction on other subjects. Success with this procedure would allow to calculate a *general PCA matrix* once on a set of subjects and use it for dimensionality reduction ever after without the necessity of calculating an own PCA matrix for a new subject.

Second, in a similar fashion, and as already suggested in Kaper and Ritter (2004a), the generalization capabilities of the classifiers are investigated. Data from a set of subjects are used to train a classifier, which will then be applied to classify data from another subject, whose data were not included in the training set. If this procedure succeeds, it would even be possible to use a *general classifier* which does not need to be trained on data from the individual subject. A new user could therefore *immediately start to operate* the system when employing the general classifier.

6.5.1 Principal Component Analysis

Methods

In a first step, the Principal Components from the previous section of the single subjects were investigated with respect to interindividual differences in the PCA structures. For this purpose, the Euclidean distances between the components were calculated and they were subsequently assorted according to these values:

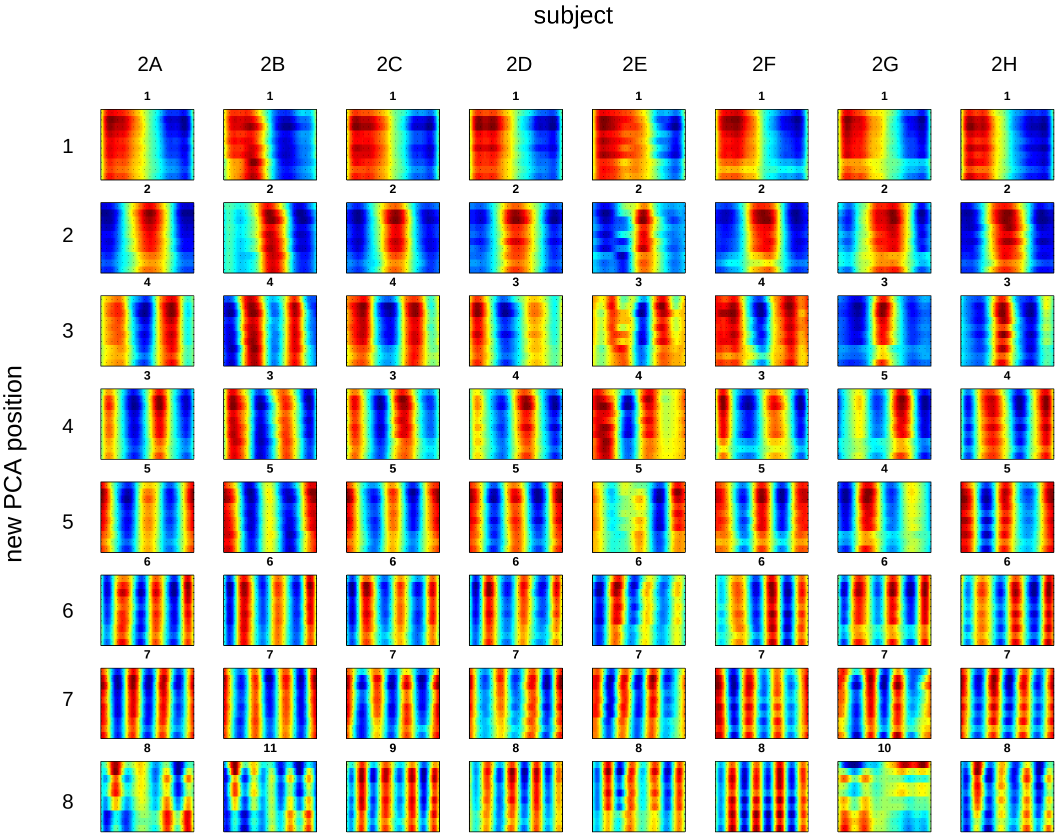


Figure 6.22: Comparison of Principal Components from different subjects. Each row contains similar Principal Components from the different subjects which are assorted in the columns. The original position of each Principal Component within each subject is drawn on top of each component.

First, the index k of the Principal Component $\mathbf{PCA}_{\mathrm{2B,k}}$ of subject 2B which exposes the minimum distance to the first Principal Component $\mathbf{PCA}_{\mathrm{2A,1}}$ of subject 2A was computed. Since the sign of Principal Components can be arbitrary, this distance was measured regardless of polarity:

$$k_{\mathrm{2B,1}} = \arg\min_{i}\Big(\min(\; \|\, \mathbf{PCA}_{\mathrm{2A,1}} - \mathbf{PCA}_{\mathrm{2B,i}} \,\|_2,\; \|\, -\mathbf{PCA}_{\mathrm{2A,1}} - \mathbf{PCA}_{\mathrm{2B,i}} \,\|_2 \,)\Big). \tag{6.8}$$

This procedure was iterated over all 8 subjects, such that the minimum distances of the Principal Component $\mathbf{PCA}_{\mathrm{2A,1}}$ to all other subjects were assessed. Afterwards, this assortion algorithm was repeated for each initial Principal Component $\mathbf{PCA}_{\mathrm{2A,j}}$ with $j \in \{1, ..., 1600\}$ of subject 2A. The new position $\hat{k}_j$ of the resulting Principal Components was then calculated by assorting the components according to the mean of the indices $k_{s,\cdot}$:

$$\hat{k}_j = \frac{1}{8} \sum_{s \in \{2A, ..., 2H\}} k_{s,j}. \tag{6.9}$$

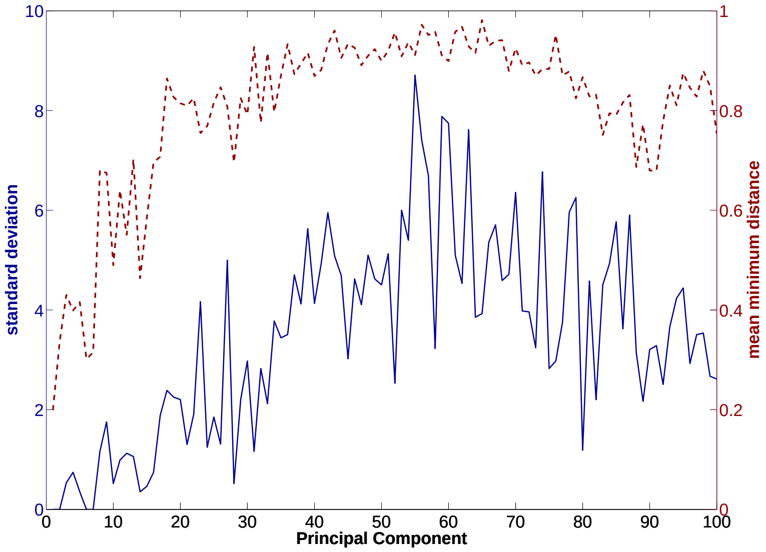


Figure 6.23: Standard deviations (solid, left scale, blue) of the original positions and means of minimum distances to each other (dashed, right scale, red) of the Principal Components from the eight subjects.

Results

The first eight assorted Principal Components for each subject are depicted in Figure 6.22. Similar structures in terms of frequency and phase relationship can be observed among subjects at each new PCA position. With increasing numbers of Principal Components, the frequency of the components is also raising within the considered range. The standard deviations of the original sources of the Principal Components within a certain position k_{new} as well as the means of the Euclidean distances are depicted in Figure 6.23 and expose that the first 100 Principal Components also stem from proximal sources of less than 10 positions away from each other.

Discussion

Similar Principal Components were found for the different subjects. This encourages to follow the suspicion that it might be possible to have one Principal Component matrix for dimensionality reduction purposes for a large variety of subjects.

For that purpose, an equal number of 900 $\mathcal{P}^+$ and 900 $\mathcal{P}^-$ samples from each of the eight subjects was collected, and Principal Component Analysis was performed on the whole collection of data to calculate the *general PCA matrix*, as it will also be used in chapter 7.

As an outcome of this analysis, the Eigenvalue distribution shows that 90% variance are captured within the first 22 Principal Components, while 99% and 99.9% variance are reflected by the first 66 and 91 Principal Components, respectively. The first 25 Principal Components are shown in Figure 6.24. Similar structures to the Principal Components from the single subjects can be identified: Again, different frequencies for the whole set of electrodes are reflected by the different Principal Components. Differentiations between electrode sites occur for PCA 10 and for PCA 12 and the following. Mostly, a cluster representing $Oz, PO7$ and $PO8$ together exposes a different structure compared to the other electrode positions.

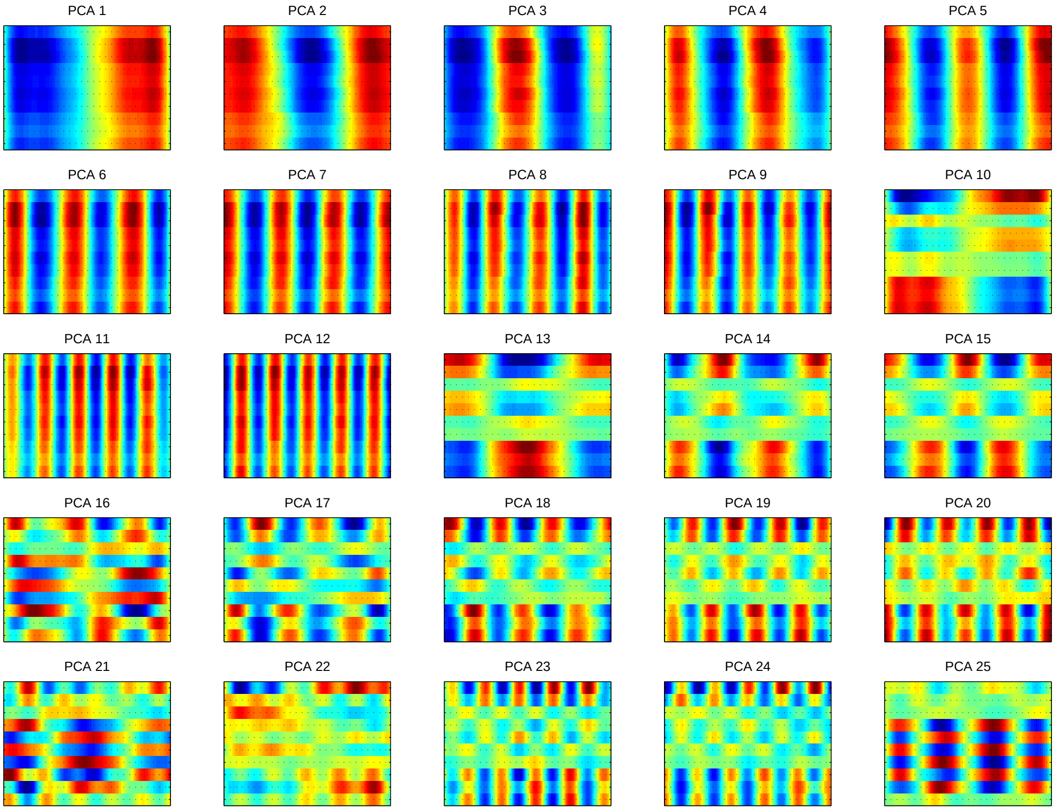


Figure 6.24: The first 25 Principal Components as derived from data collected from all eight subjects. Figure 6.13 explicates the structure of the images.

6.5.2 Binary Classification and Symbol Inferences using PCA Matrices from disjoint Subjects

The previous analyses revealed similar Principal Components for different subjects, which supports the assumption that a general PCA matrix for the purpose of dimensionality reduction can be constructed. In the following, based on the data from the subjects 2A-2H, the generalization capabilities of the PCA matrix will be investigated.

Methods

In a leave-one-case-out scheme, data from seven subjects were employed to calculate a PCA matrix. Afterwards, this matrix was applied to reduce the dimensionality of the data from the omitted 8th subject down to 100 dimensions. The procedure was repeated for each subject serving as the omitted test subject. The dimensionality reduced data from the latter subject were then analyzed in the same way as in previous sections (see, e.g., section 6.3.2 for binary classification and section 6.2.5 for symbol inference). For binary classification, the data were divided into two balanced halfs of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples, one half serving as a training, and the other half serving as a test set. Within a 5-fold cross-validation on the training set, appropriate parameters were found, with which the classifiers were trained on the whole training set. Afterwards, data from the test set were classified. On the other hand, for symbol inference, the data were split into two halfs, one half serving as a training set, and the other half serving as a test set and vice versa. Training was performed as in binary classification, and symbol inferences were computed as described in previous sections. The procedure was repeated for each of the eight subjects.

Results

Binary classification of $\mathcal{P}^{+}$ and $\mathcal{P}^{-}$ samples yielded results as listed in Table 6.9. In the mean, accuracies of 0.844±0.050 (FLDA), 0.832±0.057 (LSVM) and 0.830±0.059 (RBF SVM) were achieved. Classification results for the single subjects ranged from 0.746 (subject 2B, RBF SVM) to 0.923 (subject 2F, FLDA and RBF SVM).

As it is depicted in Figure 6.25 (top row), for **symbol inference**, an accuracy level of 80% was consistently achieved in the mean after 3 subtrials, irrespective of the classification technique. For reaching an accuracy of 90%, the aggregation of 5 subtrials was necessary for each classifier. The best performing subject (2F) reached the 80% criterion after only one subtrial and the 90% criterion after two subtrials while the worst performing subject (2D) reached the 80% criterion after only 5 subtrials and the 90% criterion using 6 subtrials. Figure 6.26 summarizes the *mean* classification accuracies obtained with the different classifiers.

Best information transfer rates of 64.69 bits/min, 70.20 bits/min, and 65.18 bits/min were obtained in the mean for the classifiers FLDA, LSVM, and RBF SVM, respectively (see Figure 6.25, bottom row). With three subtrials (which correspond to reaching 80% classification accuracy), transfer rates of 44.98 bits/min, 46.68 bits/min, and 43.87 bits/min were obtained with the three classifiers. The best transfer rate was 136.00 bits/min (subject 2F, RBF SVM) and the worst rate for reaching 80% accuracy was 26.96 bits/min (subject 2G, RBF SVM).

Discussion

When comparing the results to those obtained using a PCA matrix from the same subject, similar outcomes resulted. Thus, for the purpose of dimensionality reduction, a general PCA matrix can be computed which yields no remarkable loss in classification accuracy for a new subject Like in the previous experiments, the classification accuracies among classifiers were comparable.

6.5.3 Binary Classification and Symbol Inferences using PCA Matrices and Training Sets from disjoint Subjects

After a general PCA matrix provided reasonable results, it is now to be investigated whether it is even possible to use a classifier in a pretrained fashion without the necessity of training it on previously recorded data from the same subject. Within this section, therefore not only the PCA matrix, but also the training set stems from a disjoint set of subjects. Again, similar to the previous investigation for calculating the general PCA matrix, the classifiers are trained and tested in a leave-one-case-out scheme.

Methods

Principal Component Analysis was performed on data from 7 subjects, and the different classifiers FLDA, LSVM, and RBF SVM were also trained on balanced sets of data from these subjects as in the previous section. Then, the PCA as well as the classifiers were applied on the whole data from the omitted 8th subject for binary classification (for which a balanced set was extracted) and for symbol inferences as described above in

Table 6.9: Classification results based on projections onto the first 100 Principal Components. Data from 7 subjects were taken to train the classifier and to calculate the PCA. The trained classifier was then applied on the omitted subject. This procedure was performed for each subject.

	Subject								
Method	2A	2B	2C	2D	2E	2F	2G	2H	**Mean**
FLDA	0.878	0.828	0.821	0.810	0.896	0.923	0.774	0.825	0.844 ± 0.050
LSVM	0.850	0.753	0.823	0.824	0.902	0.917	0.777	0.806	0.832 ± 0.057
RBF SVM	0.854	0.746	0.806	0.830	0.898	0.923	0.778	0.807	0.830 ± 0.059

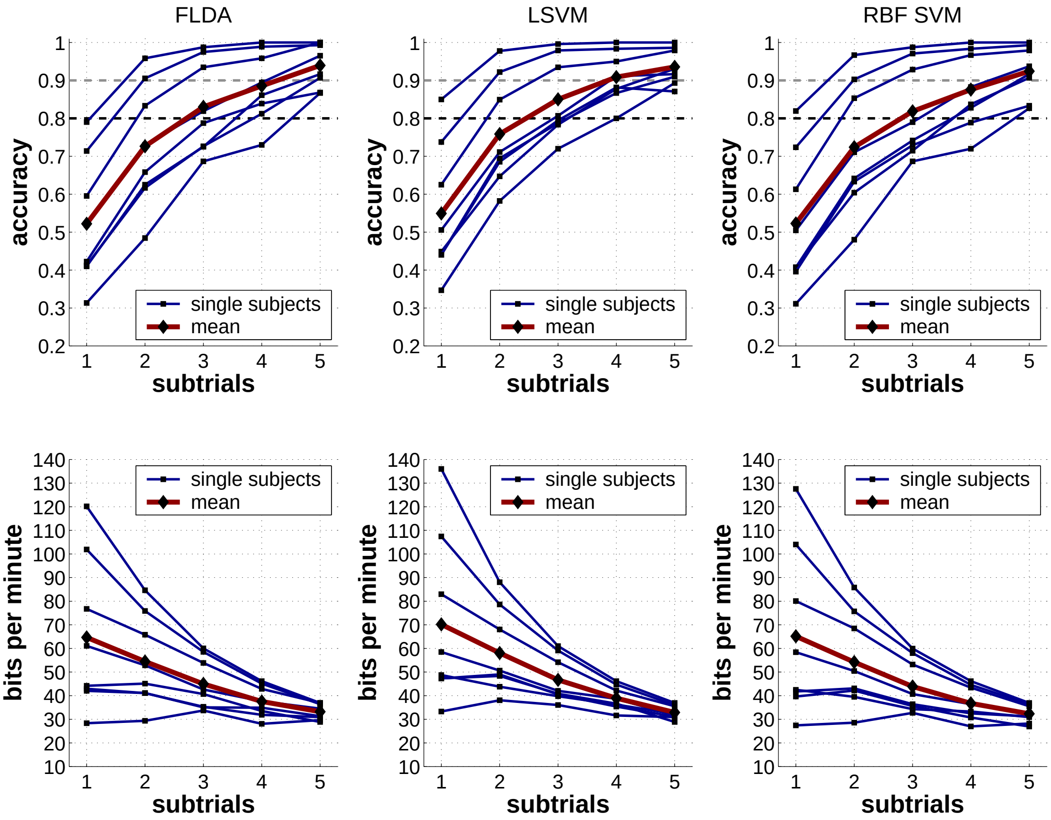


Figure 6.25: Classification accuracies (top row) and information transfer rates (bottom row) for eight subjects as yielded by the classifiers FLDA, LSVM, and RBF SVM on 100-dimensional feature vectors. The PCA matrix for dimensionality reduction was calculated from data from 7 subjects and applied on data from the omitted subject while the training sets stem from the same subject as the test set. For certain numbers of aggregated subtrials for symbol inference, thick red lines reflect the mean performances among subjects, while their single performances are drawn blue thin.

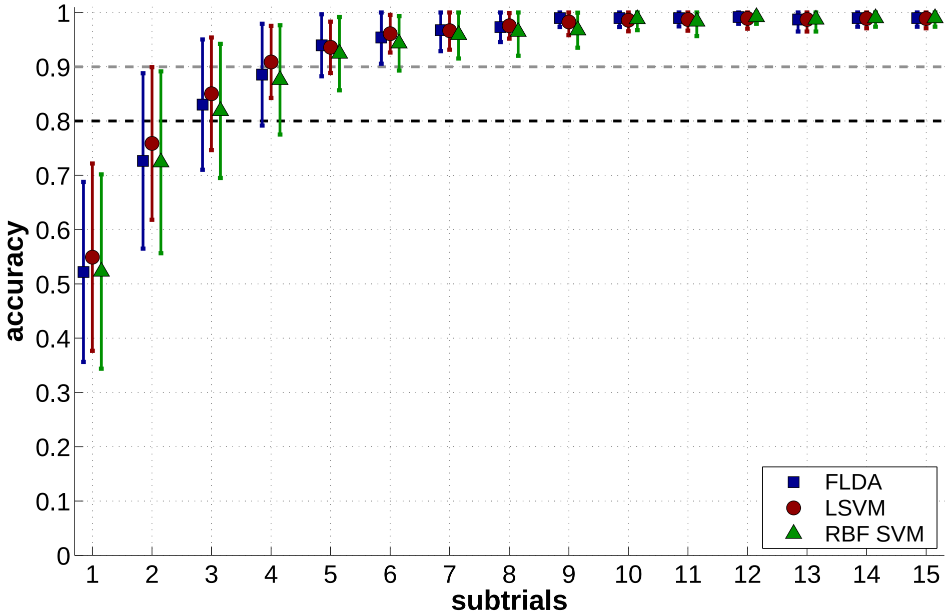


Figure 6.26: Mean classification accuracies and standard deviations as obtained for the classifiers FLDA (blue squares), LSVM (red circles), and RBF SVM (green diamonds) for computing the PCA on data from disjoint subjects.

more detail. This procedure was repeated for each of the eight subjects.

Results

In **binary classification** of $\mathcal{P}^+$ and $\mathcal{P}^-$ samples, results as listed in Table 6.10 were obtained. Mean accuracies of 0.726±0.054 (FLDA), 0.750±0.075 (LSVM) and 0.751±0.076 (RBF SVM) resulted and the classification results ranged between 0.661 (subject 2B, RBF SVM) and 0.862 (subject 2F, LSVM).

For **symbol inference**, in the mean, an accuracy level of 80% was achieved with 7 subtrials for FLDA and LSVM, but only with 8 subtrials for the RBF SVM (see Figure 6.27, top row). For 90% accuracy, at least 12 subtrials were needed (FLDA and LSVM), but RBF SVM did not succeed to reach this criterion within 15 subtrials. However, with 15 subtrials, an accuracy of $p_{\text{acc}} = 0.899$ was nevertheless achieved. For a direct comparison, mean accuracies and standard deviations are summarized in Figure 6.28. The best performing subjects (subjects 2A and 2H) reached the 80% criterion after two subtrials (FLDA and LSVM) and the 90% criterion utilizing three subtrials (with the RBF SVM as well), while the worst performing subject failed to reached the 80% criterion within 15 subtrials (subject B).

As Figure 6.27 (bottom row) exposes, the best mean information transfer rates were 38.89 bits/min (LSVM), and 20.89 bits/min (LSVM) when exceeding 80% accuracy after 6 subtrials. The best information transfer rate was 92.61 bits/min (subject H, LSVM), and at worst, 80% accuracy was never achieved within 15 subtrials.

Discussion

Applying a classifier which was trained on data from a set of subjects and used for classifying data from a new, disjoint subject, yielded a lower performance than training

Table 6.10: Classification results in binary classification based on feature vectors derived from the first 100 Principal Components. Data from 7 subjects were taken to train the classifier and to calculate the PCA. The trained classifier was then applied on the omitted subject.

	Subject								
Method	2A	2B	2C	2D	2E	2F	2G	2H	**Mean**
FLDA	0.772	0.686	0.664	0.704	0.792	0.804	0.698	0.689	0.726 ± 0.054
LSVM	0.801	0.693	0.655	0.719	0.840	0.862	0.707	0.722	0.750 ± 0.075
RBF SVM	0.804	0.681	0.661	0.729	0.843	0.861	0.707	0.718	0.751 ± 0.076

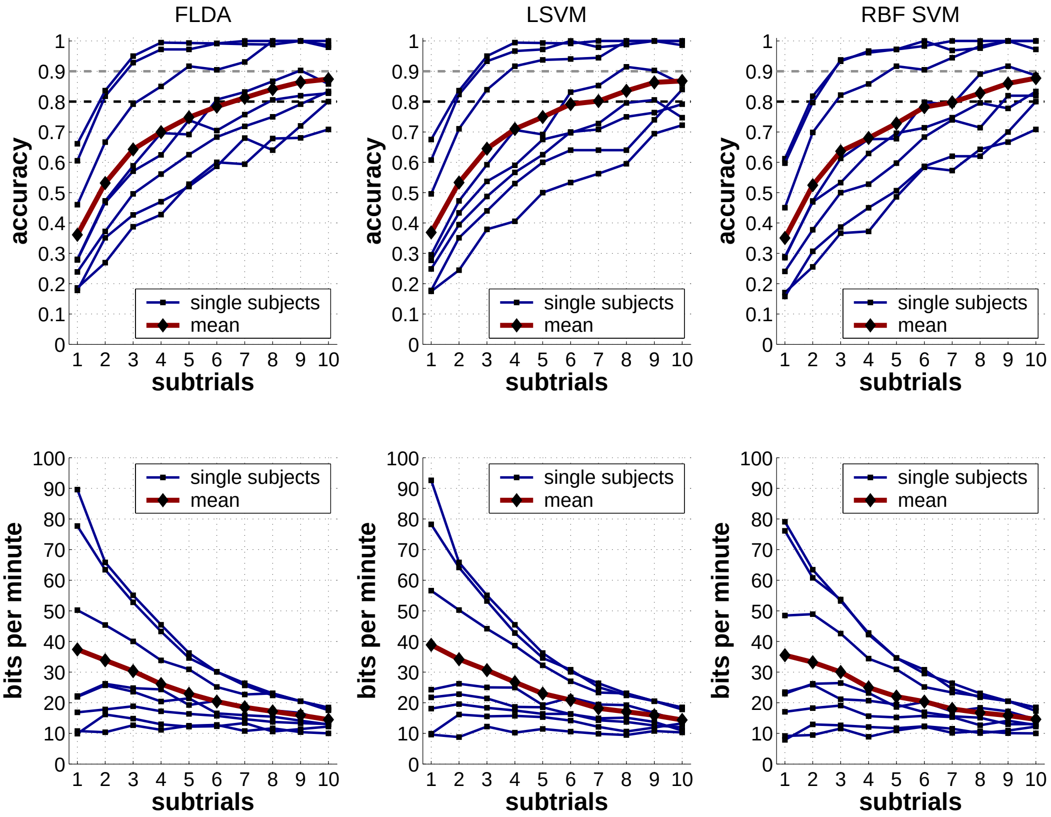


Figure 6.27: Classification accuracies (top row) and information transfer rates (bottom row) for eight subjects as yielded by the classifiers FLDA, LSVM, and RBF SVM. The PCA matrix for dimensionality reduction down to 100 dimensions and the training set as well were based on data from 7 subjects. The trained predictor was then applied on data from the omitted subject. For certain numbers of aggregated subtrials for symbol inference, thick red lines reflect the mean performances among subjects, while their single performances are drawn blue thin.

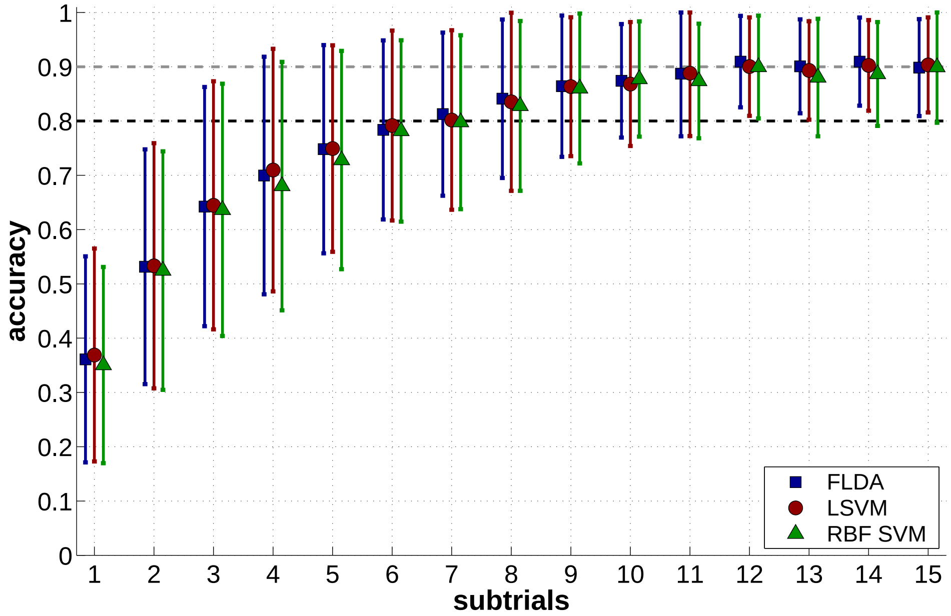


Figure 6.28: Mean classification accuracies and standard deviations as obtained for the classifiers FLDA (blue squares), LSVM (red circles), and RBF SVM (green diamonds) for training the classifier on data from disjoint subjects.

the classifier on data from the same subject.

However, classification accuracies only decreased to a level where operating the BCI is slow, but still possible. According to the outcome of the experiment, in the mean, a subject would achieve a transfer rate of 38.89 bits/min. With a criterion of at least 80% accuracy, 20.89 bits/min would still result.

Note that this information transfer rate nevertheless lies within the range of several other Brain-Computer Interfaces as the overview in section 3.3 reveals. The variance between the subjects was high: While single subjects could achieve transfer rates up to 92.61 bits/min, others remained at a level of about 10 bits/min.

6.5.4 Conclusion

As it is summarized in Figures 6.29 and 6.30, it was possible to apply PCA matrices computed on data from a set of subjects to data from new subjects for the purpose of dimensionality reduction. Thereby, almost the same results compared to utilizing a PCA matrix calculated on data from the same subject were achieved.

In contrast, applying a Machine-Learning classifier which was trained on data from other subjects produced worse performances. Nevertheless, high transfer rates could still be achieved in some cases and even in the mean, operating the device would still be possible, but only comparably slow.

Therefore, a P300-based BCI could be created to be operated by an user without *any* prior training. Neither the subject, nor the PCA, nor the classifier would need to be trained for the individual subject, and in the mean, a theoretical transfer rate of 38.89 bits/min would still be possible, which could rise up to more than 90 bits/min for some subjects.

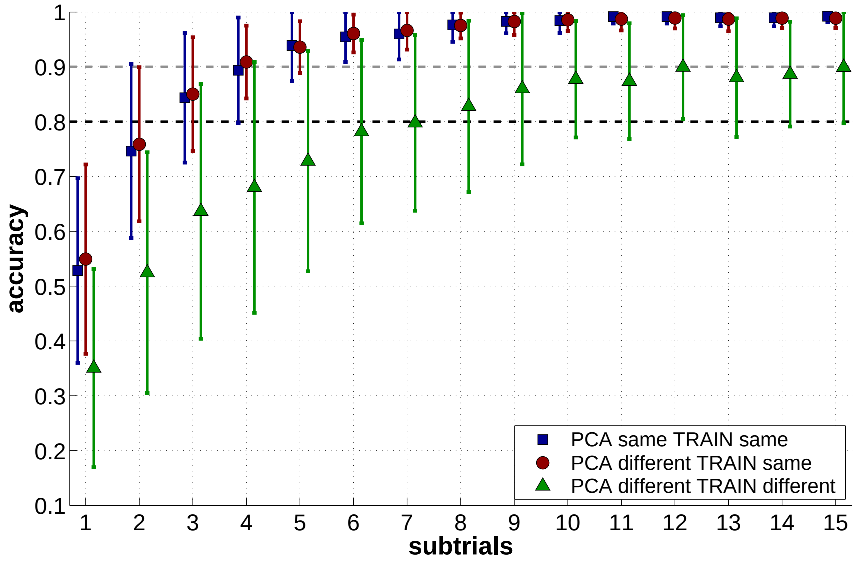


Figure 6.29: Mean classification accuracies and standard deviations obtained using data from the *same* subject or from *different* subjects for computing the PCA matrix and/or training the FLDA classifier.

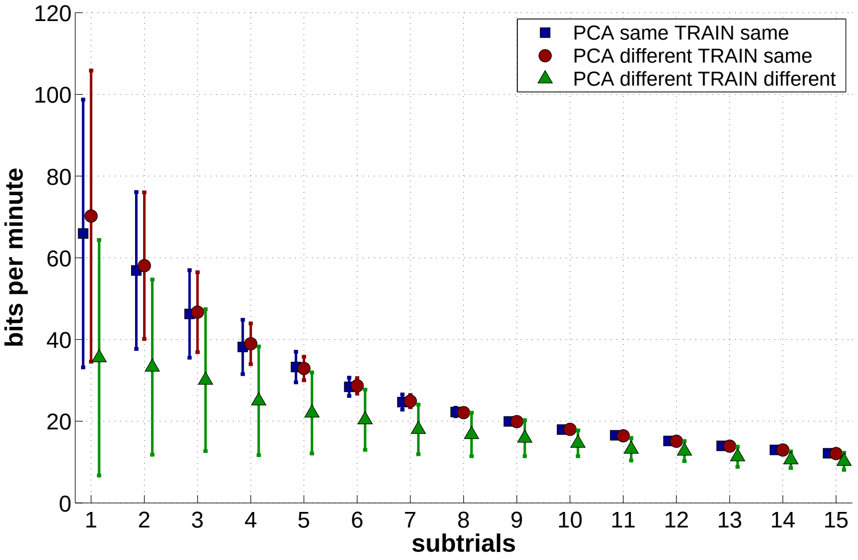


Figure 6.30: Mean information transfer rates and standard deviations obtained using data from the *same* subject or from *different* subjects for computing the PCA matrix and/or training the FLDA classifier.

6.6 BCI Competition 2003

Whenever a research group is conducting own BCI experiments and use own algorithms for classification, a confoundation of experimental methods (e.g., data recording device, characteristics of the subjects, experimental environment) and the algorithm's performance exists. In order to become able to compare the performance of algorithms, Blankertz et al. (2004) conducted a competition for best classification algorithms on given BCI data sets[6].

The group published data sets for several kinds of BCI approaches. Each set was subdivided into two further sets, a training and a test set. While the labels were included in the training set, they were missing in the test set. The objective in the competition was to infer the correct labels for the test set, by training a classifier only on data from the training set.

Methods

EEG data were derived from 64 channels at 240Hz sampling rate from one subject and collected within three sessions, denoted as session 10, 11, and 12. While sessions 10 and 11 provided training data accompanied by valid labels, the 12th session came without labels, and it was the challenge of the competitors to infer these labels. In each session, a 6×6 matrix was presented to the subject, and an investigator provided a word on which characters the subject should sequentially focus attention to. Rows and columns were flashed with a frequency of 5.7Hz, corresponding to an ISI of 175 ms. For each character, 15 subtrials were recorded.

In the author's contribution (Kaper et al., 2004), an earlier version of the RBF SVM variant of the classification procedure was utilized: Epochs of 600ms were extracted and for preprocessing, a band pass filter of 0.5-30Hz was used before normalizing the data to [-1,1]. From the provided training set, balanced data from the same 10 electrode sites as in the previous sections were extracted for the data analysis procedure. Within a 5-fold cross-validation, suitable values for the hyperparameters C and γ were found for the RBF SVM on the whole training data. Subsequently, the classifier was trained on the training set using these hyperparameter values. The trained classifier was then applied on data from the test set, and symbol inference was computed as explained in section 5.4.

Results

For the calculated hyperparameter values $C = 20.007$ and $\gamma = 6.68 \cdot 10^{-4}$, a 5-fold cross-validation on the training set revealed an accuracy of 0.845 for separating $\mathcal{P}^+$ from $\mathcal{P}^-$ epochs. When analyzing the test set for the different numbers of subtrial combinations n_{combined}, this resulted in the inferred symbols shown in Table 6.11. The accuracy p_{acc} increased with the number of combined subtrials from 0.645 to 1.000, and the correct solution was found after five subtrials. When choosing 80% correct classification as satisfying (Farwell and Donchin, 1988), only three repetitions would be necessary.

Discussion

The objective of the competition was to infer the correct symbols with maximal 15 subtrials, and the algorithm of the author therefore qualified, next to four other contributions, as a winner. But furthermore, the algorithm required only 5 subtrials to

[6] A further competition was conducted in 2005, the author did not attend to.

Table 6.11: Inferred words and associated accuracies p_{acc} from the test set for different numbers of aggregated subtrials n_{combined} (Kaper et al., 2004).

n_{combined}	**Inferred words**	p_{acc}
1	`FOOD MOOT BBM PIE CAXE NCNA N5AO6 X5Z7`	0.645
2	`FOOD MOOT BBM PIE CALE TCBA Z5AOT X5Z7`	0.710
3	`FOOD MOOT HAM PIE CALE TCNA ZYAON X567`	0.839
4	`FOOD MOOT HAM PIE CALE TUNA ZYGOT 4567`	0.968
5	`FOOD MOOT HAM PIE CAKE TUNA ZYGOT 4567`	1.000

reach perfect accuracy. Only one competitor (Xu et al., 2004) reached a comparable performance. In summary, the algorithm did not only work well on own data, but also on external data and proved its performance in an objective comparison with other algorithms.

6.7 Summary

In this chapter, offline BCI experiments with the P300 speller paradigm, originally proposed by Farwell and Donchin (1988) were conducted. Several aspects were analyzed with the goal to enhance the classification performance and the speed of the BCI device:

First, comparisons between the Model-Based classifiers *area* and *peak picking* and the Machine-Learning classifiers Linear SVM and RBF SVM were performed on data derived from the Pz electrode from two subjects. It turned out that the Machine-Learning approaches outperformed the Model-Based techniques. Furthermore, it was possible to reduce the dimensionality of the data by projecting them onto the first 11 Principal Components, reflecting 99.9% data variance, which in turn allowed to use the computational rather inexpensive Fisher's Linear Discriminant. Comparable results were obtained for the different Machine-Learning classifiers, regardless of the employed dimensionality of the feature vector (11 or 160).

Second, inspired by the ERP scalp distribution of $\mathcal{P}^+$ and $\mathcal{P}^-$ epochs, data from a *set* of 10 electrodes were employed for classification. Again, strong dimensionality reduction could be obtained by employing Principal Component Analysis.

Third, information transfer rates could be improved by increasing the presentation speed through Interstimulus Interval reduction. Theoretical information transfer rates of up to 136.77 bits/min and of 73.22 bits/min in the mean were achieved in experiments with eight subjects, which decreased to 62.44 bits/min and 33.43 bits/min, respectively, when assuming a delay of 2 seconds between trials. All three classification techniques yielded results within the same range.

Fourth, it was shown that it is possible to use a general PCA matrix and still achieve almost the same results as obtained with a PCA matrix calculated for the individual subjects. Thus, for future experiments, no prior PCA calculation is necessary for the individual subject when using the same setup. When trying to apply classifiers in a similar fashion, it turned out that they produced worse results. Nevertheless, in some cases, competitive information transfer rates could still be achieved, and in the mean, theoretical transfer rates within the range of other approaches resulted. Thus, it is in principle possible to use the classifiers without any prior training, such that a subject could immediately start to *operate* the system.

Finally, an earlier variant of the algorithm won the BCI Competition 2003 for the P300 speller section, in which algorithms were competing for best classification results on an independently published dataset. Among the winners, the algorithm proposed by the author required the least number of subtrials.

Chapter 7

From Offline to Online Analysis

The analyses of the offline experiments in the previous chapter have provided valuable insights which can now be employed for constructing an online Brain-Computer Interface (BCI) where an user can actually *operate* the BCI device. For this purpose, data need to be processed and classified directly after the recording from the subject's scalp (see Figure 7.1). Several outcomes from the previous chapter are useful for designing the device. For instance, the finding that a general PCA matrix can be used for dimensionality reduction, which must not be computed for the individual subject conveniently allows to perform dimensionality reductions with little computational costs, allowing to directly employ Fisher's Linear Discriminant Analysis (FLDA). Another outcome of the previous chapter has been that with an adequate preprocessing, the FLDA classifier achieves similar performance as the state-of-the-art Support Vector Machine.

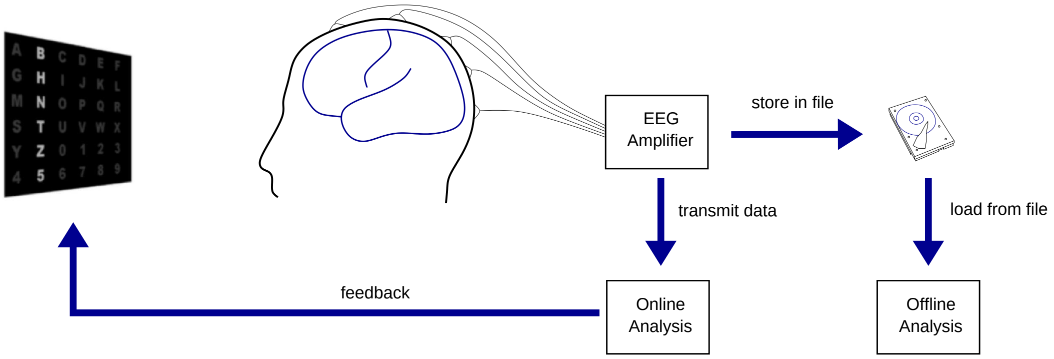


Figure 7.1: Scheme for *online* BCI analysis as performed in this thesis. The EEG data are recorded from the subject's scalp and processed by an EEG amplifier. After each trial, the incoming EEG data are classified and the results, i.e., the predicted symbol, can directly be presented as a feedback.

This chapter describes the construction of an online P300-based Brain-Computer Interface from the scratch. After considerations about mandatory properties of the system in section 7.1, the hardware environment is discussed (see section 7.2). Afterwards, the layout of the software for driving the BCI is explained in section 7.3 and an experiment is conducted with this setup in section 7.4.

7.1 Deriving Specifications for Designing and Driving an Online System

Driving an online Brain-Computer Interface is accompanied by different requirements for the system compared to offline analysis. For instance, data need directly to be processed and analyzed after recording. The data analysis procedure in turn should not be too time consuming, such that it is desirable to employ a classification algorithm with only little computational costs. Another substantial difference to offline analysis is the fact that operating a Brain-Computer Interface with a machine-learning classifier requires a previously trained classifier. Therefore, either training data from the same subject must have been recorded in advance, or the classifier must have been trained on data from different subjects as proposed in section 6.5.

In the following, several considerations are outlined concerning designing and driving the soft- and hardware of an online P300-based BCI system.

Access Data from the Amplifier

In order to become able to process the data online, it is necessary to directly record the data from the EEG amplifier with the BCI program. With the *Neuroscan Synamps 5083* EEG amplifier used for the offline experiments in this thesis, data can be accessed only indirectly from the proprietary data acquisition program `Acquire`.

The setup for the BCI system in this thesis employs 10 channels, such that the EEG amplifier should also offer at least this number of channels. Some additional channels for, e.g., EOG acquisition (cf. section 2.3) would be useful.

Communication

In the previous chapter, the three operating systems `MS-DOS`, `Windows 98`, and `Linux` were used for the different tasks *stimulus presentation*, *data acquisition*, and *data analysis*, respectively. Within an online device, data analysis would need to be performed immediately after the data acquisition, and the classification results should be presented within the stimulus presentation environment. One way to perform this task is to establish communication channels between the three different operating systems, which can be very nasty with `MS-DOS`. On the other hand, one could try to employ only one operating system, and preferably even only one computer program. This would further allow to use a single computer (even without virtual operating systems), making it also easier to transport the system.

Data Preprocessing

As introduced in section 5.2, preprocessing in the setup of this thesis consists of *band pass filtering*, *dimensionality reduction* and *scaling*.

Utilizing the Fast Fourier transform allows to construct efficient band pass filters as is detailed out in section 5.2.1. The offline experiments have shown that reducing the dimensionality using Principal Component Analysis to a degree that 99.9% data variance is still captured, does not negatively affect the classification results (see sections 6.2 and 6.3). The experiments further indicated that such a PCA matrix can be calculated from a set of subjects and then applied to different subjects. Therefore, from the data of the offline experiments, a *general PCA* can be computed and subsequently be used for dimensionality reduction purposes in new subjects in the online context. No expensive calculations besides simple matrix multiplications with the general PCA

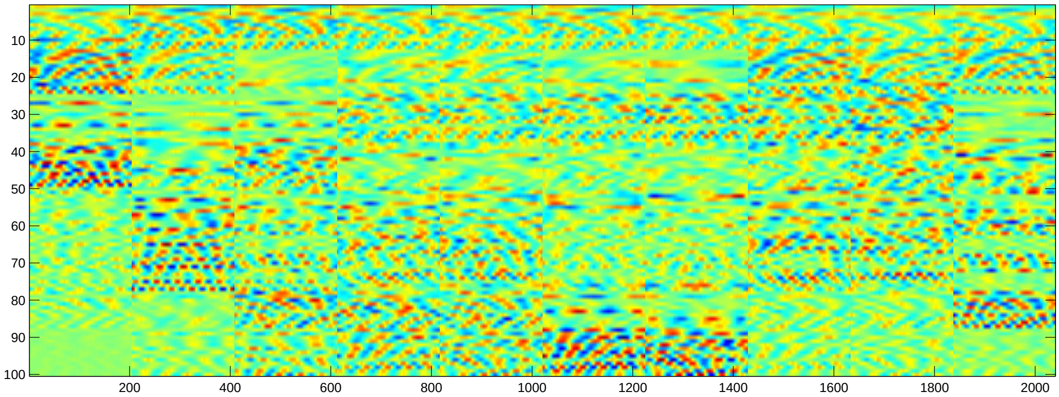


Figure 7.2: Transpose of the extraction of 2040 × 100 dimensions from the general PCA matrix as derived from the collection of the eight subjects in section 6.5.1. The input dimension was linearly expanded from 1600 to 2040 because the sampling frequency raised from 200Hz to 256Hz from offline to online analysis. This matrix reduces 2040-dimensional input vectors to 100-dimensional feature vectors.

matrix are required. The general PCA matrix as derived from the eight subjects in section 6.5.1 is depicted in Figure 7.2. Employing lower-dimensional feature vectors decreases subsequent computational costs and allows to utilize FLDA, which is computationally less expensive than, e.g., Support Vector Machines. Offline experiments nevertheless revealed similar performances for the FLDA compared to Support Vector Machines as is also discussed in section 6.4.

Classification Procedure

The classification procedure should be implemented under consideration of the temporal constraints for training and testing as well. While the computational costs were not very important for offline analysis, it would be desirable to have a classifier with only little computational demands to make classifications fast and to become able to provide a direct feedback about the classification results. This constraint can be realized by employing the FLDA classifier, which has proven to yield high accuracies for this domain in the previous chapter.

Rather than performing a parallel continuous analysis in real-time as it is, e.g., necessary for motor imagery devices (cf. section 3.2.4), it is sufficient to perform classifications *after* each trial for the P300 speller device. Before being able to perform classifications with a machine-learning classifier, such a classifier needs to be trained first. Section 6.5 has shown that it is possible to train a machine-learning classifier on data from a set of subjects, apply it to new subjects and still achieve reasonable results. On the other hand, information transfer rates experience substantial decreases when using this strategy. In order to perpetuate good classification results, training data from the *same* subject should preferably be employed for classifier training.

Trigger and Jitter

It is necessary to augment the EEG time series with information about latencies and types of the presented visual stimuli. In the previous experiments, such *trigger signals* were realized by transmitting this information via the serial port to the EEG amplifier.

Thereby, it is important to keep fluctuations (jitter) in temporal distance between stimulus onset and the temporal position of the according trigger signal low.

Further Requirements

To a large extent, the hardware should consist of standard components or at least be compatible with those. A portable system would be preferable, such that it could be tested and used in environments outside the laboratory. The experimental environment should be chosen with the goal to avoid artifacts. In the optimal case, the room should be sound attenuated and shielded for electromagnetic influences. On the other hand, it would also be desirable to make it possible to drive the BCI system in an environment which is not shielded, allowing to use it in common rooms.

7.2 Hardware Environment

From the considerations in the previous section, specific requirements to the hardware environment can be derived. First of all, it is necessary to directly record the EEG data with own programs. Thus, software interfaces for a common programming language, preferably C++, should exist or easy to be implemented. Second, to make the system portable, it would be desirable to work with a conventional laptop. In order to achieve a high signal quality, impedances are to be kept low (cf. section 2.3). Therefore, a way for assessing impedances is needed - either the EEG amplifier itself should offer such a possibility or an external device is to be utilized for this purpose. It is sufficient to have an amplifier with a limited number of channels since data from only 10 channels are acquired in this experimental setup. In the following, the hardware components engaged for driving the Brain-Computer Interface which satisfy these criteria are presented.

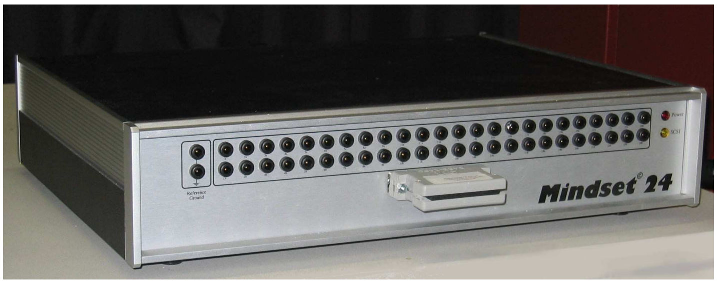


Figure 7.3: The *Mindset24* EEG amplifier with 24 differential input channels.

EEG Amplifier

The EEG amplifier *Mindset24*, depicted in Figure 7.3, was engaged in this BCI setup. It offers 24 differential input channels which second channels can be interconnected to measure electrode signals with respect to one common reference (cf. section 5.1). The data from the amplifier are transmitted via a SCSI port, allowing to operate the EEG amplifier with a conventional laptop when using, e.g., a SCSI PCMCIA adapter. For this purpose, the *Adaptec APA-1460D* PCMCIA adapter was utilized. Data from the EEG amplifier can be recorded with a frequency of up to 512Hz. The system can either be operated with single electrodes or a cap. The device comes with some basic C sources providing direct access to the data in own programs under the Windows operating system. Since no distinct trigger channel is provided with the amplifier, the software needs to be carefully designed with respect to this topic to gain EEG signals with only a small jitter.

Computer

A *Samsung X30* laptop with an Intel Pentium M 1.5 GHz CPU and 512MB RAM with a 64 MB *NVIDIA Ge ForceFX 5200 go* graphic card was chosen. The laptop is reasonably silent ($\leq$0.7 sone), offers a 15.4" WXGA TFT display with a maximal resolution of 1280 × 800 pixels and a brightness of up to 150cd/m^2.

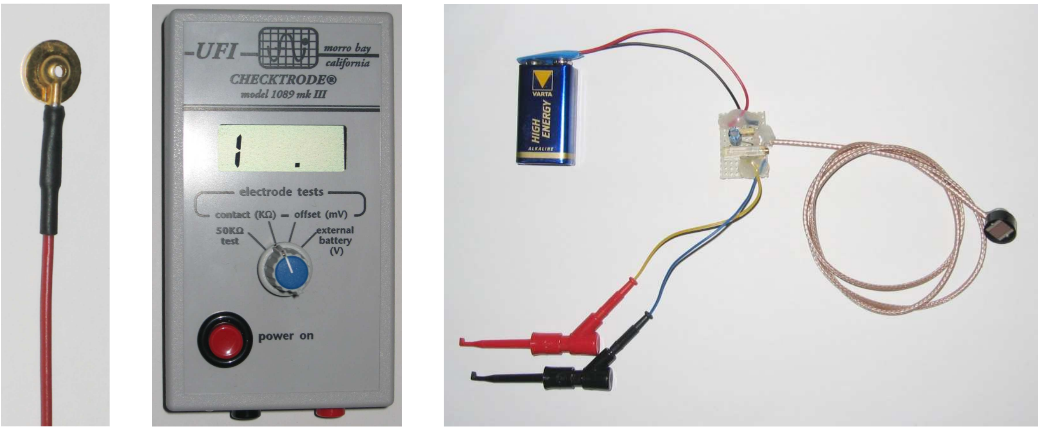


Figure 7.4: **Left:** Gold cup (Au) electrode as used in the experiment in this chapter. **Middle:** *Checktrode 1089 mk III* electrode tester for assessing impedances. **Right:** Circuit with an adjustable photo resistor for measuring temporal relationships between screen presentations and EEG recordings.

Accessories

Single gold cup (Au) electrodes as depicted in Figure 7.4 (left) were used for data acquisition from the scalp (see also Figure 7.8). For measuring impedances, the device *Checktrode 1089 mk III* (Figure 7.4, middle) was employed. An electrical circuit with an adjustable photo resistor (see Figure 7.4, right) was constructed to measure synchronization between stimulus presentations and data acquisitions from the EEG device[1]. The tool transforms light signals from the computer screen into electrical currents which can be delivered to the EEG amplifier. The EEG amplifier can be calibrated with the *Mindset Calibrator*, a device which simultaneously provides a precise 16Hz, 50μV signal to all channels, used to normalize signals derived from the channels. For grounding the subjects, a wrist strap was employed. It turned out that the refresh time of the laptop's TFT screen was too high and not constant due to the *rising and falling* of the LCD pixels, such that an external 20" CRT computer screen (*SONY Multiscan 20sf II*) was used instead for stimulus presentations. Thus, in this point, the specifications of the previous section could unfortunately not be fulfilled, and it is not sufficient to employ the display of the laptop at present. Today's high-quality stand-alone TFT displays offer better response times (of down to below 4ms), and if laptop displays will also be designed with comparable displays, the system could be operated without the CRT

[1] The author thanks Risto Kõiva and Oliver Lieske for this indispensable tool.

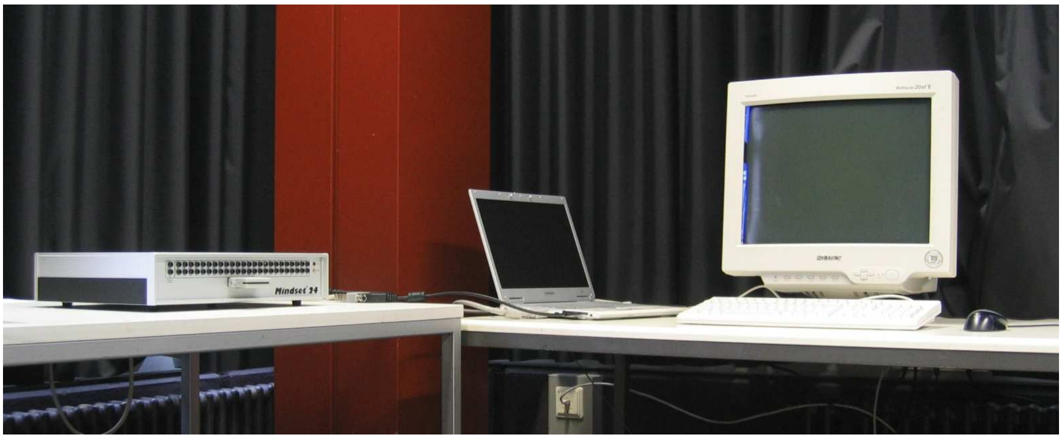


Figure 7.5: Experimental setup with EEG amplifier, laptop and computer screen.

screen. The whole system including the EEG amplifier, the laptop and the CRT screen is depicted in Figure 7.5.

7.3 Software

As for the hardware environment, requirements can also be derived from the specifications of section 7.1 for designing the software.

To begin with, it is necessary to have a system which acquires the data from the EEG amplifier and *synchronously* steers the stimulus presentation. Thereby, the point in time for stimulus presentation and recording the according EEG time series should be the same. Differences between the trigger event, i.e., the time marker for stimulus presentations, and the onset of the according EEG time series result in a jitter.

For data analysis, the processing stages *preprocessing* and *classification* together with *symbol inference* must be incorporated (cf. chapter 5). Preprocessing includes *band pass filtering*, *dimensionality reduction*, and *scaling*. Classification is performed with the FLDA classifier, since it is fast, simple and yields similar performances compared to Support Vector Machines (cf. section 6.4). The system was realized with the C++ compiler `gcc 3.3.1` for Windows XP, and the basic tasks of the system are divided into the three major modules `Graphical User Interface (GUI)`, `Communication`, and `Classification` which will be explained in the following. An overview of the communication processes between the modules is given in Figure 7.6.

7.3.1 Graphical User Interface (GUI)

One task of the `GUI` module is to provide an input mask for adjusting experimental variables concerning aspects of the subject, parameters for stimulus presentation and for controlling the EEG amplifier (see Figure 7.7). The module further organizes and steers the presentation of the stimuli as well as the results for the subject. For the latter purpose, it receives the classification results from the `Classification` module after each trial.

The experimental sequence is controlled by the `GUI` module: By simultaneously starting the presentation and the EEG recordings within separate threads, a trial is initiated.

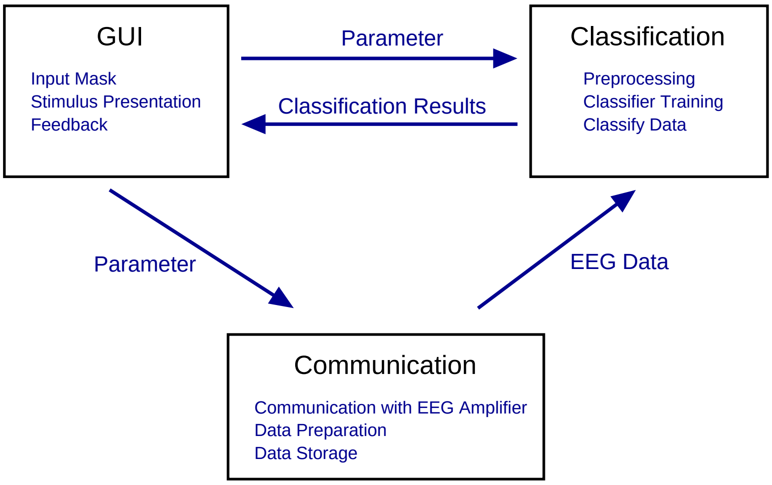


Figure 7.6: Tasks and communication processes of the modules `GUI`, `Classification`, and `Communication`.

Trigger information, i.e., the sequence of flashings and their temporal occurrences is furthermore computed for each epoch within this module. With this concept a jitter of about 1ms is achieved. The `GUI` module includes a port of `QT 3.3.4` for Windows (Trolltech, 2006).

7.3.2 Communication

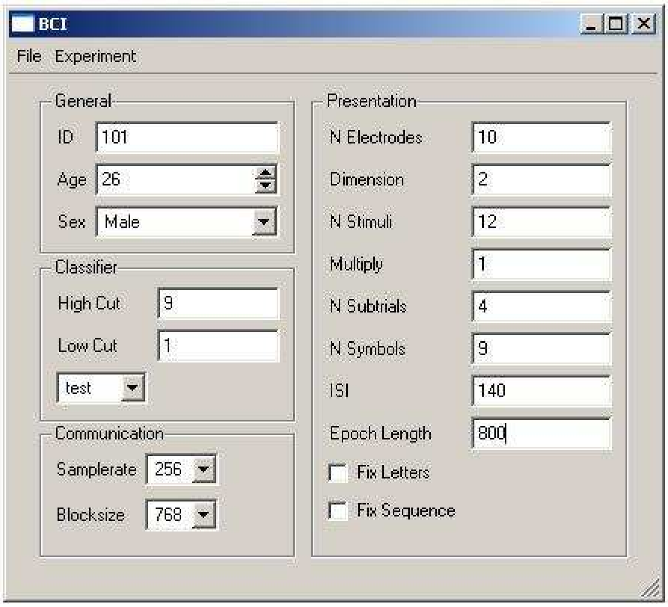
The module `Communication` initializes and controls the EEG amplifier and reads data from this device via the SCSI port. From the module `GUI`, it receives commands for these actions as well as parameters like sampling frequency and trial duration for steering the amplifier. Data are written into a file, dissected into epochs of 800ms time series, which are transmitted to the `Classification` module.

7.3.3 Classification

The `Classification` module is adjusted by parameters from the `GUI` module and receives EEG data from the `Communication` module. It incorporates the processing stages *preprocessing*, *binary classification* and *symbol inference*.

Preprocessing consists of band pass filtering, scaling and dimensionality reduction. As it is described in section 5.2.1, band pass filtering is realized by Fourier transforms using `FFTW 3.0.1` (Frigo and Johnson, 2006) allowing a brickwall filter operation in the frequency space. The dimensionality of the 2040-dimensional input data vectors is reduced to 100 dimensions by multiplying the input vector with the general PCA matrix (cf. section 6.5.1, see Figure 7.2), and scaled to an interval of [-1,1].

The `Classification` module is capable to perform train as well as testing. It employs the FLDA classifier, scans a number of values for its hyperparameter b in the training section and performs a 5-fold cross-validation to find an appropriate value for b. In contrast to chapter 6, the values for b are now evaluated in 500 equidistant steps in between the means of both classes (see also section 5.3.3). If the histograms of both classes do not overlap (which is seldom), the mean of the inner edges of both histograms



General
ID 101
Age 26
Sex Male

Classifier
High Cut 9
Low Cut 1
test

Communication
Samplerate 256
Blocksize 768

Presentation
N Electrodes 10
Dimension 2
N Stimuli 12
Multiply 1
N Subtrials 4
N Symbols 9
ISI 140
Epoch Length 800
□ Fix Letters
□ Fix Sequence

Figure 7.7: User interface for adjusting parameters of the BCI system.

is selected for b and no scanning for b is conducted. Binary classification is performed as, e.g., described in section 6.2.4 and it is the basis for the subsequent symbol inference, which is computed as described in section 5.4. The classification results are transmitted to the `GUI` module. Data objects and the classification framework are included from an in-house machine-learning library.

7.4 Online Classification with the P300 Speller Paradigm

In this section, the hardware environment and the developed software were engaged to perform a BCI experiment with online classification. For this purpose, seven subjects (age 20-30), denoted as 3A to 3G, participated in the experiment, each conducting two blocks: A training block of 50 trials with 5 subtrials each, and a test block of 90 trials with altogether 414 subtrials (see below). Within the test block, the word `INTERFACE` should be spelled by the subjects using the online Brain-Computer Interface. For that purpose, the number of subtrials employed in a trial for symbol inference was systematically decreased from 10 to 1 subtrials (omitting 9)[2] such that a trial lasted between 1.68 seconds (1 subtrial) and 16.8 seconds (10 subtrials) and the 90 trials result in 414 subtrials. Thereby, it is important to keep in mind that it becomes hard for the subject to prevent blinking and eye movements with a prolongated trial duration. In between the two blocks for training and testing, the FLDA classifier was trained on a balanced set of the training data. Only the FLDA classifier was employed, granting fast training and testing procedures.

The Interstimulus Interval was set to 140ms. Data were recorded from the scalp sites $Fz, Pz, Cz, C3, C4, P3, P4, PO7, PO8$, and Oz and recorded with a sampling rate

[2] It was initially planned to perform trials with 10,8,6,5,4,3,2, and 1 subtrials, but after performing 10 and 8 subtrials with the first subject, it turned out that a higher resolution would be desirable.

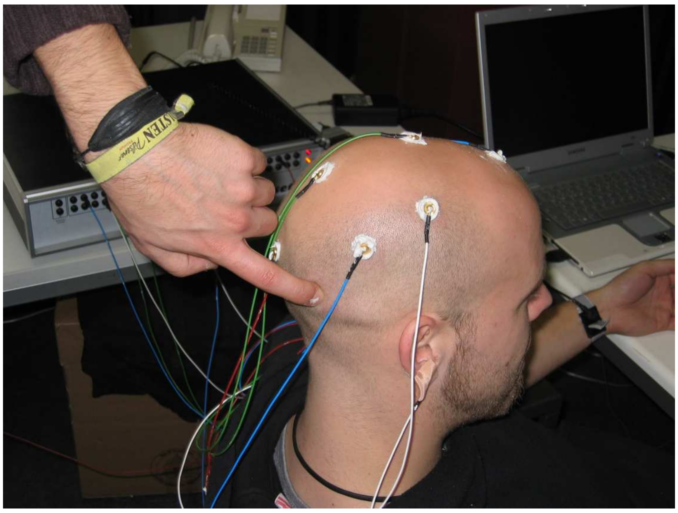


Figure 7.8: Electrode application on a subject with easy scalp access.

of 256Hz (see Figure 7.8 for an example of electrode application and Figure 6.2 for locations). Impedances below $2k\Omega$ were anticipated, which could not be achieved in all cases. Nevertheless, impedances always stayed below $5k\Omega$. Epochs of 800ms, i.e., 204 sampling points, were extracted from each of the 10 channels, resulting in data vectors of 2040 samples for an epoch. The data were band pass filtered within the range 0.5-8Hz and normalized to the interval [-1,1]. Afterwards, under usage of the general PCA matrix as derived from offline experiments (cf. section 6.5.1, see Figure 7.2), the data's dimensionality was reduced down to 100 dimensions. The input dimensionality of the PCA matrix was linearly expanded to 2040 dimensions because the data in this experiment was sampled with 256Hz, instead of 200Hz as in the offline sections.

7.4.1 Preliminary Analysis: Offline Binary Classification

Methods

A binary classification of the training set was performed as a preliminary analysis to become able to compare the results with those obtained from offline analysis. For this purpose, the data were analyzed analogously to offline analysis: The 1000 data samples were divided into two halfs of balanced $\mathcal{P}^+$ and $\mathcal{P}^-$ data. On one set, parameter optimization was performed in a 5-fold cross-validation, and on the other set, classification performance was assessed (see section 6.2.4 for details).

Results

Binary offline classification results are listed in Table 7.1. In the mean, a binary classification accuracy of 0.796 ± 0.056 was achieved using the FLDA classifier on the 100-dimensional feature vectors. While for the best performing subject an accuracy of 0.880 was reached, the worst performing subject yielded an accuracy of only 0.710.

Discussion

Compared to the offline results from section 6.5.2, the classification results for binary classification decreased from a mean accuracy of 0.844 to a mean accuracy of only

Table 7.1: Binary classification accuracies obtained within the new environment using the FLDA classifier on 100-dimensional feature vectors as computed under the usage of the general PCA for subjects 3A to 3G.

	Subject							
Method	3A	3B	3C	3D	3E	3F	3G	**Mean**
FLDA	0.798	0.818	0.710	0.744	0.880	0.788	0.832	0.796±0.056

0.796. Different factors can be responsible for this outcome. First, it is possible that the subjects did not perform as well as those used in chapter 6. Second, the experimental environment provided more noise than the previous one because the room was neither sound attenuated nor shielded from electromagnetic influences in the online experiment (cf. section 2.3). However, the results are on a level that should allow to perform online classifications.

7.4.2 Online Analysis

Methods

In the online analysis block, the subjects were instructed to spell the word `INTERFACE` with varying numbers of subtrials in a trial. The first number of subtrials was 10, and then the number of subtrials was decreasing from 8 to 1. In advance of each trial, the experimenter told the subject the particular symbol to spell. From the recorded EEG data of a trial, the classifier computed a symbol and presented it immediately after each trial on the computer screen.

Results

For spelling the word `INTERFACE`, Figures 7.9 and 7.10 reveal the classification results for accuracies and information transfer rates, respectively.

In the mean, an accuracy level of 80% was exceeded when using 4 subtrials ($p_{\text{acc}} = 0.810$). However, employing 5, 6 and 7 subtrials yielded lower accuracies ($p_{\text{acc}} = 0.683$, $p_{\text{acc}} = 0.762$, and $p_{\text{acc}} = 0.778$). Only when using 8 subtrials, 80% accuracy was reached again ($p_{\text{acc}} = 0.825$). Table 7.2 exposes the letters spelled by the different subjects when employing 4 subtrials. For single subjects, 80% accuracy was reached after 2 subtrials ($p_{\text{acc}} = 1.000$, subject 3E). Accuracies for the worst performing subject never exceeded 80% accuracy within the 10 subtrials ($p_{\text{acc}} = 0.778$ after 8 subtrials, subject 3D).

The best *mean* information transfer rate was 32.17 bits/min (4 subtrials, $p_{\text{acc}} = 0.810$). Table 7.3 exposes the highest information transfer rates from each of the 7 subjects. The best performing subject achieved 92.32 bits/min (subject 3E), and the second best performing subject reached 61.55 bits/min (subject 3G). The highest information transfer rate for the worst performing subject (subject 3C) was 22.70 bits/min and 18.46 bits/min under the precondition $p_{acc} > 0.8$. The mean of the highest information transfer rates was 46.71 ± 24.07 bits/min, and with the precondition $p_{acc} > 0.8$, it was 42.63 ± 26.46 bits/min.

Discussion

Compared to the offline experiments in chapter 6.5.2, information transfer rates were low: For exceeding 80% *mean* accuracy in *offline* experiments, transfer rates of

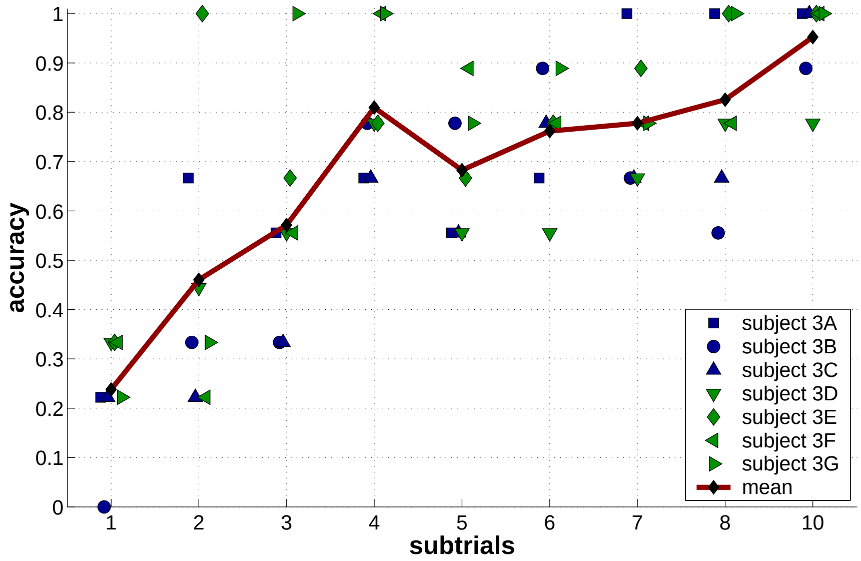


Figure 7.9: Classification accuracies obtained for spelling the word `INTERFACE` with different numbers of subtrials employed for spelling a symbol. The red line reflects the mean accuracies of the 7 subjects.

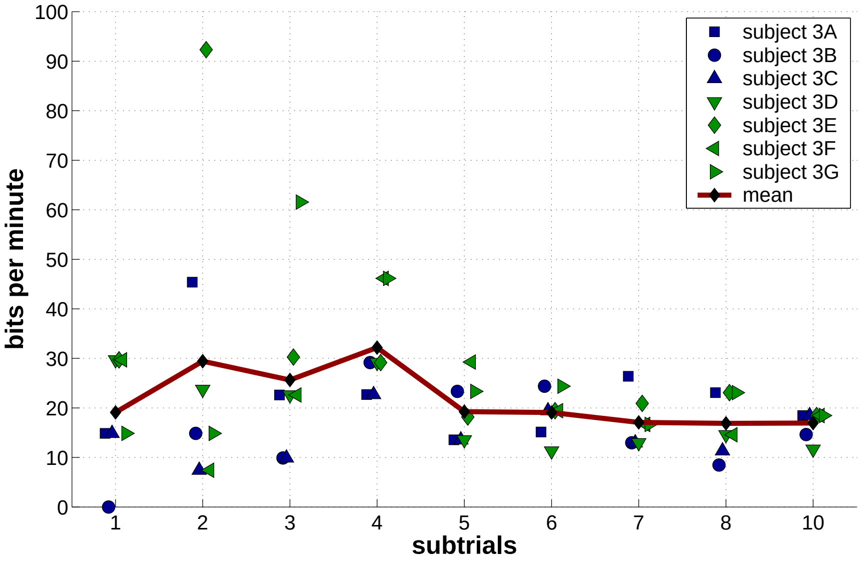


Figure 7.10: Information transfer rates for spelling the word `INTERFACE` for different numbers of subtrials. The red line reflects the mean transfer rates.

Table 7.2: Results produced by the individual subjects for spelling the word `INTERFACE` employing 4 subtrials in the online block.

Subject	**Inferred Symbols**	$p_{\mathbf{acc}}$
3A	`ONTEFEACE`	0.667
3B	`IMTERLACE`	0.778
3C	`CNSERFACE`	0.778
3D	`HNZERFECE`	0.667
3E	`INTERFHIE`	0.778
3F	`INTERFACE`	1.000
3G	`INTERFACE`	1.000

Table 7.3: Highest information transfer rates in bits/min for each of the 7 subjects for spelling the word `INTERFACE` in the online block. The second row exposes the highest information transfer rates under the precondition $p_{acc} > 0.8$. Note that subject 3D did not reach 80% accuracy, such that $p_{acc} = 0.778$ was employed in the second row for this subject.

	Subject							
Method	3A	3B	3C	3D	3E	3F	3G	**Mean**
FLDA	45.39	29.16	22.70	29.72	92.32	46.16	61.55	46.71±24.07
FLDA ($p_{acc} > 0.8$)	26.38	24.39	18.46	(29.16)	92.32	46.16	61.55	42.63±26.46

44.98 bits/min resulted with the FLDA classifier and a disjoint PCA matrix. In contrast, only 32.17 bits/min were achieved in the *online* experiment in this condition. A similar picture can be drawn for the *best* information transfer rates: The best information transfer rate in the offline experiment using FLDA was 120.17 bits/min. This rate decreased to 92.32 bits/min in the online experiment.

Thus, the findings from the online block are in line with the results obtained for the previous binary offline classification. Therefore, the reasons for the lack in performance are likely to be the same as already discussed for binary classification: Foremost, an enhanced noise due to the experimental environment and the performances of the subjects are possible sources for the performance deficits. But by not conducting the experiment in a special environment, comparable results can be expected when driving the Brain-Computer Interface in common rooms in a house or a hospital.

Despite the decrease in accuracies and transfer rates from offline to online analysis, the information transfer rates obtained in online classification are still high compared to other approaches as shown in section 3.3.

Although the results are convincing, these outcomes must be interpreted with care for several reasons: First, the number of 9 symbols predicted in this experiment in each subtrial condition for each subject is probably too low to allow general conclusions. Second, it is difficult to incorporate the findings of, e.g., Wolpaw and McFarland (2004) in the comparisons, since information transfer rates in bits/min might not be an appropriate measure for this two-dimensional device for continuous movements (cf. section 3.2.4). Third, delays between trials were omitted for computing the information transfer rates. This calculation is quite common (Donchin et al., 2000; Serby et al., 2005) but overestimates the speed of the devices. For example, assuming an average delay of 2 seconds between trials, each employing 4 subtrials would yield a decrease in information transfer rates of 22.94%. Thus, the mean information transfer

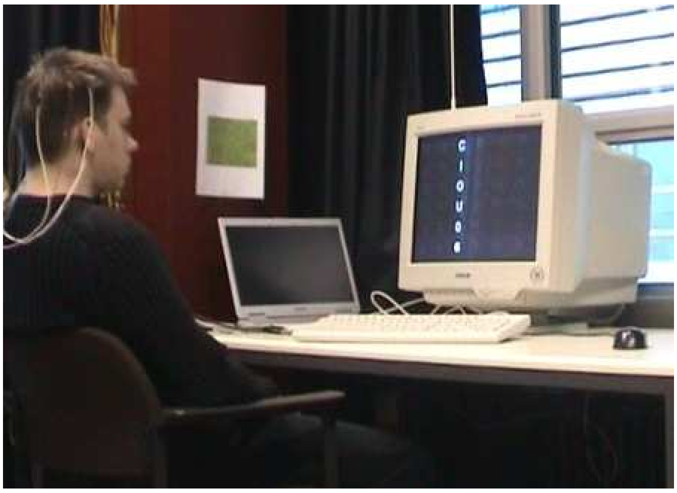


Figure 7.11: Subject operating the online Brain-Computer Interface.

rate of 32.17 bits/min would decrease to 24.79 bits/min. The best information transfer rate of 92.32 bits/min would even decrease to 57.88 bits/min. Fourth, information transfer rates should not solely be employed for evaluating the performance of a Brain-Computer Interface. Different devices are appropriate for different application purposes. Compared to other approaches, the P300 speller paradigm would probably not be very suitable for, e.g., steering a wheelchair since it does not provide a continuous signal and depends on stimulations. On the other hand, the P300 speller device can be valuable for producing text or controlling an internet browser. Therefore, the different approaches can complement each other and it is desirable to achieve improvements within each of the different approaches.

For the approach under investigation, the P300 speller paradigm, it can be stated that the presented classification strategy in this thesis has yielded high performances in an online version while employing rather simple classification techniques.

7.5 Summary

Within this chapter, the conclusions derived from the *offline* experiments in the previous chapter were exploited for constructing an *online* P300-based Brain-Computer Interface. This interface was built from the scratch such that hardware had to be set up carefully and software needed to be written to drive the BCI with respect to special requirements to the system. Most important, it was necessary to directly access the data from the amplifier to allow immediate data analysis and feedback after each trial. Furthermore, precise synchronizations of stimulus presentations and EEG recordings in order to avoid jitters were inevitable.

One of the constraints for an online BCI is that the data analysis procedure needs to operate fast. This prerequisite is fulfilled in the presented software by employing the efficient Fast Fourier transform for band pass filtering. Furthermore, dimensionality reduction is realized by using the general PCA as derived from the offline experiments. Finally, classification is conducted in an efficient way by using Fisher's Linear Discriminant, which was shown to yield similar results as the Support Vector Machine in the previous chapter. After designing the BCI system for driving the P300 speller paradigm, an own experiment was performed with this device.

It was possible to drive the system in an online fashion and achieve mean theoretical information transfer rates of 32.17 bits/min ($\approx$ 6 symbols per minute), which is lower than in the offline analysis, but still competitive to other systems. When considering a delay of 2s between trials, this rate decreases to 24.79 bits/min ($\approx$ 4 symbols per minute). In the best case, a theoretical information transfer rate of 92.32 bits/min ($\approx$ 17 symbols per minute) was achieved, which is reduced to 57.87 bits/min ($\approx$ 11 symbols per minute) with a 2s delay between trials.

A possible reason for the performance deficits compared to the offline experiments conducted in chapter 6 is the experimental environment. While, e.g., the experimental room was sound attenuated and shielded from electromagnetic influences in offline experiments, only a conventional room was used for online experiments. On the other hand, results obtained in the latter environment can thus better be generalized to common house or hospital environments.

Although high information transfer rates were achieved, these findings have to be interpreted with care because only a restricted number of inferences was performed.

Information transfer rates should not be the only measure for comparing Brain-Computer Interfaces. Rather, different Brain-Computer Interfaces are appropriate for different applications and should complement each other. For the case of the P300 speller device, a simple but powerful classification framework was developed.

Chapter 8

Conclusion

Within this thesis, an efficient data analysis procedure incorporating Machine-Learning techniques was developed for the P300 speller Brain-Computer Interface (BCI). First, in own *offline* experiments, data were produced which were subsequently engaged to optimize classification algorithms. Afterwards, based on these findings, an *online* BCI was designed, implemented and evaluated.

In the first chapters, basic aspects of Electroencephalography, Brain-Machine Interfaces in general and the P300 speller paradigm BCI in particular were introduced. Afterwards, data analysis techniques relevant for the P300 speller device proposed in this thesis were discussed. Then, in a series of experiments and accompanying analytical investigations, the data analysis procedure for the P300 speller paradigm was improved and finally incorporated in an online BCI. These steps and quintessential findings will be outlined in the following.

Better Classification Accuracies by Machine-Learning Techniques

An experiment with two subjects was conducted and after deriving adequate preprocessing parameters, it was investigated whether it is possible to increase the performance of common Model-Based approaches by employing Machine-Learning techniques. Furthermore, the impact of dimensionality reduction based on Principal Component Analysis (PCA) as a preprocessing step on classification results was examined.

It turned out that Machine-Learning techniques have yielded higher performances than the Model-Based approaches and that dimensionality reductions have not affected classification accuracies negatively when capturing at least 99.9% variance. No substantial performance differences for the three Machine-Learning classification techniques Support Vector Machine with Gaussian kernel (RBF SVM), Linear Support Vector Machine (LSVM), and Fisher's Linear Discriminant Analysis (FLDA) have been found.

Classification Improvements by Employing Multiple Electrodes

Inspired by the scalp distribution of Event-Related Potentials from EEG time series belonging to the (*oddball*) events which elicit a P300, instead of *one* electrode as in the previous investigation, a set of ten electrodes was chosen to constitute the feature vector. The ability of Machine-Learning techniques to adapt to a data structure allowed to use these new feature vectors with only minor modifications. In contrast, for the Model-Based methods, the model assumptions would need to be augmented. Therefore, these methods were not employed any more and only the performances of the three Machine-Learning classifiers were compared with different (PCA-reduced) dimensions.

The experiment indicated that strong increases in classification performance were achieved when using data from ten channels. Similar to the previous investigations, the three Machine-Learning classification techniques reached almost the same performance, and PCA-based dimensionality reduction capturing 99.9% variance had no negative impact on the classification results.

Information Transfer Improvements by Enhancing the Presentation Speed

After these improvements, the speed of the BCI was still unsatisfying such that the temporal distance of two consecutive stimuli was reduced from 500ms to 140ms. Furthermore, to allow for more general assertions, eight subjects were employed. More than 99.9% data variance was captured within the first 100 Principal Components of each subject, such that only projections of the EEG data on these 100 Principal Components were subsequently employed as feature vectors.

As an outcome, although the temporal distance between events decreased, resulting in less pronounciated P300 components due to overlaps of consecutive time series of EEG signals (cf. section 4.1), comparable high classification accuracies were achieved with the Machine-Learning classifiers. All three classification methods yielded mean theoretical information transfer rates of about 70 bits/min (disregarding delays between trials). The best subject achieved up to 136.77 bits/min.

Generalizations to New Subjects

Two directions of generalization capabilities were investigated. Inspired by a high degree of similarity of the first Principal Components between subjects, it was examined whether it would be possible to employ a *general* PCA matrix for dimensionality reduction which was trained on data from a set of subjects and applied for dimensionality reduction to data from new subjects. The same strategy was then pursued for the Machine-Learning classifiers which were also trained on a set of subjects and applied to a new subject. Success with such a strategy would allow to use a pretrained PCA matrix and/or classifier for a new subject, avoiding the requirement to record data from the individual subject for classifier training.

The results indicate that using a general PCA matrix did not affect the classification results negatively. In contrast, accuracies decreased when using a Machine-Learning classifier which was not trained on data from the individual subject.

Performance in the BCI Competition 2003

An early version of the classification method was applied on data provided with the BCI Competition 2003 (Blankertz et al., 2004). The goal of the competition was to let different algorithms compete for best classification accuracies. For this purpose, a labeled data set of EEG data as recorded during a BCI experiment was published which could be used for classifier training. For another unlabeled data set, the algorithms should infer the labels on the basis of the EEG data. The author's contribution managed to find the correct labels with the least numbers of trials among the contributions for the P300 speller domain.

Constructing and Driving an Online BCI

The results of the offline analyses were utilized for designing an online BCI. Particularly the findings that a general PCA can be employed and that it is sufficient to employ a rather simple classifier like Fisher's Linear Discriminant were valuable for constructing the online system. After defining hard- and software specifications, the online BCI system was built from the scratch. Most important, direct access to EEG data from the amplifier in own programs had to be performed, and an exact synchronization of the amplifier and the stimulus presentation was mandatory. Additionally, a fast classification was desirable which could be realized by employing Fisher's Linear Discriminant classifier and the previously computed general PCA matrix for dimensionality reduction. With this system, an online experiment with seven subjects was conducted. Each subject performed 50 training trials on which the classifier was trained and, afterwards, operated the BCI device for spelling the word `INTERFACE` with varying numbers of subtrials for each letter.

In the online experiment, theoretical information transfer rates of 32.17 bits/min (mean) and 92.32 bits/min (best) could be achieved, which are competitive compared to existing P300-based approaches. Considering a delay of 2s between trials, these rates decrease to 24.79 bits/min ($\approx$ 4 symbols per minute) and 57.87 bits/min ($\approx$ 11 symbols per minute), respectively .

The rates obtained in the online experiment are lower than in offline experiments, which might be caused by a lack of sound attenuation and shielding from electromagnetic influences in the laboratory. On the other hand, the experimental conditions are therewith closer to a real-world environment as provided in rooms in a normal house, such that the system could presumably also work there with a comparable performance.

Although promising information transfer rates have been obtained, the online results rely on a limited set of only 9 sample letters for each subject, such that general conclusions can only be drawn with care. Furthermore, the information transfer rate should not be the only basis for comparing BCIs. Depending on the purpose of the system, other variables like the question whether the interface depends upon stimulation or not, or whether it provides a continuous signal are also important.

However, it can be stated that for the P300 speller device an efficient and powerful classification strategy was found within this thesis which has also yielded high information transfer rates in an online experiment.

In summary, the investigations conducted in this thesis succeeded in enhancing the classification performance of a P300 speller Brain-Computer Interface and, therewith, its speed for transferring information. Additionally, the generalization capability of the proposed method to new subjects has been demonstrated. The findings were incorporated in an online BCI system which was constructed from the scratch and for which own software was written. This BCI proved that the classification strategies derived from the offline experiments also work well under real-world circumstances, and still allow for high information transfer rates.

8.1 Outlook

After designing an online Brain-Computer Interface with the P300 speller paradigm allowing for high classification accuracies, further directions in which the performance of the system could be improved can be investigated. Additionally, one can include the proposed P300 recognition mechanism for alternative applications.

Enhancing the Performance of the P300 Speller Device

A first attempt to further improve the speed of the current P300 speller device would be to employ a variable number of subtrials. Rather than reaching different accuracies with a predetermined constant number of subtrials, the mechanism can be reversed, such that a specific level of certainty could be defined which is to be reached with a variable number of subtrials. For this purpose, classification results are to be computed after each subtrial, providing a basis for deciding whether further subtrials need to be performed. Since efficient strategies for band pass filtering (Fast Fourier transform), dimensionality reduction (general PCA matrix) and classification (FLDA) were employed, fast classifications would be possible[1].

A way to improve the spelling of *words* would be to employ a T9-like system as commonly known from the Short Messaging System (SMS) for cell phones. By exploiting the facts that only a limited number of words are commonly used in conversations and that neighboring letters of a word are probabilistically related, a word completion algorithm could be designed, such that after, e.g., the first three letters of a word were determined, a selection of possible words could be presented as stimuli. Similar approaches for a BCI were already included for other BCIs (Hinterberger et al., 2004a; Pfurtscheller et al., 2003). Note that a prerequisite would be a perfect classification of the three letters. A similar approach would target into another direction: Imperfect classifications of single letters in a word could be compensated by calculating the minimum distance to words in a dictionary. If still ambiguities exist, the possible words could be presented as choices after the initial spelling. Furthermore, it could be considered for this strategy that misclassifications mostly result in symbols from the same row or column. For example, subject 3C in section 7.4 spelled the word `CNSERFACE` instead of `INTERFACE`. The letter `C` is in the same column as `I`, and the letter `S` lies in the same row as `T`, such that the correct word `INTERFACE` could easily be determined.

Section 6.5 revealed that it is possible to construct a *general* classifier which was trained on data from a number of subjects and subsequently applied to new subjects. With such a classifier, no training for the individual subject is necessary and the subject does not need to perform training trials. Unfortunately, compared to classifiers adapted to the individual, only minor performance was achieved with this strategy. But to avoid exhausting training sessions, one could combine both classifiers: Initially, a general classifier could be used to achieve classifications even after the first trials. Therewith, data samples from the individual can be collected and a classifier can be trained on *that* data. A large number of subtrials would initially be needed to guarantee valid labels. Since the computational inexpensive classifier FLDA proved to produce high performance results, such an online training would consume only little computational costs and could easily be performed in the background. In Machine-Learning, gradually increasing the number of training samples is denoted as *online learning*, and several publications can be found regarding this topic (Schoelkopf and Smola, 2002). While operating the BCI device, the classification strategy can gradually be faded from the general to the individual classifier. This procedure could also be performed for an individual subject among sessions to compensate for intraindividual variations. By using a dynamic number of subtrials, relying on reaching a certainty level for symbol inferences as described above, no further control of the numbers of subtrials would be necessary. This strategy would result in a continuously adapting self-accelerating BCI.

Another direction for improving the performance would be to include artifact elimi-

[1] Note that a delay of 660ms between subtrials would result because a time series of 800ms must be recorded for the last stimulus of a subtrial, of which 140ms are consumed for presentation.

nation techniques. For example, artifacts in the training samples could be detected and the according EEG time series be dropped for the training procedure. On the other hand, the impact of artifacts could be reduced by appropriate preprocessing techniques as e.g. suggested by Guan et al. (2004) and Jung et al. (2000).

Further Applications

A drawback of the current P300 speller device is that it relies on visual stimulation, which could lead to serious problem when working with locked-in patients which might loose control of their visual attention. As stated in section 2.5, the P300 component is modality independent and a first alternative to visual stimuli are auditory stimuli. However, transferring the matrix style of the P300 speller paradigm to auditory stimuli is a delicate task, since two dimensions of an auditory stimulus need to be varied independently. While it would be possible to e.g. change the volume and the frequency of an auditory stimulus in such a way, it would be hard for the subject to mentally construct the desired combination of volume and frequency to choose a symbol in the auditory matrix. Therefore, it would be more promising to employ single stimuli, e.g. spoken letters. The SD speller design as introduced in section 4.3 proved that such an approach could also work with reasonable speed (Guan et al., 2004). Beside using auditory stimuli, also tactile stimulation could be employed by e.g. mechanical or electrical stimulation of the finger tips.

Using the presented Machine-Learning classification framework, P300-based detection of deception as suggested by Farwell and Donchin (1991) and discussed in section 4.3 could possibly be enhanced. The Machine-Learning classifier could learn which EEG patterns result for familiar and unfamiliar stimuli without requiring a concept like e.g. MERMER. Furthermore, adaptations to the EEG patterns of the individual subjects would result when training trials with familiar/unfamiliar stimuli would be recorded from the specific subject and the Machine-Learning classifier would be trained on that data. The stimuli in the present online BCI software are stored as bitmaps, making it possible to easily substitute them by pictures from any scene.

Once equipped with an online classification procedure, real interactions are possible which allow for a broad range of applications: For example, as introduced by Mellinger et al. (2003) for the Slow Cortical Potential BCI, a webbrowser could be controlled with such a device: The links of a website could be flashed and induce a P300 component for the link the user wishes to follow. Furthermore, several kinds of games could be implemented. Obiously, for checker board games, which naturally expose a matrix style, the P300 speller approach could be utilized to select fields in the checker board. But also more dynamic interactions in virtual environments are possible as Bayliss (2003) has shown.

The subjects in the online experiment reported that it highly depends on their degree of attention whether a classification succeeds or not. Inspired by Lalor et al. (2004), the logic of the BCI could be reversed and used as a measure for attention. Combined with feedback, such a device could then be useful for e.g. children with attentional deficits (Holtmann et al., 2004). The degree of attention could be provided as a feedback for the children in a game with the goal to improve their attention.

Bibliography

Alfa, R. (2005). Brain-machine interfaces: Reinventing sensory and motor functions after injury or disease. *Saltman Quarterly*, 2(1).

Allison, B. and Pineda, J. (2003). ERPs evoked by different matrix sizes: implications for a brain computer interface (BCI) system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):110–113.

Anderson, C. (1997). Effects of variations in neural network topology and output averaging on the discrimination of mental tasks from spontaneous electroencephalogram. *Journal of Intelligent Systems*, 11:423–431.

Anderson, C. and Sijercic, Z. (1996). Classification of EEG signals from four subjects during five mental tasks. In Bulsari, A., Kallio, S., and Tsaptsinos, D., editors, *Proceedings of the Conference on Engineering Applications in Neural Networks (EANN'96)*, pages 405–414. Systems Engineering Association.

Anderson, V., Burchiel, K., Hogarth, P., Favre, J., and Hammerstad, J. (2005). Pallidal vs subthalamic nucleus deep brain stimulation in parkinson disease. *Archives of Neurology*, 62:554–560.

Babiloni, F., Cincotti, F., Lazzarini, L., and Marciani, M. G. (2000). Linear classification of low-resolution EEG patterns produced by imagined hand movements. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 8(2):186–188.

Bauby, J.-D. (1998). *The Diving Bell and the Butterfly.* Vintage, New York.

Bayliss, J. (2003). Use of the evoked potential p3 component for control in a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):113–116.

Bayliss, J. and Ballard, D. (1999). Single trial P300 recognition in a virtual environment. *Soft Computing in Biomedicine (CIMA).*

Ben-Shakhar, G. and Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *The Journal of Applied Psychology*, 88(1):131–151.

Bennett, K. and Campbell, C. (2000). Support vector machines: Hype or halleluya? *SIGKDD Explorations*, 2:1–13.

Berger, H. (1929). Über das elektroenkephalogramm des menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87:527–570.

Berger, H. (1931). Über das elektroenkephalogramm des menschen (teil 2). *Archiv für Psychiatrie und Nervenkrankheiten*, 94:16.

Birbaumer, N. (1990). *Physiologische Psychologie.* Springer, Heidelberg.

Birbaumer, N., Elbert, T., Canavan, A., and Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behaviour. *Physiological Reviews*, 70(1):1–41.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H. (1999). A spelling device for the paralysed. *Nature*, 398:297–298.

Birbaumer, N., Hinterberger, T., Kübler, A., and Neumann, N. (2003). The thought translation device (TTD): Neurobehavioral mechanisms and clinical outcome. *IEEE Transactions on Rehabilitation Engineering*, 11(2):120–123.

Birbaumer, N., Kübler, A., Ghanayim, N., Hinterberger, T., Perelmouter, J., Kaiser, J., Iversen, I., Kotchoubey, B., Neumann, N., and Flor, H. (2000). The thought translation device (TTD) for completely paralyzed patients. *IEEE Transactions on Rehabilitation Engineering*, 8(2):190–193.

Birbaumer, N. and Schmidt, R. (2005). *Biologische Psychologie.* Springer, Berlin.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

Blankertz, B., Dornhege, G., Schäfer, C., Krepki, R., Kohlmorgen, J., Müller, K.-R., Kunzmann, V., Losch, F., and Curio, G. (2003). Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11:100–104.

Blankertz, B., Müller, K.-R., Curio, G., Vaughan, T. M., Schalk, G., Wolpaw, J. R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schröder, M., and Birbaumer, N. (2004). The BCI competition 2003. *IEEE Transactions on Biomedical Engineering*, 51(6):1044–1051.

Bledowski, C., Prvulovic, D., Hoechstetter, K., Scherg, M., Wibral, M., Goebel, R., and Linden, D. (2004). Localizing P300 generators in visual target and distractor processing: A combined event-related potential and functional magnetic resonance imaging study. *Journal of Neuroscience*, 24:9353–9360.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Carmena, J., Lebedev, M. A., Crist, R., O'Doherty, J., Santucci, D., Dimitrov, D., Patil, P., Henriquez, C., and Nicolelis, M. (2003). Learning to control a brain-machine interface for reaching and grasping by primates. *Public Library of Science (PLoS) Biology*, 1:1–16.

Caspers, H. and Speckmann, E. (1974). Cortical DC shifts associated with changes of gas tensions in blood and tissue. In *Handbook of Electroencephalography and Clinical Neurophysiology*, pages A10–A65. Elsevier.

Caton, R. (1875). The electric currents of the brain. *British Medical Journal*, 2:278.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.* Software available at `www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapin, J., Moxon, K., Markowitz, R., and Nicolelis, M. (1999). Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience*, 2(7):664–670.

Cheng, M., Gao, X., Gao, S., and Xu, D. (2002). Design and implementation of a brain-computer interface with high transfer rates. *IEEE Transactions on Biomedical Engineering*, 49:1181–1186.

Cincotti, F., Mattia, D., Babiloni, C., Carducci, F., Salinari, S., Bianchi, L., Marciani, M. G., and Babiloni, F. (2003). The use of EEG modifications due to motor imagery for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):131–133.

Clark, G., Tong, Y., Martin, L., and Busby, P. (1981). A multi-channel cochlear implant. an evaluation using an open-set word test. *Acta Otolaryngol*, 91:173–175.

Clark, K., Naritoku, D., Smith, D., Browning, R., and Jensen, R. (1999). Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nature Neuroscience*, 2:94–98.

Coles, M. (1989). Modern mind-brain reading: Psychophysiology, physiology and cognition. *Psychophysiology*, 26(3):251–269.

Coles, M., Gratton, G., and Fabiani, M. (1990). Event-related brain-potentials. In *principles of psychophysiology*, pages 413–455. Cambridge University Press.

Coles, M., Gratton, G., Kramer, A., and Miller, G. (1986). Principles of signal acquisition and analysis. In Coles, M., Donchin, E., and Porges, S., editors, *Psychophysiology - System, Processes and Applications*, New York. The Guilford Press.

Coles, M. and Rugg, M. (1995). Event-related brain potentials: An introduction. In Rugg, M. and Coles, M., editors, *Electrophysiology of Mind*, Oxford. Oxford University Press.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301.

Cooper, R., Winter, A., Crow, H., and Grey, W. (1965). Comparison of subcortical, cortical and scalp activity using chronically indwelling electrodes in man. *Electroencephalography and clinical Neurophysiology*, 28:217–228.

Coyle, S., Ward, T., Markham, C., and McDarby, G. (2004). On the suitability of near-infrared (NIR) systems for next-generation brain-computer interfaces. *Physiological Measurement*, 25:815–822.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines.* Cambridge University Press, Cambridge, UK.

Cyberkinetics (2006). Cyberkinetics - neurotechnology systems, inc. Website, retrieved February 2, 2006, from http://www.cyberkineticsinc.com.

Cyberonics (2005). VNS therapy fact sheet. Website, retrieved June 27, 2005, from http://www.cyberonics.com.

Dobelle, W. (2000). Artificial vision for the blind by connecting a television camera to the visual cortex. *American Society of Artificial Internal Organs Journal*, 46:3–9.

Dobelle, W., Mladejovsky, M., Evans, J., Roberts, T., and Girvin, J. (1976). 'Braille' reading by a blind volunteer by visual cortex stimulation. *Nature*, 259:111–112.

Dobelle, W., Mladejovsky, M., and Girvin, J. (1974). Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science*, 183:440–444.

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425.

Donchin, E. (1981). Surprise! . . . surprise? *Psychophysiology*, 18:493–513.

Donchin, E., Spencer, K., and Wijeshinghe, R. (2000). The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8(2):174–179.

Donoghue, J. (2002). Connecting cortex to machines: Recent advances in brain interfaces. *Nature Neuroscience supplement*, 5:1085–1088.

Dornhege, G., Blankertz, B., Curio, G., and Müller, K.-R. (2004). Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993–1002.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification.* Wiley-Interscience Publication.

Duncan, D. and Friedman, R. (2005). Fernsteuerung durch gedanken. *MIT Technology Review*, 3:72–78.

Duncan-Johnson, C. C. and Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, 14:456–467.

Ebe, M. and Homma, I. (1994). *Leitfaden für die EEG Praxis: Ein Bildkompendium.* Gustav Fischer, Stuttgart.

Elbert, N. B. T., Rockstroh, B., Daum, I., Wolf, P., and Canavan., A. (1991). Clinical-psychological treatment of epileptic seizures: a controlled study. In Ehlers, A., editor, *Perspectives and promises of clinical psychology.*

Engel, A., Moll, C., Fried, I., and Ojemann, G. (2005). Invasive recordings from the human brain: Clinical insights and beyond. *Nature Neuroscience*, 6:35–47.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50.

Farwell, L. (2006). Brain fingerprinting - home. Website, retrieved February 4, 2006, from http://www.brainwavescience.com.

Farwell, L. and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(S2):510–523.

Farwell, L. and Donchin, E. (1991). The truth will out: Interrogative polygraphy (lie detection) with event-related brain potentials. *Psychophysiology*, 28(5):531–547.

Fawcett, A., Moro, E., and Lang, A. (2005). Pallidal deep brain stimulation influences both reflexive and voluntary saccades in huntington's disease. *Movement Disorders*, 20(3):371–376.

Fernandez, E., Pelayo, F., Ahnelt, P., Ammermüller, J., and Normann, R. A. (2005). Cortical visual neuroprostheses for the blind. *Restorative Neurology and Neuroscience*, page in press.

Fetz, E. (1999). Real-time control of a robotic arm by neuronal assemblies. *Nature Neuroscience*, 2:583–584.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley and Sons.

Flexer, A., Gruber, G., and Dorffner, G. (2005). A reliable probabilistic sleep stager based on a single EEG signal. *Artificial Intelligence in Medicine*, 33:199–207.

Friedman, D., Putnam, L., and Sutton, S. (1989). Event-related potentials in children, young adults and senior citizens. *Developmental Neuropsychology*, 5:33–60.

Frigo, M. and Johnson, S. (2006). FFTW home page. Website, retrieved January 25, 2006, from http://www.fftw.org.

Fukunaga, K. (1982). Intrinsic dimensionality extraction. In Krishnaiah, P. and Kanal, L., editors, *Handbook of Statistics*, volume 2, pages 347–360. Amsterdam.

Gao, X., Xu, D., Cheng, M., and Gao, S. (2003). A BCI-based environmental controller for the motion-disabled. *IEEE Transactions of Neural Systems and Rehabilitation Engineering*, 11(2):137–140.

Golland, P. (2002). Discriminative direction for kernel classifiers. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(286):531–537.

Goodglass, H. and Geschwind, N. (1976). Language disorders. In Carterette, E. and Friedman, M., editors, *Handbook of Perception: Language and Speech*, volume 7, New York. Academic Press.

Gratton, G. and Coles, M. (1989). Generalization and evaluation of eye-movement correction procedures. *Journal of Psychophysiology*, 3:14–16.

Guan, C., Thulasidas, M., and Wu, J. (2004). High performance P300 speller for brain-computer interface. In *Proceedings of IEEE Biological Circuits and Systems (BioCAS)*, Singapore.

Guger, C., Edlinger, G., Harkam, W., Niedermayer, I., and Pfurtscheller, G. (2003). How many people are able to operate an EEG-based brain-computer interface (BCI)? *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):145–147.

Guyon, I. (2006). SVM application list. Website, retrieved February 4, 2006, from http://www.clopinet.com/isabelle/Projects/SVM/applist.html.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Hamani, C., Hodaie, M., and Lozano, A. (2005). Present and future of deep brain stimulation for refractory epilepsy. *Acta Neurochirurgica*, 147(3):227–230.

Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3:61–70.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Heidemann, G. (2006). The principal components of natural images revisited. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. accepted.

Heuser, U., Göppert, J., Rosenstiel, W., and Stevens, A. (1997). Classification of human brain waves using self-organizing maps. In Lavrac, N., Karavnou, E., and Zupan, B., editors, *Intelligent Data Analysis in Medicine and Pharmacology*, pages 279–294, Boston. Kluwer Academic Publishers.

Hillyard, S., Hink, R., Schwent, V., and Picton, T. (1973). Electrical signs of selective attention in the human brain. *Science*, 182:177–180.

Hinterberger, T., Schmidt, S., Neumann, N., Mellinger, J., Blankertz, B., Curio, G., and Birbaumer, N. (2004a). Brain-computer communication and slow-cortical potentials. *IEEE Transactions on Biomedical Engineering*, 51(6):1011–1018.

Hinterberger, T., Weiskopf, N., Veit, R., Wilhelm, B., Betta, E., and Birbaumer, N. (2004b). An EEG-driven brain-computer interface combined with functional magnetics resonance imaging. *IEEE Transactions on Biomedical Engineering*, 51(6):971–974.

Holtmann, M., Stadler, C., Leins, U., Strehl, U., Birbaumer, N., and Poustka, F. (2004). Neurofeedback in der behandlung der aufmerksamkeitsdefizit-hyperaktivitätsstörung (adhs) bei kindern. *Zeitschrift für Kinder- und Jugendpsychiatrie*, 32(3):187–200.

Hoppe, C., Helmstaedter, C., Scherrmann, J., and Elger, C. (2001). No evidence for cognitive side effects after 6 months of vagus nerve stimulation in epilepsy patients. *Epilepsy and Behavior*, 2:351–356.

Hoppe, F. and Kaper, M. (2003). *EEG-Datenanalyse zur Entwicklung einer Gehirn-Computer Schnittstelle*. Bielefeld University, Faculty of Technology. Diploma Thesis.

Horch, W. and Dhillon, G. S. (2004). *Neuroprosthetics: Theory and Practice*. World Scientific Publishing Company, Singapore.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Suverys*, 2:94–128.

Jaskowski, P. and Verleger, R. (2000). An estimation of methods for single-trial estimation of P3 latency. *Psychophysiology*, 37:153–162.

Jasper, H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and clinical Neurophysiology*, 10:371–375.

Jung, T., Humphries, C., Lee, T., McKeown, M., Iragui, V., S.Makeig, and Sejnowski, T. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178.

Kaper, M., Meinicke, P., Grossekathoefer, U., Lingner, T., and Ritter, H. (2004). BCI competition 2003 - dataset IIb: Support vector machines for the P300 speller paradigm. *IEEE Transactions Biomedical Engineering*, 51:1073–1076.

Kaper, M., Meinicke, P., Müller, H. M., Weiss, S., Bekel, H., Hermann, T., Saalbach, A., and Ritter, H. (2006). Neuroinformatic techniques in cognitive neuroscience of language. In Rickheit, G. and Wachsmuth, I., editors, *Situated Communication*, pages 265–286, Berlin. Mouton de Gruyter.

Kaper, M. and Ritter, H. (2004a). Generalizing to new subjects in brain-computer interfacing. In *Proceedings of the 26th IEEE EMBS Annual International Conference (EMBC)*, San Francisco, USA.

Kaper, M. and Ritter, H. (2004b). Progress in P300-based brain-computer interfacing. In *Proceedings of IEEE Biological Circuits and Systems (BioCAS)*, Singapore.

Kaper, M., Saalbach, A., Finke, A., Müller, H. M., Weiss, S., and Ritter, H. (2005). Exploratory data analysis of EEG coherence using self-organizing maps. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*.

Karhunen, K. (1947). Über lineare methoden in der wahrscheinlichkeitsrechnung. *Annales Academiae Scientarium Fennicae*, 37:3–79.

Keerthi, S. S. and Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15:1667–1689.

Keirn, Z. and Aunon, I. (1990). A new mode of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering*, 37(12):1209–1214.

Kennedy, P., Bakay, R., Moore, M., Adams, K., and Goldwaithe, J. (2000). Direct control of a computer from the human central nervous system. *IEEE Transactions on Rehabilitation Engineering*, 8(2):198–202.

Kirchner, A., Birklein, F., Stefan, H., and Handwerker, H. (2000). Left vagus nerve stimulation suppresses experimentally induced pain. *Neurology*, 55:1167–1171.

Klinke, R., Frühstorfer, H., and Finkenzeller, P. (1968). Evoked responses as a function of external and stored information. *Electroencephalography and Clinical Neurophysiology*, 25:119–122.

Koles, Z. and Soong, A. (1998). EEG source localization: implementing the spatio-temporal decomposition approach. *Electroencephalography and Clinical Neurophysiology*, 107:343–352.

Kornhuber, H. and Deecke, L. (1965). Hirnpotentialänderungen bei willkürbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und reafferente potentiale. *Pflüger's Archiv für die gesamte Physiologie*, 284:1–17.

Krepki, R., Blankertz, B., Curio, G., and K.R.-Müller (2004). The berlin brain-computer interface (BBCI): towards a new communication channel for online control of multimedia applications and computer games. *Journal of Multimedia Tools and Applications*.

Kronegg, J., Voloshynovskiy, S., and Pun, T. (2005). Analysis of bit-rate definitions for brain-computer interfaces. *International Conference on Human-Computer Interaction (HCI'05)*.

Kübler, A., Kotchoubey, B., Hinterberger, T., Ghanayim, N., Perelmouter, J., Schauer, M., Fritsch, C., Taub, E., and Birbaumer, N. (1999). The thought translation device: a neurophysiological approach to commincation in total motor paralysis. *Experimental Brain Research*, 124:223–232.

Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., and Birbaumer, N. (2001). Brain-computer communication: Unlocking the locked in. *Psychological Bulletin*, 127(3):358–375.

Kumar, R., Lozano, A. M., Sime, E., Halket, E., and Lang, A. E. (1999). Comparative effects of unilateral and bilateral subthalamic nucleus deep brain stimulation. *Neurology*, 53:561–571.

Kumar, R., Lozano, A. M., Sime, E., and Lang, A. E. (2003). Long-term follow-up of thalamic deep brain stimulation for essential and parkinsonian tremor. *Neurology*, 61:1601–1604.

Kutas, M. and Hillyard, S. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207:203–205.

Laitinen, L. (2003). *Neuromagnetic sensorimotor signals in brain computer interfaces.* Helsinki University of Technology. Master's Thesis.

Lalor, E., Kelly, S., C.Finucane, R.Burke, Reilly, R., and McDarby, G. (2004). Brain computer interface based on the steady-state vep for immersive gaming control. *Biomedizinische Technik*, 49(1):63–64.

Lantz, G., Peralta, R., Spinelli, L., Seeck, M., and Michel, C. (2003). Epileptic source localization with high density EEG: How many electrodes are needed? *Clinical Neurophysiology*, 114:63–69.

Lemm, S., Blankertz, B., Curio, G., and K.R.-Müller (2005). Spatio-spectral filters for improved classification of single trial EEG. *IEEE Transactions on Biomedical Engineering.* to appear.

Levine, S. P., Huggins, J. E., BeMent, S. L., Kushwaha, R. K., Schuh, L. A., Rohde, M. M., Passaro, E. A., Ross, D. A., Elisevich, K., and Smith, B. J. (2000). A direct brain interface based on event-related potentials. *IEEE Transactions on Rehabilitation Engineering*, 8:180–185.

Löwenstein, K. and Borchart, M. (1918). Symptomatologie und elektrische reizung bei einer schubverletzung des hinterhauptlappens. *Deutsche Zeitung für Nervenheilkunde*, 58:264.

Lykken, D. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43(6):385–388.

Makeig, S., Debener, S., and Onton, J. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, 8(5):204–210.

Margalit, E., Maia, M., Weiland, J., Greenberg, R., Fujii, G., Torres, G., Piyathaisere, D., O'Hearn, T., Liu, W., Lazzi, G., Dagniele, G., Scribner, D., de Juan, E., and Humayun, M. (2002). Retinal prosthesis for the blind. *Survey of Ophthalmology*, 47(4):335–356.

Marques de Sa, J. (2001). *Pattern Recognition.* Springer, New York.

Martin, J. H. (1991). The collective electrical behavior of cortical neurons: The electroencephalogram and the mechanisms of epilepsy. In *Principles of Neural Science*, pages 777–790. Elsevier.

Maynard, E. (2001). Visual prostheses. *Annual Review of Biomedical Engineering*, 3:145–168.

McFarland, D. J., Sarnacki, W. A., and Wolpaw, J. R. (2003). Brain-computer interface (BCI) operation: Optimizing information transfer rates. *Biological Psychology*, 63:237–251.

Medtronics (2005). Questions and answers about activa parkinson's control therapy. Website, retrieved June 27, 2005, from http://www.medtronic.com.

Mehring, C., Rickert, J., Vaadia, E., de Oliviera, S. C., Aertsen, A., and Rotter, S. (2003). Inference of hand movements from local field potentials in monkey motor cortex. *Nature Neuroscience*, 6:1253–1254.

Meinicke, P., Hermann, T., Bekel, H., Müller, H. M., Weiss, S., and Ritter, H. (2004). Identification of discriminative features in eeg. *Journal for Intelligent Data Analysis*, 8(1):97–107.

Meinicke, P., Kaper, M., Hoppe, F., Heumann, M., and Ritter, H. (2003). Improving transfer rates in brain computer interfacing: a case study. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.

Mellinger, J., Hinterberger, T., Bensch, M., Schröder, M., and Birbaumer, N. (2003). Surfing the web with electrical brain sigfnals: The brain web surfer (BWS) for the completely paralysed. In Ring, H. and Soroker, N., editors, *Proceedings of the 2nd World Congress of the International Society of Physical and Rehabilitation Medicine (ISPRM)*.

Mellinger, J., Nijboer, F., Pawelzik, H., Schalk, G., McFarland, D. J., Vaughan, T. M., Wolpaw, J. R., Birbaumer, N., and Kübler, A. (2004). P300 for communication: Evidence from patients with amyotrophic lateral sclerosis (ALS). *Biomedizinische Technik*, 49:71–74.

Middendorf, M., McMillan, G., Calhoun, G., and Jones, K. (2000). Brain-computer interface based on the steady-state visual-evoked response. *IEEE Transactions on Rehabilitation Engineering*, 8(2):211–214.

Mika, S., Rätsch, G., and Müller, K.-R. (2001). A mathematical programming approach to the kernel fisher algorithm. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 591–597, Cambridge, MA. MIT Press.

Morgan, S., Hansen, J., and Hillyard, S. (1996). Selective attention to stimulus location modulates the steady state visual evoked potential. *Proceedings of the National Academy of Science USA*, 93:4770–4774.

Müller, H. M. (2006). Neurobiological aspects of meaning constitution during language processing. In Rickheit, G. and Wachsmuth, I., editors, *Situated Communication*, pages 243–264, Berlin. Mouton de Gruyter.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Müller, M., Picton, T., Valdes-Sosa, P., Riera, P., Teder-Sälerjärvi, W., and Hillyard, S. (1998). Effects of spatial selective attention on the steady-state visual evoked potential in the 20-28hz range. *Cognitive Brain Research*, 6:249–261.

Mullis, R., Holcomb, P., Diner, B., and Dykman, R. (1985). The effect of aging on the P3 component of the visual event-related potential. *Electroencephalography and Clinical Neurophysiology*, 62:141–149.

Neuper, C. and Pfurtscheller, G. (1999). Motor imagery and erd. *Handbook of Electroencephalogram and Clinical Neurophysiology*, 6:3–74.

NeuralSignals (2006). Neural signals inc. - cutting edge assistive technology. Website, retrieved February 2, 2006, from http://www.neuralsignals.com.

Nicolelis, M. (2001). Actions from thoughts. *Nature*, 409:403–406.

Nicolelis, M. (2003). Brain-machine interfaces to restore motor function and probe neural circuits. *Nature Neuroscience*, 4:417–422.

Nicolelis, M., Birbaumer, N., and Müller, K.-R. (2004). Editorial. *IEEE Transactions on Biomedical Engineering*, 51/6:877–879.

NIDCD (2006). Cochlear implants [NIDCD health information]. Website, retrieved February 4, 2006, from http://www.nidcd.nih.gov/health/hearing/coch.asp.

Nieuwenhuys, R., Voogd, J., and van Huijzen, C. (1988). *The Human Central Nervous System*. Springer, Berlin.

Nykopp, T., Laitinen, L., Heikkonen, J., and Sams, M. (2005). Statistical methods for MEG based finger movement classification. *IEEE Transactions on Neural Systems and Rehabiliteering Engineering*. in press.

Näätänen, R. (1982). Processing negativity: An evoked potential reflection of selective attention. *Psychological Bulletin*, 92:605–640.

Obermaier, B., Guger, C., Neuper, C., and Pfurtscheller, G. (2001). Hidden markov models for online classification of single trial EEG data. *Pattern Recognition Letters*, 22:1299–1309.

Patel, S. and Azzam, P. (2005). Characterization of N200 and P300: Selective studies of the event-related potential. *International Journal of Medical Sciences*, 2(4):147–154.

Patterson, J. and Grabois, M. (1986). Locked-in syndrome: a review of 139 cases. *Stroke*, 17:758–764.

Pfingst, B. (2000). Auditory prostheses. In Chapin, J. and Moxon, K., editors, *Neural Prostheses for Restoration of Sensory and Motor Function*, pages 3–43. CRC Press.

Pfurtscheller, G. and Neuper, C. (1997). Motor imagery activates primary sensorimotor area in man. *Neuroscience Letters*, 239:65–68.

Pfurtscheller, G., Neuper, C., Guger, C., Harkam, W., Ramoser, H., Schlögl, A., Obermaier, B., and Pregenzer, M. (2000). Current trends in graz brain computer interface (BCI) research. *IEEE Transactions on Rehabilitation Engineering*, 8(2):216–219.

Pfurtscheller, G., Neuper, C., Müller, G., Obermaier, B., Krausz, G., Schlögl, A., Scherer, R., Graimann, B., Keinrath, C., Skliris, D., Woertz, M., Supp, G., and Schrank, C. (2003). Graz-BCI: State of the art and clinical applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):177–180.

Picton, T., Alain, C., Otten, L., Ritter, W., and Achim, A. (2000). Mismatch negativity: Different water in the same river. *Audiology and Neuro-Otology*, 5:111–139.

Picton, T., Lins, O., and Scherg, M. (1995). The recording and analysis of event-related potentials. In Boller, F. and Grafman, J., editors, *Handbook of Neuropsychology*, pages 3–74. Elsevier.

Pinel, J. P. (1990). *Biopsychology.* Allyn & Bacon, Boston.

Polich, J. (1998). P300 clinical utility and control of variability. *Journal of Clinical Neurophysiology*, 15(1):14–33.

Polich, J. (2003). Overview of P3a and P3b. In Polich, J., editor, *Detection of Change: Event-Related Potential and fMRI Findings*, pages 83–98, Boston. Kluwer Academic Press.

Posner, M., Snyder, C., and Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, New York, NY, USA.

Pritchard, W. (1981). Psychophysiology of P300. *Psychological Bulletin*, 89:506–540.

Pritzel, M., Brandt, M., and Markowitsch, H. (2003). *Gehirn und Verhalten.* Spektrum, Heidelberg.

Qian, H., Loizou, P., and Dorman, M. (2003). A phone-assistive device based on bluetooth technology for cochlear implant users. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(3):282–287.

Rappelsberger, P. and Petsche, H. (1988). Probability mapping: Power and coherence analysis of cognitive processes. *Brain Topography*, 1:46–54.

Regan, D. (1989). *Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine.* Elsevier.

Ripley, B. (1999). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, UK.

Rosenfeld, J. (2005). Brain fingerprinting: A critical analysis. *Scientific Review of Mental Health Practice.* in press.

Sajda, P., Gerson, A., Müller, K.-R., Blankertz, B., and Parra, L. (2003). A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):184–185.

Salek-Haddadi, A., Friston, K., Lemieux, L., and Fish, D. (2003). Studying spontaneous EEG activity with fMRI. *Brain Research Reviews*, 43:110–133.

Schachter, S. (2002). Vagus nerve stimulation therapy summary - five years after fda approval. *Neurology*, 59:15–29.

Schack, B. and Weiss, S. (2005). Quantification of phase synchronization phenomena and their importance for verbal memory processes. *Biological Cybernetics*, 92(1):275–287.

Schandry, R. (1981). *Psychophysiologie.* Urban and Schwarzenberg, München.

Scherer, R., Müller, G., Neuper, C., Graimann, B., and Pfurtscheller, G. (2004). An asynchronously controlled EEG-based virtual keyboard: Improvement of the spelling rate. *IEEE Transactions on Biomedical Engineering*, 51(6):979–984.

Schmidt, E., Bak, M., Hambrecht, F., Kufta, C., O'Rourke, D., and Vallabhanath, P. (1996). Feasibility of a visual prosthesis for the blind based on intracortical microstimulation of the visual cortex. *Brain*, 119:507–522.

Schoelkopf, B. and Smola, A. J. (2002). *Learning with Kernels.* The MIT Press, Cambridge, MA.

Serby, H., Yom-Tov, E., and Inbar, G. (2005). An improved P300-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 3(1):89–98.

Serruya, M., Hatsopoulos, N., Paninski, L., Fellows, M., and Donoghue, J. (2002). Instant control of a movement signal. *Nature*, 416:141–142.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Smith, S. W. (1999). *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, San Diego.

Sokolow, Y. (1963). *Perception and the Conditioned Reflex.* Pergamon Press, Oxford.

Sutton, S., Braren, M., Zubin, J., and John, E. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, 150:1187–1188.

Tallon-Baudry, C. and Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences*, 3(4):151–162.

Talwar, S., Xu, S., Hawley, E., Weiss, S., Moxon, K., and Chapin, J. (2002). Rat navigation guided by remote control. *Nature*, 317:37–38.

Taylor, D., Tillery, S., and Schwartz, A. (2002). Direct cortical control of 3d neuroprosthetic devices. *Science*, 296:1829–1832.

Trolltech (2006). Trolltech - cross-platform C++ gui development, and embedded linux solutions. Website, retrieved February 3, 2006, from http://www.trolltech.com.

Tucker, D. M. (1993). Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*, 87:145–163.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86.

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., Goodale, M., and Mansfield, R., editors, *Analysis of visual behior*, pages 549–586, Cambridge. MIT Press.

Uthman, B., Reichl, A., and Dean, J. (2004). Effectiveness of vagus nerve stimulation in epilepsy patients: A 12-year observation. *Neurology*, 63:1124–1126.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Varela, F., Lachaux, J.-P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: Phase and synchronisation and large-scale integration. *Nature Neuroscience*, 2:229–239.

Vaughan, H. and Arezzo, J. (1988). The neural basis of event-related potentials. *Handbook of Electroencephalogram and Clinical Neurophysiology*, 3:45–96.

Verleger, R. (1988). Event-related potentials and cognition: A critique of the context updating hypothesis and an alternative interpretation of P3. *Behavioral and Brain Sciences*, 11:343–427.

Vetter, R., Williams, C., Hetke, J., Nunamaker, E., and Kipke, D. (2004). Chronic neural recording using silicon-substrate microelectrode arrays implanted in cerebral cortex. *IEEE Transactions on Biomedical Engineering*, 51(6):896–904.

Weiskopf, N., Mathiak, K., Bock, S., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., and Birbaumer, N. (2004). Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). *IEEE Transactions on Biomedical Engineering*, 51(6):966–970.

Weiss, S., Müller, H., Schack, B., King, J., Kutas, M., and Rappelsberger, P. (2005). Increased neuronal synchronization accompanying sentence comprehension. *International Journal of Psychophysiology*, 57:129–141.

Weiss, S. and Müller, H. (2003). The contribution of EEG coherence to the investigation of language. *Brain and Language*, 85:325–343.

Wessberg, J., Stambaugh, C., Kralik, J., Beck, P., Chapin, J., Kim, J., Biggs, S., Srinivasan, M., and Nicolelis, M. (2000). Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408:361–365.

Wickelgreen, I. (2003). Tapping the mind. *Science*, 299:496 – 499.

Wolpaw, J., Birbaumer, N., Heetderks, W., McFarland, D., Peckham, P., Schalk, G., Donchin, E., Quatrano, L., Robinson, C., and Vaughan., T. (2000). Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, 8:161–163.

Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G., and Vaughan, T. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113:767–791.

Wolpaw, J. and McFarland, D. (1994). Multichannel EEG-based brain-computer communication. *Electroencephalographie and Clinical Neurophysiology*, 90:444–449.

Wolpaw, J. and McFarland, D. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Science (PNAS)*, 101:17849–17854.

Wolpaw, J., McFarland, D., Vaughan, T., and Schalk, G. (2003). The wadsworth center brain-computer interface (BCI) research and development program. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):204–207.

Wolpaw, J. R., McFarland, D. J., Neat, G. W., and Forneris, C. A. (1991). EEG-based brain-computer interface for cursor control. *Clinical Neurophysiology*, 78:252–259.

Wolpe, P., Foster, K., and Langleben, D. (2005). Emerging neurotechnologies foer lie-detection: Promises and perils. *American Journal of Bioethics*, 5(2):39–49.

Xu, N., Gao, X., Hong, B., Miao, X., Gao, S., and Yang, F. (2004). BCI competition 2003 - dataset IIb: Enhancing P300 wave detection using ICA-based subspace projections for BCI applications. *IEEE Transactions Biomedical Engineering*, 51:1067–1072.

Zschocke, S. (1995). *Klinische Elektroenzephalographie*. Springer, Berlin.