Feature Selection for Brain-Computer Interfaces

Dissertation
der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. Michael W. Tangermann
aus Berlin

Tübingen
2007

Tag der mündlichen Prüfung:	11.7.2007
Dekan	Prof. Dr. Michael Diehl
1. Berichterstatter:	Prof. Dr. Wolfgang Rosenstiel
2. Berichterstatter:	Prof. Dr. Bernhard Schölkopf (MPI Tübingen)

Dedicated to the patients.

Acknowledgment

The author would like to thank Prof. Dr. Wolfgang Rosenstiel and all the collegues in Tübingen and Berlin for fruitful discussions, for support and funding of the BCI project. Special thanks go to my project partners Navin Lal, Jeremy Hill and Prof. Dr. Bernhard Schölkopf at the MPI Tübingen as well as to Dr. Thilo Hinterberger and Prof. Dr. Niels Birbaumer at the Institut für medizinische Psychologie und Verhaltensneurobiologie Tübingen for the long and fruitful interdisciplinary collaboration on this difficult field we work on. Thanks goes also to Ulrike Romberg, my students at the University of Tübingen that performed restless data processing and to Bernd Battes and Prof. Dr. Kuno Kirschfeld for their support with the EEG recordings. The work was kindly supported by the Deutsche Forschungsgemeinschaft DFG under the SFB 550, B5 and grant RO 1030/12, and by the National Institute of Health (D.31.03765.2). I am indebted to my wife, my friends and my family for the support in those more difficult times of writing.

Abstract

Brain-Computer Interface (BCI) systems are a means of establishing communication for severely paralyzed patients. Based on the brain activity signals during the execution of mental tasks by a user, a computer system translates those signals first into higher-level features and finally into control commands for communication interfaces. This involves a number of algorithmic steps that have to be optimized individually for each patient in order to attain high performance of the BCI. One of these steps is the choice of a suitable set of brain signal features. This set is supposed to provide good discriminability of the mental tasks, allow for introspection, simplify the experimental effort and thus increase acceptance in patients. In terms of EEG electrodes, a smaller feature set entails a reduction of the immense effort of the daily setting-up of electrodes prior to the start of an experiment. The problem of feature selection is hard to tackle as optimal features may vary between subjects and even between sessions with the same subject.

This thesis proposes a new signal processing framework for BCI that incorporates a quick and fully algorithmic feature selection step combined with an SVM classification in embedded form. For the evaluation of this new methodology, the results of own online and offline studies with electroencephalogram (EEG), electrocorticogram (ECoG) and magnetoencephalogram (MEG) will be presented, including the first ever implementation of a motor-imagery BCI system in MEG. In addition, the framework has been evaluated against several existing filter and wrapper approaches for feature selection. According to these results, the new method is capable of adapting to the changing signal characteristics of BCI users, can be used without prior neurophysiological knowledge about underlying mental tasks, reduces the number of features from several hundreds to just about 10% of the original features all while remaining highly accurate in terms of classification performance. Furthermore, the results of the feature selection step prove to be plausible in terms of neurophysiology, i.e. chosen EEG channels agree well with the expected underlying cortical activity patterns during the mental tasks. Under restricted conditions, it is shown that the optimized feature sets determined by the new signal processing framework can be transferred across subjects with only a small drop in performance.

Zusammenfassung

Ein Brain-Computer-Interface-System (BCI) bietet schwerstgelähmten Patienten eine Möglichkeit zu kommunizieren. Während sie verschiedene mentale Aufgaben ausführen, was kurzzeitig zu unterschiedlichen mentalen Zuständen führt, werden die Gehirnsignale der Patienten aufgezeichnet. Aus diesen Signalen extrahiert ein Computersystem zuerst komplexe Merkmale und übersetzt diese in einem zweiten Schritt in Kontrollsignale zur Steuerung einer Kommunikationsanwendung. Extraktion und Übersetzung der Signale werden durch z.T. lernende Algorithmen realisiert, welche für jeden Benutzer individuell angepasst werden müssen, um eine optimale Leistungsfähigkeit des BCI-Systems zu erzielen. Einer der zu optimierenden Schritte ist die Selektion einer geeigneten Menge von Merkmalen. Diese Merkmalsmenge soll eine möglichst exakte Unterscheidung der mentalen Zustände ermöglichen, durch Einblick in die Lösung das Verständnis für diesen Lösungsansatz erhöhen, den experimentellen Aufwand absenken und damit die Akzeptanz für das BCI-System bei den Benutzern erhöhen. Sieht man EEG Elektroden als Merkmale an, so ist die Auswahl einer kleineren Merkmalsmenge äuSSerst wünschenswert - sie verkleinert den immensen täglichen Aufwand für das Setzen der Elektroden vor dem Beginn der Kommunikation per BCI. Das Problem der Merkmalsselektion ist außerordentlich schwierig zu lösen, da optimale Merkmale sowohl zwischen Benutzern als auch zwischen Sitzungen des gleichen Benutzers variieren.

Diese Doktorarbeit schlägt zur Lösung des Problems eine neue Methode für die Signalverarbeitung in BCI Systemen vor. Sie umfasst eine schnelle algorithmische Merkmalsauswahl, die mit der Signalklassifikation durch Support-Vektor-Maschinen (SVM) in eingebetteter Form kombiniert wird. Zur Evaluierung dieses neuen Ansatzes werden die Ergebnisse eigener Studien mit unterschiedlichen Signalquellen präsentiert. Sie umfassen BCI Experimente mit Messungen des Elektroezephalogramms (EEG), des Elektrokortikogramms (ECoG) und des Magnetenzephalogramms (MEG). Für die Experimente dieser Doktorarbeit wurde außerdem erstmalig ein auf motorischen Vorstellungen basierendes BCI-System mit MEG-Signalen realisiert und online getestet. Zur Validierung der neu vorgeschlagenen Methode wurde sie mit bekannten Filter- und Wrappermethoden für die Merkmalsselektion bei BCI verglichen. Die Ergebnisse zeigen, dass sich die neue Methode an die wechselnde Signalcharakteristika von Benutzern anpassen kann. Sie kann ohne neurophysiologisches Vorwissen über die zugrundeliegenden Hirnvorgänge angewendet werden und ist damit auch für neuartige mentale Aufgaben geeignet. Sie schafft es, die Anzahl der Merkmale von initial mehreren Hundert auf ungefähr 10% der Ausgangsmenge zu reduzieren, während eine hohe Klassifikationsgenauigkeit beibehalten wird. Die von der Methode ausgewählten Merkmale stimmen mit denjenigen überein, die man durch neurophysiologisches Grundlagenwissen erwarten kann - die gewählten EEG-Kanäle etwa liegen über denjenigen kortikalen Gebieten, von denen

man erwartet, dass sie für die Vorstellungsaufgaben relevant sind. Für die erschwerte Aufgabenstellung der Übertragung von Merkmalsmengen von Benutzer zu Benutzer konnte gezeigt werden, dass dies mit Merkmalsmengen, die durch die neue Methode bestimmt wurden, mit nur kleinen Abstrichen in der Klassifikationsgenauigkeit möglich ist.

Contents

List of Figures

List of Tables

1 Introduction

According to today's knowledge, emotions as well as conscious or unconscious cognitive accomplishments of man are realized by the cells of the brain via electrical and chemical interactions. While most chemical processes are very difficult to observe in the active brain, at least some broad electric activity patterns of the brain have been discovered rather early: Hans Berger recorded the first Electroencephalogram (EEG) in 1924 and published these observations in 1929 [Ber29]. He realized that EEG signals that depend on the synchronous activity of millions of neurons can reflect physiological states like sleep or wakefulness. The recording of EEG signals and, later, their treatment in early digital signal processing systems opened further possibilities to the field. For example, the averaging of hundreds of short EEG signal sequences (so-called trials) that had been gained from repetitions under constant experimental conditions allowed to reduce the signal noise so that the average electrical reaction of a brain upon various stimuli could be investigated.

During the last two decades, even newer techniques for recording and visualizing brain activity emerged. The availability of increased computing power and improved signal analysis techniques brought about the rise of single trial EEG analysis. By judging EEG in single trial, real time feedback experiments became possible and led to the development of the first Brain-Computer Interfaces (BCIs) [RBEL84],[WM94], although the idea to use electrical brain signals for control or communication had already existed since Berger's discovery.

Among the first applications of BCI systems was the self-regulatory training of subjects to control slow cortical potentials of the EEG [HKK$^+$00]. Later followed other applications like relaxation training for epileptic patients to avoid seizures [KSU$^+$01], feedback training for patients suffering from chronic pain [LLD76], ADHD [LL95] and many more.

Today most BCI research has the goal to enhance the abilities of severely disabled and even completely paralyzed patients. Such patients may suffer from paralysis for different reasons but have the loss of that motor control and the ability to communicate in common. If the sensory abilities of these patients, like the perception of sounds, speech, their vision ability and their proprioception are intact or only partly damaged and if in addition some higher cognitive functions are present, the patient faces a very unsatisfying so-called completely locked-in (CLIS) state. Less severe CLIS with remaining eye movement control is called locked-in state (LIS). The patient perceives the world, processes this information on the basis of his personal background and cognitive abilities, but is not able to communicate the results to anyone or to initiate any kind of change of the (personal) environment. It is easy to imagine that the CLIS syndrome is highly difficult to bear for the patient and that the re-establishment of communication abilities must be one of the primary goals of treatment.

BCI systems try to overcome the CLIS or LIS state. Using brain activity directly as an output channel, a BCI system can classify brain patterns and interpret them as simple commands. These commands can be mapped by a computer to control attached devices. For paralyzed patients, these devices can range from rather simple switches of e.g. the TV, the door opener or the room light to specialized powerful communication software. The latter can open a nearly endless number of information sources to the patient via a web browser or allow for participating in decisions, for pursuing personal communication and personal relationships in spite of the paralysis. BCI systems can therefore bring CLIS patients back to communication and end their involuntary isolation.

BCI is an interdisciplinary field that needs combined effort of medicine, psychology, physics, and computer science. For EEG-based BCI, two main design approaches can be observed. The oldest and best-established approach relies on long term human feedback training to improve control. In such a setting, the signals processing, the exploited EEG features, and the mapping of these features to control signals are fixed. The necessary training of a patient is based on conditioning with feedback and has several weeks' duration. The constant motivation of patients during such long training periods is very difficult. Furthermore, this approach is very sensitive to the correct choice of the initial BCI setting - there is a certain risk that the (possibly severely handicapped) patient is unable to produce the desired EEG signals that are necessary to control the fixed BCI system.

The newer approach puts the load of learning on the computer as much as possible before human training is involved. This machine learning process involves the adaptation of real-time signal processing algorithms, the choice of suitable EEG features and the training of classification algorithms that map features to control signals. Although not all of these steps can up to now be made by automatic methods of machine learning (a supervisor still has to take some of the decisions, e.g. which features to use), advantages of this second approach are obvious: not only is the training process for the user sped up by a quicker start, but the approach additionally allows an adaption of the BCI system to the individual abilities and EEG characteristics of the user. Still, there are many data analysis problems as well as problems related to the experimental situation to be improved. This encourages computer scientists and machine learners to enter

the field and participate in the design of specialized algorithms and provides challenging tasks for the machine learning community.

The focus of this thesis is a specific aspect of machine learning for BCI: the identification of suitable brain signal features that help solving the learning task optimally for a new subject or a new experimental setting and at the same time reduce the need of the very time-consuming application of many sensors at the beginning of every session. The latter goal is considered very important for achieving higher acceptance rates of BCI systems among the users.

From a machine learning point of view, feature selection with noisy data is a hard problem [MKD+04], especially if the number of possible features is high compared to the number of examples the algorithms can learn from. For learning problems in BCI, this is typically the case - hundreds or thousands of possible features are opposed to only a few hundred data points. A simple solution would be to decide on some fixed standard features beforehand and not to change them during the machine learning process. But there are several reasons why the use of predefined feature sets usually does not lead to optimal results. First, although the classification-relevant signal features (e.g. spatial patterns, recording positions or frequency bands) are roughly known for some mental tasks, these features show quite some amount of variation from subject to subject. The reason for this is that the folding of the cortex and the position as well as the fine granular functional structures vary between individuals due to their ontogenesis and that the characteristic way, in which mental tasks are executed, are different from user to user. Secondly, it is known that paralysis of patients is often accompanied with cortex damages and possibly reorganizations. In addition, smaller cognitive lacks due to these damages are not unusual. In these cases, the classical BCI paradigms have to be adapted to the abilities of the individuals. In these cases standard features might not be useful any more. Third, BCI research in general is in need of investigating new and more mental tasks for which typical patterns or useful features are not known. This necessity arises when typical cortex areas (e.g. motor cortex) show lesions or are degenerated. Future research comprises the exploration of additional new mental tasks for BCI. They might even increase the information rate of existing BCI systems, e.g. by expanding the widely used robust two-class BCIs to multi-class BCIs.

The thesis is structured as follows: The basic concepts related to the recording and interpretation of brain activity are introduced in Section 2. This comprises an introduction to brain physiology, recording techniques for BCI, an introduction into the types of features that can be derived from such recordings, an introduction into classification and feature selection. State of the art BCI systems are introduced in Section 3. Together with the BCI applications, the living conditions of paralyzed patients are introduced and several aspects of BCI paradigms are highlighted. After a reflection on general difficulties that arise in BCI, the analysis of four existing BCI systems follows. The thesis is centered around Section 4, which proposes a novel algorithm for individual feature selection (IFS) in BCI to cope with the shortcomings of the existing BCI approaches. The algorithm is part of a new signal processing concept, and evaluation and implementation aspects are discussed. Section 5 explains the procedure of a series of BCI experiments performed with three different recoding techniques. The experiments were designed to investigate three main aspects of the newly introduced individual feature selection (IFS) concept in detail: the classification performance of the new system, the comparison of automatically chosen feature subsets to those chosen with *a priori* knowledge and the robustness of the method in terms of different signal sources and of the transfer of feature subsets to new subjects. Section 6 shows the results and discussions of these investigations, broken down into the above mentioned three main aspects. Finally Section 7 finishes with a summary of the merits of this thesis to the topic of feature selection in BCI.

2 Fundamentals

2.1 Relevant Structures of the Human Brain

This section introduces the structures relevant for the generation of brain signals in a bottom-up structure. First, the neurons and their electric abilities are described, before the effect of bigger ensembles of neurons is familiarized on a cortical level.

2.1.1 Neurons and their Electrical Activity

The human nervous system consists of approximately 10^{10} to 10^{11} neurons, cells specialized in information processing, and of about the same number of neuroglia cells that support the neurons' activities in various ways [EAW93]. Most of the neurons are situated in the central nervous system consisting of the brain and the spinal chord.

In a single neuron, information processing in its simplest form takes place. The information processing is based on the special characteristics of the membrane of neurons. Without external excitation of the neuron, the passive semi-permeable membrane with its selective ion channels and the active transport of ions by the Na^+/K^+ pump establish a constant polarization. This resting cell potential between the intra- and extracellular space has approximately -70 mV. Figure 2.1 shows the prototype of a pyramidal cell. It is the prevalent neuron type of the cerebral cortex (see below). At its apical synapses of the dendritic tree, a neuron receives information either from sensory cells or from other neurons in the form of electrical or chemical stimulations which can be excitatory or inhibitory. If excitatory stimulations prevail, an

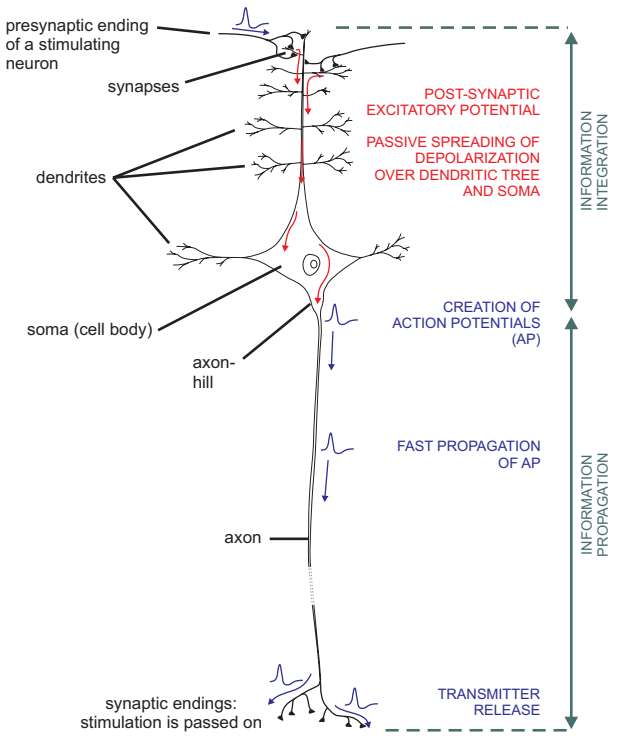


Figure 2.1: *Sketch of a pyramidal cell of the human motor cortex. The red arrows indicate the intracellular spreading of a depolarization (EPSP) caused by Na^+ influx at the dendritic endings. The green segments indicate different functional areas of the neuron. In the dendritic tree and the soma the spatio-temporal integration of potentials takes place. If the overall depolarization at the axon hill is strong enough to trigger the generation of action potentials (APs), the neuron propagates stimulation in the form of APs along the axon to distant α-motorneurons which in turn stimulate muscle fibers. Please note that the length of the axon is reduced strongly for this illustration.*

inflow of Na^+ ions through the membrane occurs. This inflow transiently disturbs the resting cell potential, depolarizes

the membrane, and leads to a so-called excitatory postsynaptic potential (EPSP). This depolarization only lasts for 1-2 ms before the influx of K^+ ions reestablishes the original polarization. The red arrows in figure 2.1 show how the depolarization spreads from the apical end of the dendritic tree towards the cell body (soma) and finally reaches the axon hillock. When the depolarization arrives at the axon hillock with sufficient amplitude, the neuron produces a series of all-or-nothing action potentials (AP) which is fired through the axon to other neurons. In the case of a pyramidal cell of the motor cortex, this axon is very large. It runs down to the spinal cord and connects to an α-motorneuron which in turn stimulates muscle fibers.

During the resting state of the neuron, a polarization of approximately -70 mV is established at the membrane (see figure 2.2). In the resting state, the membrane is in equilibrium, as is the intracellular and the extracellular space.

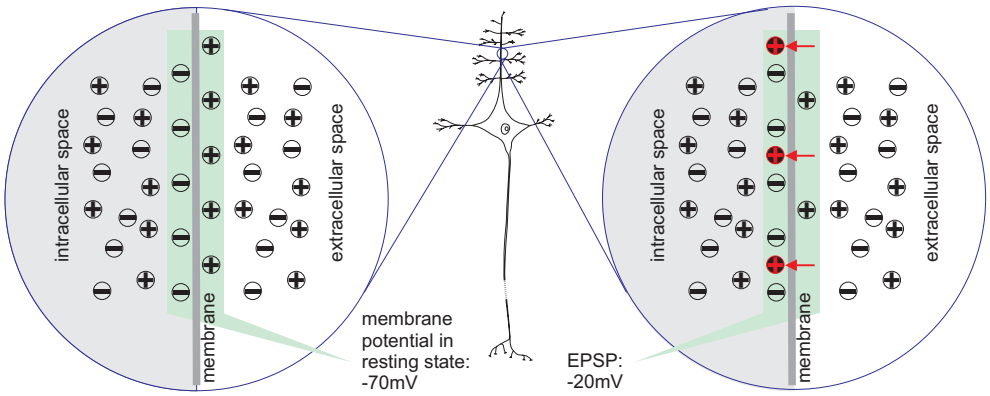


Figure 2.2: *Situation at the membrane close to firing excitatory synapses at the dendrite tree. The distribution of ions close to the membrane, of the intracellular- and the extracellular space is shown. Left illustration: During the cell resting state (before synaptic excitation) the membrane is polarized to approximately -70 mV. The intra- and extracellular spaces are in equilibrium. The positive and negative ions close to the membrane attract each other. The right illustration shows the situation of an excitatory postsynaptic potential (EPSP). Due to the opening of Na^+ channels an inflow of positive ions has reduced the polarization to approximately -20 mV.*

If the neuron is excited by a predecessor neuron, many firing synapses at the apical end of the dendrite tree cause an excitatory postsynaptic potential (EPSP) of approximately -20 mV at the membrane at this part of the neuron. Compared to the distant membrane potentials of the soma and the basal dendritic endings, which still show the resting potential of -70 mV, the apical dendritic tree membrane becomes transiently positively charged (see figure 2.3). The resulting

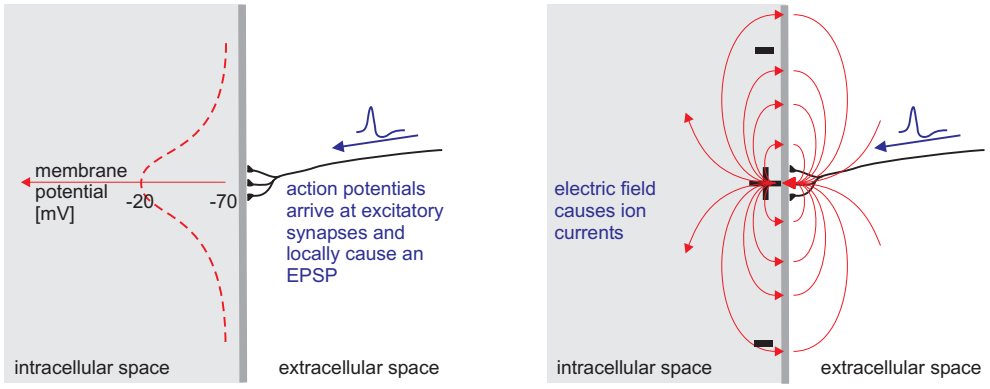


Figure 2.3: *The left illustration shows a sketch of the local distribution of an EPSP. While the membrane close to the stimulating synapses is depolarized to approximately -20 mV, distant membranes e.g. at the base of the dendritic tree still show the resting potential of -70 mV. The right illustration shows the resulting electric field and ion currents.*

electric field along the membrane (and a magnetic field orientated perpendicularly with respect to the electric field) starts local compensating currents inside the dendrite and in the extracellular space [BML01].

If the electric activity of neighboring neurons is synchronized, it is reflected by so-called local field potentials (LFP). Changes of these electromagnetic fields can be picked up by invasive and - if the group of neurons is large enough - also

by non-invasive recording techniques (see Section 2.2) that are commonly used for BCI systems.

2.1.2 Functional Organization of the Cerebral Cortex

The cerebral cortex (also called neocortex, isocortex or cortex cerebri) is the part of the brain that has evolved most intensively during the evolution of vertebrates. It embraces the growing cerebrum and has gained more and more new functionality over evolutionary time. For simple vertebrates, whose cerebral cortex is rather small, it only implements motor control and sensory functions. In higher developed vertebrates like mammals and especially in the case of primates and cetacea (whales and dolphins) the growth and specialization of the cerebral cortex culminated in a folded, bilateral structure that enlarged its surface significantly. The structure of sulci (grooves of the folding, singular *sulcus*) and gyri (bumps, singular *gyrus*) are useful to coarsely recognize functional areas but differ in details between individuals.

The neurons of the two hemispheres are strongly interconnected via the corpus callosum and connect to deeper structures of the cerebrum (e.g. to the limbic system, several nuclei), to the cerebellum and also via the medulla oblongata towards the spinal chord.

The human cerebral cortex has an unfolded area of approximately $0.5m^2$. In addition to areas specialized in motor and sensory functions (the so-called motor and sensory cortex, see figure 2.4) it shows regions involved in various other cognitive tasks like language, vision, auditory perception, memory, consciousness, planning, reasoning etc.

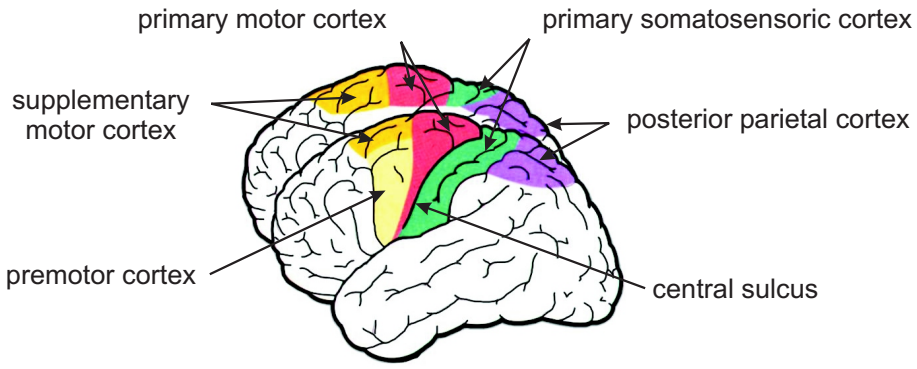


Figure 2.4: *Sensory areas and motor areas of the human cerebral cortex (according to [Goh97].), seen from the left side. For improved introspection, the left and right cortex are plotted with a small gap.*

Especially for the higher-level cognitive tasks, the involved cortex areas are not easy to localize and can vary between individuals. But even simple tasks demand the exchange of information of several regions (neuron clusters) coupled in a functional network. In the special case of the sensory and the motor cortex, some of these connections are luckily known in detail. For these two areas a map can be derived (see the plot of the homunculus in Figure 2.5). It reflects the observation that more or less well-defined patches of the motor- and sensory cortex are tightly coupled with body parts. As the exact mapping is subject to the ontogenesis of a human, the mappings can vary in details between individuals. The mapping was discovered when electrical stimulation of neurons of a patch led to the illusion of a touch (for sensory neurons) or even to the movement (for motor neurons) of the respective body part [PR50].

2.1.3 Neuronal Activity in the Cerebral Cortex

During the excitation of a neuron, a magnetic field arises as well. It is oriented normally to the electric field. As the fields of a single neuron are very weak, they can only be detected with close-by sensors that penetrate the neural tissue. Non-invasive recording techniques cannot detect the electric activity of single neurons but only the activity of large clustered groups of neurons that have correlated activity. Their sensors record electric or magnetic activity from outside the neural tissue and are placed either inside the skull (Electrocorticogram, ECoG) or outside the skull (Electroencephalogram, EEG, and Magnetoencephalogram, MEG)[1]. Their distance to the signal sources ranges from a few millimeters to a few centimeters.

Non-invasive recording techniques can only perform with a reasonable signal-to-noise ratio if the signal source is strong enough and located close to the sensors. The cerebral cortex (and the pyramidal cells whose dendritic trees and cell bodies are located in the cerebral cortex) show characteristics that help to fulfill these necessities:

[1] methods like EEG, ECoG and MEG are discussed in detail in section 2.2.

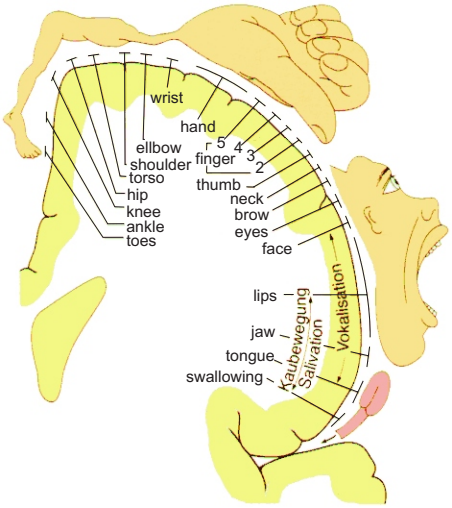


Figure 2.5: *The homunculus visualizes the mapping of body muscles to the motor cortex. The mapping is not isomorph as important areas like tongue, hands and lips are overly represented (according to [Goh97]).).*

- *orientation and position:*
  The cerebral cortex shows a layered vertical structure of about 5mm thickness. From outside to inside six layers can be distinguished: molecular layer, external granular layer, external pyramidal layer, internal granular layer, internal pyramidal layer and fusiform layer.
  Although the cell bodies of pyramidal cells are found in several of these layers, those neurons show a clear orientation: the dendritic trees are mostly situated in the outer layers while the axons are directed towards the inner layers (see figure 2.6). If a neuron is located at a gyrus of the cerebral cortex, the electric and magnetic fields caused by synaptic stimulation are close to the surface (for simplification, figure 2.6 illustrates only the electric fields). If neurons are located in one of the sulci, the fields will be more difficult to detect as they are deeper in the brain. Please note that the orientation of magnetic fields is of course normal to the electric fields.

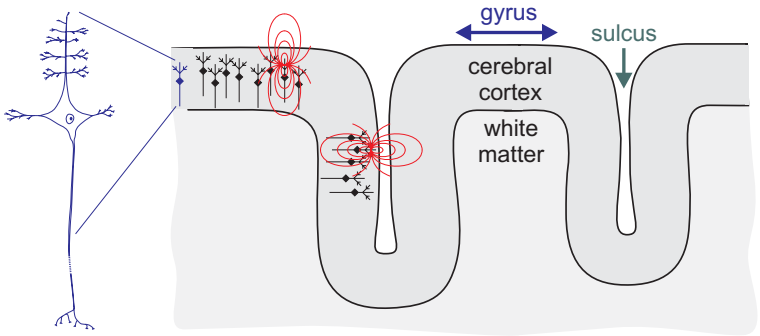


Figure 2.6: *The cerebral cortex has a folded structure with bumps (gyri) and folds (sulci). The soma and dendritic trees of pyramidal cells are situated in the cerebral cortex. These neurons have an orientation so that the dendritic trees are close to the surface. If excited, an electric field illustrated with red color emerges (and also a magnetic field with orientation normal to the electric field which is not plotted). From outside the skull, the orientation of the field changes depending on the position of the neuron (in sulcus or gyrus).*

- *field potentials*
  The cerebral cortex is organized in groups of neurons (ranging from a few neurons up to several millions per group), that form localized functional clusters or networks. Neurons belonging to the same cluster exhibit similar or even synchronized electrical activity, if the cluster is involved in a task. This functional and localized clustering leads to enlarged and stronger fields that can be detected by inductive or capacitative sensors of non-invasive methods.

Unfortunately some characteristics of the cerebral cortex and the anatomy of the human head hamper the recording:

- *folding*
  One aspect of the folding is that signal sources in sulci are possibly more distant to the electrodes than those located in gyri. Thus their signal amplitude will be smaller. Another important aspect is that sensors of some recording techniques show anisotropic sensitivity. An example are the coils of MEG, but also ring-shaped electrodes of EEG. They can best detect changes in the electric or magnetic fields if these have a matching orientation and perform worse otherwise. As the generated fields in sulci and gyri are approximately orthogonal to each other, such a recording method usually performs well for either the fields generated in gyri or the fields generated in sulci, but not for both.

- *meninges, skull and scalp*
  Three membranes separate the brain from the encapsulating skull: the dura mater, the arachnoidea and the pia mater. Together with the skull, the membranes protect the brain. Between the arachnoidea, which has a spongy structure, and the pia mater, which is situated directly on top of the cortex, an interspace is found. It is filled with cerebrospinal fluid and helps to absorb shocks. On top of the skull (thickness approximately 1cm) a layer of fat and the skin form a barrier of a few millimeters. For surface-EEG recordings that capture the electrical activity of the cerebral cortex, these obstacles have to be overcome by the signals. The combination of skin, fat, bone, meninges and cerebrospinal fluid attenuate signal amplitudes and act as a low-pass filter. Due to volume conduction, with growing distances between any sensors and the signal sources, the exact location of the sources gets more difficult.

In practice, the resulting orders of magnitude for the recording of electric field potentials with EEG is in the range of approximately ± 30 microvolt, for ECoG approximately ± 200 microvolts and approximately ± 50 femto Tesla for magnetic fields recorded with MEG.

Other imaging techniques like functional magnetic resonance imaging (fMRI) or position emission tomography (PET) try to capture the neural activity via indirect effects like changes in the blood oxygenation level (called the BOLD effect). As these techniques show very poor time resolutions they will not be considered in this work.

2.2 Recording Techniques

Available functional recording techniques differ much in their time and space resolution, in the component of the overall central neural activity that can be assessed with a technique and in their suitability for handicapped patients.

In the following sections, three recording techniques will be introduced: the electroencephalogram (EEG) which is typically used for BCI systems and two newer recording techniques, the magnetoencephalogram (MEG) and the electrocorticogram (ECoG). All three systems allow the calculation of feedback signals in real time and have a high temporal resolution. The three methods mainly show differences in the applicability, the degree of invasiveness necessary and in the component of neural activity that is reflected.

2.2.1 Electroencephalogram (EEG)

The electroencephalogram (EEG) is an extra-cranial non-invasive recording technique that is sensitive to changes in electrical fields generated by neuronal activity. It was first discovered by Hans Berger in 1924 and published in 1929 [Ber29]. Soon after the discovery of EEG it was realized that EEG signals can reflect physiological states like sleep or wakefulness. New exploratory methods opened further possibilities: by averaging hundreds of short EEG signal sequences gained from repetitions of constant experimental conditions, it became possible to examine the reaction to various stimuli.

EEG signals are picked up capacitatively by small disc-shaped electrodes that are stuck to the scalp with a contact gel. The electrodes and gel can easily be removed or washed off after the recording session. The EEG signal is usually recorded at many locations simultaneously by one electrode at each position (the term *channel* is often used to refer to a recording position).

For the application of a larger number of electrodes the use of EEG caps is a big help. The distance between neighboring electrodes is usually in the range of one to a few centimeters and available EEG caps can record up to 128 channels. The time necessary for an assistant medical technician to prepare a subject is 10 to 60 minutes and depends on the number of electrodes that have to be applied. Although caps usually follow the extended 10-20 standard positioning system [Ame91] and improve the reproducibility of an EEG montage, the recording positions can show a few millimeters of variation for successive montages. Typical recording sessions with healthy subjects can last up to a few hours if the experimental tasks do not require much concentration. For clinical reasons, extremely compact and portable EEG systems

are in use that record EEG activity for 24 hours or more. But normal EEG recording systems are also portable enough in the sense that they can be applied at a patient's home without bigger problems, even though a shielded environment is favorable (see below).

EEG signals are electrical potentials that are determined at positions relative to one or more reference electrodes. While the wanted signal channels are usually spread over interesting cortical areas, the reference electrode can be placed at the earlobe or the tip of the nose. Figure 2.7 shows a montage scheme for an EEG recording.

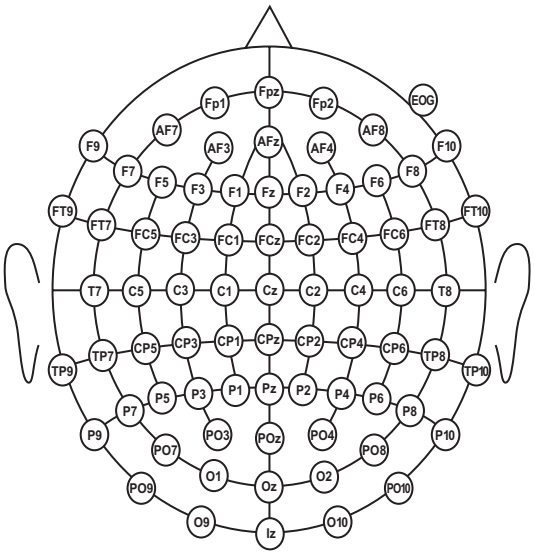


Figure 2.7: *Schematic top view of a standard EEG montage.*

As the recorded signals are in the order of ± 100 μVolt an amplifier enlarges the used signal range to e.g. ± 5 Volt before the signals are sampled for the computer. Typical sampling rates for EEG are in the range of >100 Hz and up to a few thousand Hz. Sampling frequencies above 256 Hz are enough to track typical EEG patterns that are reported to have frequency components of 0 Hz to approximately 100 Hz. Most analysis techniques can not recognize frequencies above approximately 40 Hz as the signal-to-noise ratio decreases quickly for higher frequencies and as time-locked averaging methods fail.

EEG recordings have a very good time resolution but suffer from disadvantages that are mostly caused by the skull bone, the meninges and the intra-cerebral liquor. These layers act as a low-pass filter to the electrical fields and lead to the general notion of approx. 100 Hz for the upper limit of contained frequencies. Furthermore these layers act as a spatial low-pass filter. The spatial resolution is thus not necessarily limited by the distance between electrodes (usually approx. 2 cm) but additionally by this smearing effect.

Artifacts are signal components picked up by EEG electrodes that are *not* caused by neural activity. As artifacts can be so strong in amplitude that interesting signals are not detectable any more, they are not desirable. Even if weaker in amplitude, artifacts can alias for true EEG signals (especially in low frequency ranges). The fact that artifacts are picked up with highest intensity at electrodes closest to their origin can help to identify them.

Typical artifacts in EEG comprise:

- Muscle activity
  This class of artifacts is caused by e.g. clenching jaws, tongue movements, facial and neck muscles. Except for tongue movements, they can easily be recognized and at least partially filtered due to strong frequency components bigger than 50 Hz and high amplitudes of up to several hundred μV. Tongue movements are slower and can be misinterpreted as certain kind of neural activity (slow cortical potentials, SCP). Subjects are instructed to avoid all kind of movements during the recording intervals.

- Movements of the eyeball
  This type of artifact is caused by the dipole characteristics of the eyeball. Eye movements of healthy subjects typically reflect saccades in their time series, but these can be missing in patients. The contained frequency components are rather slow and difficult to distinguish from certain types of neural signals like SCP. To avoid the masking of neural signals by eye signals, subjects are usually asked to fixate a mark during the EEG recordings.

- Eye blinks
  Eye blink signals are very short peaks that show an amplitude higher than the neural activity. They can easily be detected via visual inspection of the time series but have rather inconspicuous frequency components. Subjects are usually instructed to avoid eye blinks if possible during EEG recordings but are shown explicit time intervals that can be used instead for blinking.

- Stray pick-up from exterior signal sources
  This class of artifacts is always present for unshielded recording environments and mostly comprises 50 Hz and 100 Hz humming from electrical sources and devices, rectifiers etc as can be found in power supplies, artificial ventilation devices, food pumps etc. The influence of stray pick-up can be decreased by careful application of the EEG electrodes. The lower the impedance of the electrodes, the less exterior signals are picked up.

Most artifacts can be controlled by proper instruction of the subjects, by using additional control electrodes close to possible artifact locations and by proper frequency filtering of the recorded signals.

2.2.2 Electrocorticogram (ECoG)

The electrocorticogram (ECoG) is a newer method [JP49, TDT+94]. It records the electrical brain activity with the help of an electrode grid directly on top of the cortex. Figure 2.8 shows a recording of four channels. ECoG is an invasive method in the sense that the skull must be opened for the application of the sensor grid. As the cortex itself is not punctured or harmed in any way, the level of invasiveness is not comparable to recordings with e.g. multi-electrode arrays that are inserted into the brain matter [CLC+03, HSF+06]. However, the implantation necessary for ECoG includes the risk of infections and is thus more difficult to justify than the use of non-invasive methods.

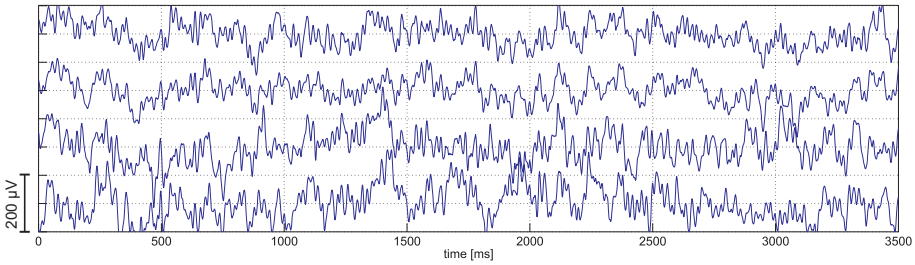


Figure 2.8: *This plot shows ECoG recordings from 4 channels in high resolution (1000 Hz). The distance of two horizontal lines decodes 100 μV. The amplitude of the recordings ranges roughly from -100 μV to +100 μV which is on the order of five to ten times the amplitude measured with extra-cranial EEG.*

The ECoG experiments presented in this work were not performed with healthy subjects or paralyzed patients but with patients suffering from certain forms of epilepsy. Those patients have many seizures and unfortunately do not respond to medication with anti-epileptica. One way of treating them is to surgically remove neural tissue that can be identified as the so-called focus, an area that is involved in the generation of epileptic seizures. This treatment usually reduces the number and intensity of epileptic seizures. The prelude of the focus removal comprises the monitoring of electrical brain activity for several days in order to identify the exact location of the focus. For this surveillance, patients undergo an implantation of ECoG grids. For collecting ECoG data, those already implanted patients were asked to participate in BCI experiments during the monitoring interval. It must be mentioned that the implantation position was entirely determined by the estimated location of the foci and thus is usually not optimal in terms of BCI. The electrode grid is a flexible gel plate. The disc shaped electrodes of approximately 0.4mm diameter are embedded in the gel plate in the form of a linear or two-dimensional array. The distance between two electrode centers is 10mm. Typical grid shapes are (8x8) or (1x4). The grid stripes can be applied in sulci of the cortex, while the two-dimensional grids cover larger parts of the cortex. The distance between an electrode and the cortex tissue is in the range of 1mm only.

The characteristics of data recorded via ECoG (sometimes called cortical EEG) are similar to that recorded with EEG (sometimes called scalp EEG) but show a few very important advantages. As the electrodes which are manufactured from either stainless steel or platinum are positioned under the skull and the meninges, there is less dampening tissue between the signal sources and electrodes. The neural activity is recorded without the low pass filtering caused by these layers. Contrary to EEG, ECoG signals additionally show stronger frequency components clearly above 40 Hz and have a higher spatial resolution. The signal-to-noise ratio of ECoG is reported to be several orders of magnitude higher than that of scalp EEG [RLD+02].

While EEG recordings are always prone to muscle artifacts, ECoG recordings are not so easily corrupted by muscle activity. For short-term implants, like in the case of the epilepsy patients, ECoG electrodes can be applied with very low impedance, and the surrounding skull bone provides shielding against muscle artifacts.

Like EEG, ECoG can in principle be applied for completely paralyzed and artificially ventilated LIS and CLIS patients as well. However, only very little experience has been collected with long-term ECoG implants in humans. In animal experiments ECoG implants have been used over periods of several months.

2.2.3 Magnetoencephalogram (MEG)

Neuromagnetic activity was first measured by David Cohen in 1968[Coh68] using a coil sensor in a shielded room. Magnetoencephalography (MEG) recordings are different from EEG and ECoG in the way that it is not the changes of the electrical component of the electromagnetic field that are registered but changes of the magnetic component. As the magnetic fields generated by neural activity are very small (in the order of some femto Tesla (fT), which is a factor of 10^{-6} compared to the magnetic field of the earth) they can only be detected in completely shielded environments. Inside a shielded MEG room no disturbing ferromagnetic materials may be used. Figure 2.9 shows the shielded room built by Cohen at MIT, where a recording with one Superconducting Quantum Interference Device (SQUID) was first performed. SQUID sensors pick up the changes in magnetic field by induction. They are contained in a big helmet-like

Figure 2.9: *The first MEG system using a SQUID in a shielded room was operated at the Francis Bitter Magnetic Laboratory at MIT by David Cohen, Ed Zimmermann and Jim Zimmermann in 1969 (reproduced with friendly permission of D. Cohen).*

construction called dewar whose interior is cooled by liquid helium to $5°K$. The dewar is covered by layers of thermal materials that allow a subject to have direct contact with the dewar surface. Currently a dewar can contain over 300 sensors. The sensors are approximately 2cm apart from each other and, due to the insulation, approximately 3cm distant from the head surface.

In section 2.2, it has been introduced, that EEG, ECoG and MEG record different components of the same underlying electric currents. While EEG/ECoG is sensitive for both radial and tangential currents, MEG is sensitive only for tangential currents that under normal circumstances take place in sulci[2]. The magnetic field is less affected by the skull bone, skin and meninges than the electric field, which prevents smearing and leads to signals with stronger components at higher frequencies. Together with the shielding of MEG rooms, the signal-to-noise ratio of MEG recordings is better than that of EEG. In addition, DC components can be recorded more effectively by MEG than by EEG [CH03]. A study performed in 1990 reported that the spatial resolution of MEG is slightly better (8mm) compared to scalp EEG (10mm) for the localization of implanted electrical sources[CCY$^+$90]. In the meantime, the technical development of both methods has improved, but also later studies report better spatial resolution of MEG (see [CH03] for an overview).

[2]This is of course reversed for areas, where the cortex turns away from the skull.

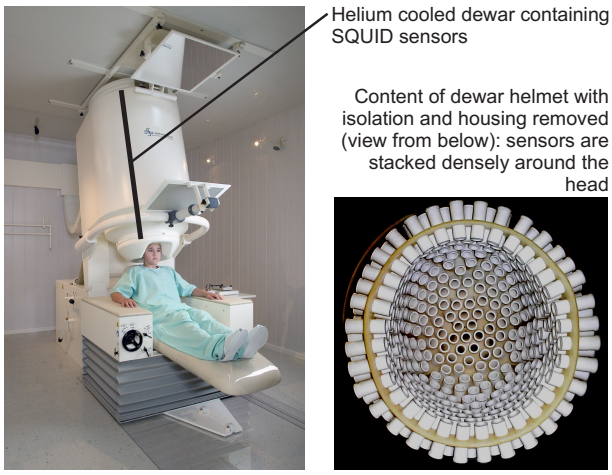


Figure 2.10: *MEG recording system. The left image shows the dewar inside a modern shielded room. It contains hundreds of SQUID sensors. To avoid noise from the sensors, the SQUIDS are kept at very low temperature. The dewar is isolated against the subjects head with a vacuum layer. The right image shows the dense arrangement of sensors inside the dewar helmet. (Images reproduced with kind permission of VSM MedTech Ltd.)*

In practice, MEG is a very quick method as it eliminates the time for electrode attachment. Algorithms developed for EEG signals are applicable for MEG without problems and vice versa. Like EEG, MEG suffers from movement and eye artifacts and the exact re-positioning for successive sessions is not possible (like in EEG). Unfortunately MEG is a very expensive method (an installation including the shielded room costs 2-3 million euros) and is not available as a compact, portable or home system. This restricts the usefulness for BCI experiments to either the developmental phase (when healthy subjects can be used) or the training of patients that are not completely paralyzed and mobile enough to be taken to an MEG room.

The experimental setup for MEG is more elaborate than for EEG/ECoG as stimuli or additional sensors have to be implemented without usual electrical components.

2.3 Extracting Features from Brain Activity Signals

The activity of the brain during an experiment is usually assessed at many positions simultaneously. These positions are determined by the applied recording technique or are defined by the user. The recording of changes of the magnetic field for example is only possible at fixed positions defined by the construction of the recording helmet of the MEG apparatus, for EEG the usage of standardized caps with e.g. 64 predefined electrode positions over the scalp or the application of fewer electrodes at desired positions is practical. For ECoG a standardized positioning system does not (yet) exist. Electrodes are usually applied at the focus of attention in the form of 1d-strips (for e.g. 8 electrodes) and 2d-grids (for e.g. $8 * 8 = 64$ electrodes). The electrodes in these grids have fixed, regular positions. The grids are applied under the meninges directly on the surface of the cortex, including the sulci. For the duration of the experiment, the changes of electric or magnetic activity are recorded at every electrode. The sampling rates are chosen problem dependent but are typically performed with at least 100 Hz, ranging up to 1 kHz. For most BCI paradigms, the recording of so-called trials, intervals of a few seconds, forms an (atomic) data sequence that describes brain activity during a short mental task (see section 5 for details on experimental paradigms). Figure 2.11 shows a time series of 5 seconds duration recorded via EEG for one single EEG channel.

Estimated conservatively, the recording of a trial leads to several thousand samples (e.g. for 32 channels, 128 Hz sample frequency and 3 seconds trial duration). The simplest representation of the data of a trial (e.g. for classification purposes) would be to concatenate the time series data from all channels and form a very high-dimensional data vector x (for the above example: $x \in \mathbb{R}^{12288}$). This representation is not only highly unpractical, memory consuming and slow, but in addition shows further disadvantages:

1. The data dimensions are not independent: neighboring channels might partly share strong signal components, and succeeding time sample points are very similar.

2. Most algorithms do not perform well on very high-dimensional data. Thus a compacting of the data space is desirable.

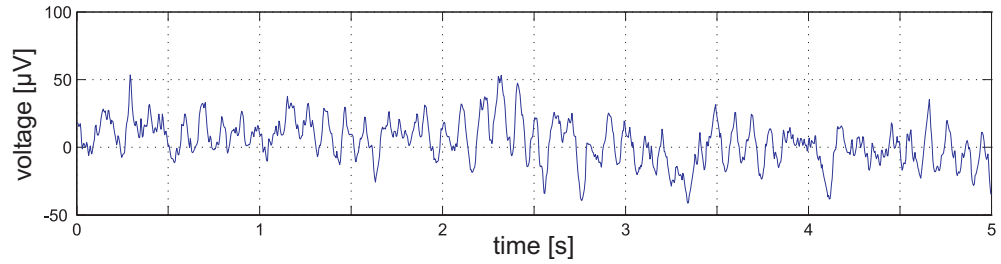


Figure 2.11: *Example of one channel of an EEG recording of a healthy person. The time series was recorded over the motor cortex with 256 Hz sampling frequency.*

3. The time series recordings can contain a good portion of noise (e.g. from analog components like sensors and amplifiers or noisy stray-like line humming picked up in unshielded environments) or artifacts (e.g. DC components from eye movements contained in recordings of frontal sensors, high frequent muscle artifacts in recordings of sensors close to jaw muscles etc.). Transforming the time series to another representation can be useful to decrease or delete these components.

The number of possible transformation methods to tackle these problems is huge. Although dimensionality reduction itself is a very important point, the transformation should in addition fulfill the objective to reveal characteristics or *features* of the data that are possibly *relevant* for the specific BCI task. From this point of view, three main feature types will be dealt with in the next sections: time series features, spatial features and frequency features. Features describing the complexity of brain signals (see for example [FLA+06]) are very rarely used in BCI context and will thus not be introduced here.

2.3.1 Time Series Features

The description of an EEG signal like the one in figure 2.11 with time series features is straight forward. Most of these features can easily be computed:

- The average of the signal (offset)
- The linear trend of the signal
- Absolute minimum and maximum values
- Number and order of local minimum and maximum values
- Weight factors describing the matching and positions of predefined patterns
- Slopes/steepness/height/width of predefined patterns

The last two types of features can be applied with either general pattern families (e.g. for wavelet analysis) or for EEG specific patterns. Possible applications are the detection of sleep stages by analysis of the α-wave intensity (see a list of typical EEG components in Section 3.4.3), the determination of spike foci via analysis of spike forms or the study of event-related P300 waves that can be observed approximately 300 ms after external auditory or visual stimulations. [Birbaume-EEG-Buch]

Linear trends and offsets are applied to the so-called slow wave user training with LIS patients [KNK+01]. The application of pattern matching methods often presumes that the experiment follows a well-known paradigm for which a typical EEG signal answer is expected. Although this is true for some BCI paradigms, another drawback arises: (except for linear trends and offsets) most of the above time series features are difficult to observe directly in single trial studies (e.g. P300) and can be clearly revealed only by averaging of many trials. Only in combination with good classification algorithms can these features be useful for BCI.

2.3.2 Spatial Features

The sensor positions of a BCI system can be considered spatial features. For recording techniques like EEG and MEG standard positions are given that are often applied by default. For experiments with ECoG usually all sensors of the implanted grid are recorded. The use of a subset of all available sensor positions can easily be performed and does not

involve computation. Apart from choosing a subgroup of sensor positions, a blind source separation (BSS) approach can be feasible for multi-channel recordings. The most popular algorithm class is the independent component analysis (ICA). It is commonly used for the separation of line humming and other artifacts.

2.3.3 Frequency Features

A time series can be described by its spectral characteristics, typically by estimating the so-called power spectral density (PSD) of the signal. The PSD (which will also be called *spectrum* in further sections) can be used to identify important frequency components, e.g. components that change with the mental activities of the user. For BCI signals, the spectral analysis is an important method as the brain is known to generate task-dependent activity in relatively small frequency bands. For EEG recordings the relevant activity is contained between 0 and approximately 40 Hz, while MEG and ECoG recordings can show activity even in higher frequencies. For an explanation of these differences see Section 2.2. A difficulty arises if such components of the PSD are hidden in the broadband noise of the time series. Figure 2.12 is an estimate of the PSD of the time series in figure 2.11. Starting with high power for the slow frequencies, the PSD

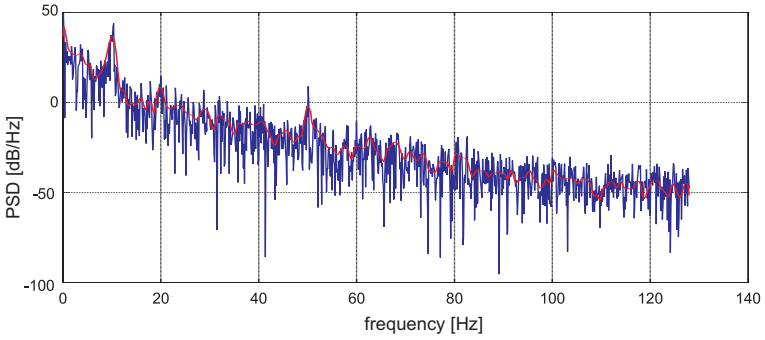


Figure 2.12: *Estimate of the power spectral density (PSD) of a 5 second time series recorded via EEG at 256 Hz sampling rate. The blue, noisy estimation was generated by a periodogram with rectangular window. The red estimate stems from the Welch algorithm [Wel67] which averages the PSD estimates from several shorter-time, overlapping periodograms with Hanning windows.*

reveals a strong decay towards higher frequencies (approximately 50dB for the first 40 Hz). In the figure two frequency components can be identified as they show relatively high powers: 10 Hz and 50 Hz. The 10 Hz component is one of the typical EEG rhythms over the motor cortex (sometimes accompanied by a further peaks at approximately 20 Hz and/or 30 Hz), while the second peak is humming induced by the power supply.

The existing methods for estimating the frequency characteristics of a time series signal can broadly be put into three groups: Fourier methods, parametric methods and subspace methods [Kay88]. The Fourier methods and the parametric methods are introduced in the following subsections. The computational complexity of the subspace methods being higher than for the other two groups, they will not be considered in this work.

Fourier Methods

The first group contains the most common methods based on the discrete Fourier transform or the fast Fourier transform (FFT). For the analysis of a time series, these methods make the (usually wrong) assumption, that the underlying random process is zero outside the interval of interest. Examples for these methods are the periodogram, the Welch method and the multitaper method. The simple example in figure 2.12 was generated by two algorithms of this group. The algorithms are generally restricted to detect frequencies that maximally are half the sampling frequency (the so-called Nyquist frequency). The PSD estimated by the periodogram (blue graph in figure 2.12) can be calculated very quickly via the FFT. If used without windowing, its PSD shows very slim peaks. A drawback of this method is the strong *leakage* effect - side lobes close to peaks in the spectrum that in practice can obscure other, smaller peaks. This negative effect unfortunately becomes stronger with shorter time series. As the time series analyzed for BCI systems are typically very short, the problem of leakage is very common. Using a windowed version of the periodogram (e.g. by applying a Hamming window to the time series) decreases the height of the side lobes (p.70 in [Kay88]) and leads to increased peak-to-noise ratios for sinusoidal signals embedded in white noise. Unfortunately the tradeoff for increased SNR is a decrease of frequency resolution as peaks of the PSD are enlarged by windowing. Thus the resolution of two neighboring peaks gets more difficult. The periodogram in addition shows a frequency bias, which means that the expectation of

estimated PSD peaks of the periodogram does not reliably converge to the true PSD peaks by repeated estimates. The PSD estimated with periodogram has a high variance, which unfortunately is independent of the length of the analyzed time series but can be tackled with averaging methods.
The Welch algorithm [Wel67] reduces the strong variance of the periodogram's PSD with an averaging approach. It divides the signal into several (8 in this example) half-overlapping segments and multiplies each segment with a function (e.g. Bartlett window or Hamming window etc.). It then estimates a separate PSD via periodogram for each segment. Averaging the PSDs leads to the red estimate in figure 2.12. The benefits of the Welch method are paid for with increased computation time and increased bias: the frequency resolution is decreased as the resulting peaks have become wider and might eventually overlap with neighboring peaks.
Multitaper methods use a combination of multiple, orthogonal windows to estimate the PSD of a signal. Although their application can reduce leakage and variance even further, they are not applicable for BCI systems due to their relatively large computation time.
For the application of Fourier based methods it is assumed that the time series is repeated (for the case of implicit rectangular windowing) or that it is zero outside the interval of interest (for the case of window functions that approach zero at the boundaries). This assumption leads to smeared spectral estimates. To conclude, all Fourier methods have to realize a trade-off between bias and variance. For long signal sequences, they may still give reasonable results, but performance deteriorates with shorter sequences.

Parametric Methods

Parametric methods employ *a priori* assumptions about the generating random process. Examples for this knowledge could be that only sinusoidal components are contained in the underlying white noise or that even the number of components are known, that only broadband or only narrowband activity is expected etc. Depending on these prior assumptions, a model class and model order can be chosen that is hopefully able to capture the characteristics of the signal. If the choice is incorrect, systematic error and bias is possibly introduced. Otherwise the parameters of the model can be estimated so that the model describes the time series well.

As an example for parametric methods the autoregressive (AR) model will be described in detail. The idea of the autoregressive (AR) method is quite simple: assumed that the time series $x[n]$ is an AR process of order p, hence called $AR(p)$; Then the so-called *linear prediction formulation* of AR predicts the unobserved sample n of the time series x based on the linear combination of its p known past samples $x[n-1]$, $x[n-s]$, ... , $x[n-p]$.

$$x[n] = -\sum_{k=1}^{p} \alpha[k]x[n-k]$$

The time series of interest x is usually not a pure AR process but rather a noisy process. In this case, $x[n]$ is approximated by an AR model $\hat{x}[n]$ plus noise:

$$x[n] = \hat{x}[n] + v[n]$$

The fitting of an $AR(p)$ model to a time series of length n turns into finding the prediction coefficients

$$A = (\alpha[1], \alpha[2], ..., \alpha[p])$$

(also called model parameters or weights of the linear combination) so that they minimize the forward prediction error e:

$$e = \frac{1}{n-p}\sum_{i=p}^{n}(x[i] - \hat{x}[i])^2$$

Usually the order p is much smaller than n so that the optimization task of finding the AR parameters leads to an overdetermined system of linear equations. The least squares solution A can be determined by various standard regression algorithms and will not be discussed in detail. Available implementations of AR models typically make use of two modifications. First, the AR parameters are determined by minimizing simultaneously the forward and the backward prediction error. This doubles the number of equations in the linear system but allows for slightly better estimates of A especially for very short time series. Secondly, a small speedup is gained by using a slightly modified variant of the above formulation. It involves the precalculation of the autocorrelation r_{xx} of the time series x. The forward-/backward prediction is then expressed based on $r_{xx}[n]$ instead of directly on $x[n]$. The resulting linear equations are called Yule-Walker equations and are solved efficiently with the Levinson algorithm [Kay88] in $O(p^2)$ compared to $O(p^3)$ for the Gaussian Elimination.

Parametric methods show substantial advantages compared to the Fourier methods:

- Applicability for short time series
  Short time series provide less information than longer time series. Thus every spectral estimation method delivers less exact estimates. Parametric methods however incorporate a priori knowledge and do not have the same degree of freedom as the Fourier methods, as their model order is usually small compared to the frequency resolution of the PSD. The problem of finding the right model parameters is therefore better defined and a resulting PSD estimate is smoother than PSD estimates of most Fourier methods. In addition, the application of parametric methods

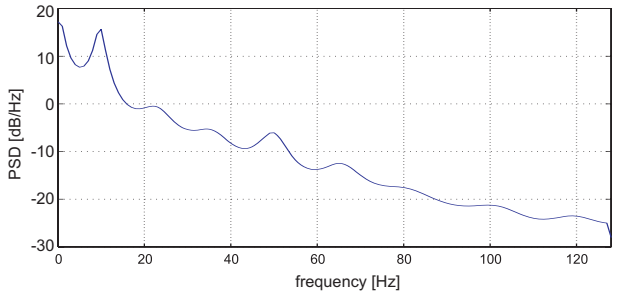


Figure 2.13: *Estimate of the power spectral density (PSD) of a 5 second time series recorded via EEG at 256 Hz sampling rate. The estimation was generated by an autoregressive process of order 20.*

  does not involve windowing of the data sequence like it is often necessary for Fourier methods. Windowing, e.g. the multiplication of a time series sequence with a smoothing function, attenuates the values at the boundary of the window close to zero. This reduces the information available in the data sequence. Thus parametric methods capture the frequency spectrum of short time series quite well compared to the Fourier methods.

- The variance of repeated PSD calculations is smaller.
- The problem of side lobes, as with periodogram, does not exist.
- Peaks in the spectrum can be more narrow for parametric methods, as the smearing induced by windowing is omitted.
- The modeling of a time series by a parametric method is a strong reduction in dimensionality. If important characteristics of the time series are captured by the model, this is a good starting point for further processing. For example, a single channel time series of 3 seconds duration sampled at 256 Hz shows 768 dimensions that of course are not independent. Fitting a model of e.g. order 20 reduces the dimensionality to 20. Although some information of the time series is lost by this transformation, the frequency spectrum is represented in a quite detailed manner by the 20 AR coefficients. For an example, see figure 2.13 which shows the PSD of an autoregressive (AR) model of order 20 to the time series in 2.11. It is much smoother than the PSD generated by the periodogram or the Welch algorithm in figure 2.12 but still captures the movement related spectral peaks and the 50 Hz humming.
- Smoothed denoised PSDs
  The AR method delivers smooth and denoised spectral estimates when applied with small model orders. With growing model orders however, the spectrum is represented with more and more details and eventually includes additional noise.
- Fast calculation
  Depending on the specific implementation, the fitting of an AR model to a time series takes about the same time as the calculation of its FFT.

For the application of parametric models it is wise to start with the removal of linear trends contained in the time series. Then a model class and a model order are chosen depending on the domain knowledge about the data generating process. In the next step, the model parameters are estimated for the time series. Either the parameter values can now directly be used as a condensed new representation of the time series (e.g. for classification, see section 2.4) or for estimating the PSD. In this case, the parameters are substituted into the theoretical PSD expressions of the model class.

Examples for parametric methods are autoregressive (AR) method, the moving average (MA) method, or the combination of these two methods (ARMA). The methods have different application areas. The AR model is all-pole (see [Kay88] for an introduction to filter theory) and thus well-suited to describe processes that consist of narrow band

(ideally: sinusoidal) activity embedded in broadband background activity. An example for such activity is the spectrum in figure 2.13 that shows two peaks at 10 Hz (a typical EEG rhythm) and 50 Hz (power line humming). For many descriptions of EEG, ECoG and MEG spectra, these type of processes are prevalent. The MA method (all zeros) is well-suited to model processes that reveal sharp valleys in the spectrum but not for processes containing sharp peaks. ARMA models are a combination of these two models (poles and zeros) and can model peaks and dips of spectra, but need more parameters than AR or MA models. For the description of non-invasive brain signals like those used in BCI systems, the autoregressive method is often applied. It seems to be sufficiently powerful to model typical rhythmic and broadband activity [Kay88]. The compact description of an EEG signal by an AR model is possible by compressing the original data but of course some information (hopefully noise) is lost during this process. The training of the model only takes the reconstruction error into account - the AR fitting process does not incorporate knowledge about the discriminative value of the information. This could in principle pose a problem for a following classification task. To avoid this, the optimal AR model order (and therefore the compression rate) can be determined by validation techniques.

2.4 Classification

This section establishes the notion of a classification task as it is understood in machine learning. After some notation issues, a very powerful learning method, the support vector machine is introduced together with issues of model selection and proper estimation of the generalization error of such a learning method.

2.4.1 Classification Task

Given a number of possibly high-dimensional data points and class labels for these points, a generalized mapping of the input space to the class labels must be learned.

The classification task is to decide for an unlabeled data point which class it does belong to depending on its features. For this decision a classifier is needed that incorporates knowledge available from the labeled examples. The acquisition of this knowledge is called either the machine learning phase or the training phase of the classifier. A trained classifier delivers a rule or set of rules that defines a mapping from the input space to the class labels. In many real valued classification tasks, the rule can be formulated as a classification function or hyperplane of the input space. Figure 2.14 illustrates a possible set of training examples for a two-class problem. Given the coordinates of some apples on the ground and their class label (here: the apple's breed), the class label of the remaining apples is wanted. The data are represented as points in $\mathbb{R}^2$ and a set of possible linear and one nonlinear classification functions (hyperplanes) that separate the data is shown. It is clear that among all possible hyperplanes, we should choose the one that not only

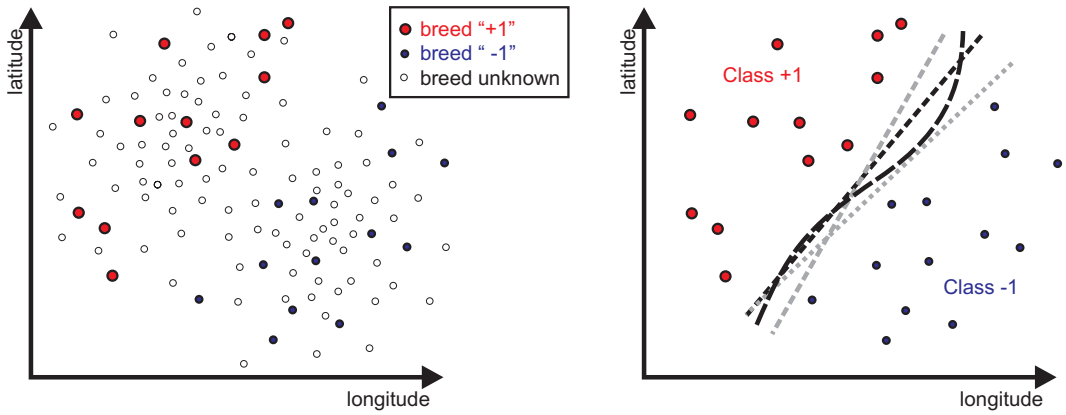


Figure 2.14: *Position of apples in the garden as an example for linearly separable training data. Left plot: red and blue dots represent positions of apples whose breed has been determined by tasting (training data), while gray dots represent apples whose breeds are to be determined by a classifier (test data). Right plot: Several classification functions (linear and nonlinear) can separate the colored training data.)*

classifies the training data well (colored points) but shows minimal error on the remaining apples. In practice, the exact class distributions of the data is unfortunately unknown and they might be overlapping such that an optimal solution of the classification problem is very difficult. Before proceeding to an introduction of the support vector machine (SVM), an example of a classification algorithm, it's worth taking a look on the situation in BCI.

2.4.2 Notation

Let n denote the number of training vectors (trials) of the data sets and let d denote the data dimension which is for example the number of channels times the number of features per channel due to the hierarchical feature representation in a BCI context. The training data for a classifier is denoted as $X = (x_1, ..., x_n)$ with $x_i \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n \times d}$. Corresponding class labels are $Y = (y_1, ..., y_n), Y \in \{-1, 1\}^n$. The terms *dimension* and *feature* are often used synonymously throughout this thesis. In some cases, the group of features that belong to a channel are called either a feature or a channel, if for example the channel is the hierarchical level of interest to work with. These cases can be easily recognized by the context.

2.4.3 Support Vector Machine

For the focus of this thesis, a stable classification method is needed that allows a comparison between different feature selection strategies. The Support Vector Machine (SVM) is a classification technique developed by V. Vapnik [Vap95].

The central idea of the SVM is to separate data $X \subset \mathbb{R}^d$ from two classes by finding a weight vector $w \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$ of a hyperplane H:

$$\begin{aligned} H : \mathbb{R}^d &\rightarrow \{-1, 1\} \\ x &\mapsto \mathit{sign}(w \cdot x + b) \end{aligned}$$

From all possible hyperplanes, the one with the largest margin is chosen (see left plot of figure 2.15). If the data X is linearly separable the margin of a hyperplane H is the distance of the hyperplane to the closest points $x \in X$, the so-called support vectors (SV). Ideally these are only a small fraction of the training data points. As the hyperplane is formulated only based on these SVs, the algorithm delivers a quite sparse solution. Apart from being intuitive, the idea of maximizing the margin has been shown to provide theoretical guaranties in terms of the generalization ability of the classifier[Vap95].

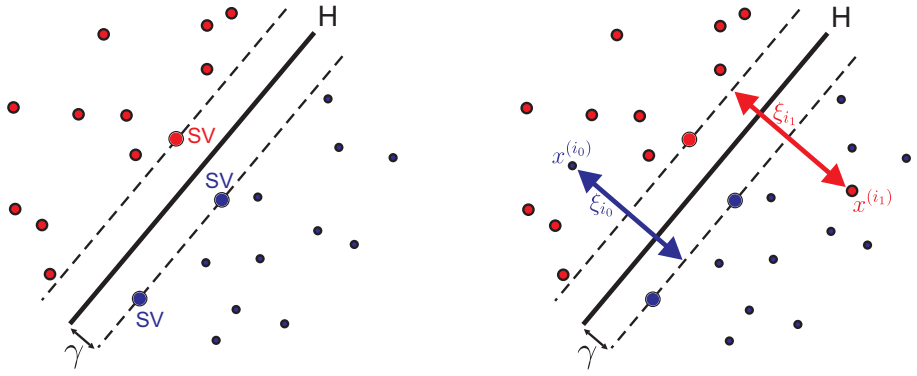


Figure 2.15: *Left plot: For linearly separable data, the support vector machine chooses the hyperplane H that leads to the largest margin. Support vectors defining the hyperplane are plotted in encircled dots. Right plot: If the training data is not linearly separable, the SVM has to perform a trade-off between large margin and small training error. Please note the addition of two points.*)

A variant of the SVM algorithm deals with the more natural case when the data is not linearly separable. The classification problem is then formulated to solve the following optimization:

$$\begin{aligned} \min_{b, w \in \mathbb{R}^d} \quad & \|w\|_2 + C \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \quad & y_i(w \cdot x^{(i)} + b) \geq 1 - \xi_i \quad (i = 1, ..., n) \end{aligned} \tag{2.1}$$

The parameters ξ_i are called slack variables. Including them in the optimization problem leads to a solution that allows misclassified training examples (small in number and with a small distance to the hyperplane). It ensures that the problem still has a solution in case the data is not linearly separable (see right plot in figure 2.15). In practice it is often not known for a problem, whether it it linearly separable or not. In this case it is a good idea to use the slack variable formulation

as it can improve the generalization ability of the solution.
The margin is defined as

$$\gamma(X, Y, C) = 1/\|w\|_2$$

so that the left part of the objective function in 2.1 leads to the maximization of the margin γ.

In practice a trade-off has to be made between two objectives. One is to gain a lower error on the training data, so that fewer and smaller slack variables ($\sum \xi_i^2$) are preferred. In the extreme case, the model is overfitted to the training data and will behave badly on unseen test data. The other is to prefer a larger margin γ that might come along with a higher error on training data but possibly results in better generalization to unseen test data. This trade-off is controlled by the regularization parameter C. Finding a good value for C is part of the model selection procedure (see section 2.4.4. The value $2/C$ is also referred to as the *ridge*. For a detailed discussion please refer to [SS02].

The SVM formulation can easily be expanded to the nonlinear case via the so-called kernel trick. As only the linear SVM is necessary for further understanding of this work the nonlinear formulation is omitted here but can be explored in [SS02].

2.4.4 The Need for Model Selection

Model selection comprises the choice of diverse parameters that can be set for algorithms involved in the classification task. This does not only comprise parameters of the classification algorithm itself (e.g. parameter C for the linear SVM) but also parameters for feature extraction (see section 2.3) or for feature selection (section 2.5).
Those parameters that are not available through expert knowledge or *a priori* knowledge of any kind have to be estimated from the data itself. Correct estimation is crucial for good classification results. It can be performed by various validation techniques like cross-validation, leave-one-out validation (LOO) or bootstrap. There is no general best approach for validation, the used strategy rather depending on the number of training data points and calculation time available. For the model selection in this thesis, cross-validation has been favored.

2.4.5 Generalization Error Estimation

The generalization error of classifiers was estimated via 10-fold cross-validation (see Figure 2.16). Throughout the thesis linear SVMs are used. For regularization purposes a ridge was applied on the kernel matrix which corresponds to a 2-norm penalty on the slack variables [CV95].

2.5 Algorithmic Feature Selection

For BCI systems, the learning problem is more complicated than for the toy example of Section 2.4.1. During experiments, the brain signals of a subject are recorded via EEG, ECoG or MEG at 100 or more positions. The class label for a specific recording is determined by a specific imagination task the subject has performed during that time. The task and with it the class label is known for each recording as the subject has been instructed during the experiment. Usually, the raw time series signals are not classified directly. Instead, higher-level frequency features like AR coefficients introduced in section 2.3.3, or others introduced in section 2.3 are extracted from the pure time series of all the recording channels before classification is performed. Compared to two dimensions that describe the apples (their position in the garden), brain recordings show a much higher dimensionality - the feature vectors easily have several hundred dimensions.

In contrast to selecting features by hand based on prior knowledge, feature selection algorithms select or omit features depending on some kind of performance measure so that the feature selection process results in a *good* subset of features according to this measure. During the selection process, different search strategies are possible. As the number of feature subsets is combinatorial, a *full search* through all possible subsets is usually not feasible. As a side remark, Blum and Rivest have shown that many problems related to feature selection are NP-hard [BR92].
The so-called *forward search* methods (starting with one feature and iteratively building larger feature sets) and the *backward elimination* (starting with all features and iteratively removing features) are the most common feature selection strategies and exist in various algorithmic implementations.
A major drawback of these simpler methods is that (nonlinear) interactions between features can be present. In this case, the problem of how to rate the relevance of a feature is not trivial since the overall performance might not be monotonic in the number of features used. Some feature selection methods try to overcome this problem by optimizing the feature selection for subgroups of fixed sizes (plus-l take-away-r search) or by implementing floating strategies (e.g. floating forward search) [PNK94] instead of only looking at single features. Only few algorithms like e.g. genetic algorithms

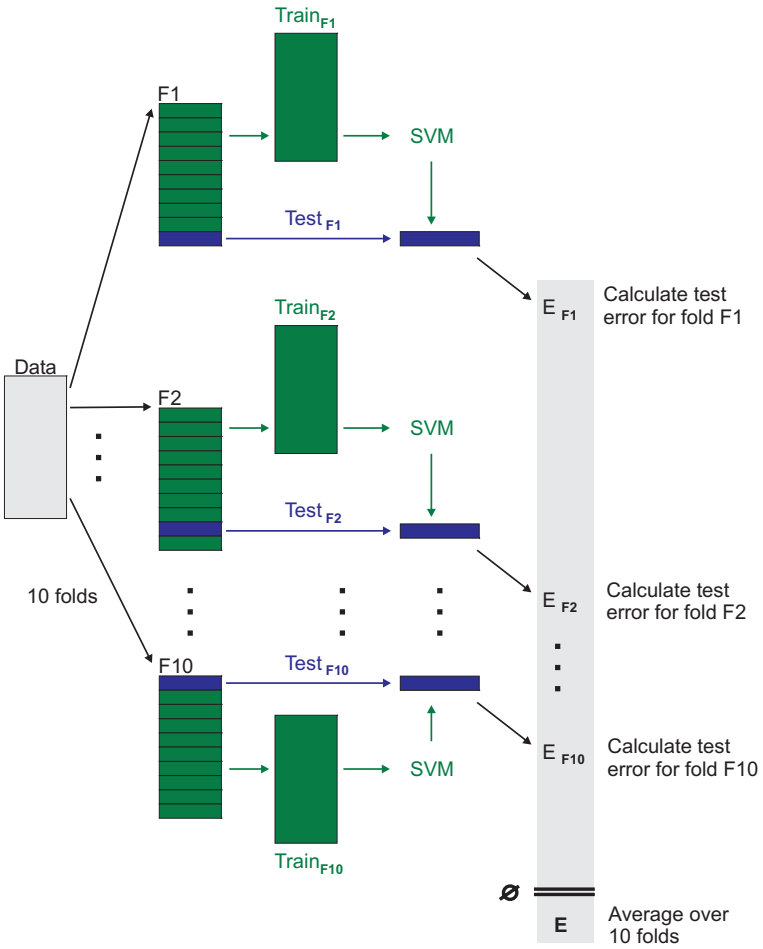


Figure 2.16: *Scheme of a 10-fold cross-validation method to estimate the error of a model (e.g. an SVM classifier) on unseen data of the same distribution. The full data set is split into 10 equally sized folds. Each fold is once used for testing an SVM that has been trained on the remaining 9 folds. The average of the 10 testing errors is reported as the cross-validation error (abbreviated as CV error).*

can choose subgroups of arbitrary size during the feature selection process. They have successfully been used for the selection of spatial features in BCI applications [SBR+03] but are computationally very demanding.

The performance measure that determines the relevance of a feature and decides if the feature is included or excluded from the final feature set can be chosen in various ways and cannot necessarily cope for nonlinear interactions between several features. The most important ones for classification purposes are:

- Linear correlation measures between a feature and the class labels. (Example: correlation coefficient).
- Nonlinear correlation measures between a feature and the class labels (Example: Measures based on entropy like information gain.) Following [YL03] continuous features need to be properly discretized before entropy based measures can be used as they can handle nominal or discrete features only.
- Correlation measures (linear or nonlinear) between a feature and other features which have already been determined as members of the final subset. This measure is based on the analysis of redundancy of features rather than on relevance [YL04].
- The gain or loss in classification performance of the current feature set compared to the enlarged/reduced set.

To judge the quality of the final feature subset chosen by an algorithm, two main measures are common: the classification performance and the size of the subset. Feature sparsity is useful, if the use of many sensors is costly, or if the online processing time is an issue. In the context of BCI (see Section 3.1) for example, the gelling of electrodes is very time-consuming.

2.5.1 Filter Methods

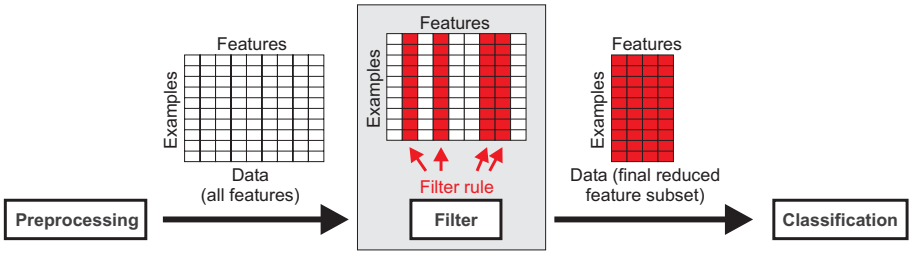


Figure 2.17: *Scheme of the filter method for feature selection. The filter rule is based either on prior knowledge about the classification problem or on statistics derived from the data. The filter is independent of the classifier.*

Filter methods select a subset of features from the data based on some filter rule before the classification algorithm is trained. If available, the filter rule is derived from prior knowledge (e.g. from an expert). If no prior knowledge about the classification problem is at hand, the rule is based on data statistics. Most filter methods require strong assumptions about the class distributions of the data to work efficiently. If the assumptions are true, filter methods are very quick methods, easy to implement and can sometimes even deliver best possible solutions. The risk of overfitting the subset selection to the classification task is rather small compared to other approaches. For this reason, a filter method is often consulted for comparison with new methods. However, if the assumptions are wrong, filter methods might perform poorly as the statistical quantities do not describe the class distribution sufficiently well and have no explanatory power for the classification task.

Filter methods usually rank the available features one-by-one according to a relevance criterion (score). It is then up to the user to define a final feature subset by e.g. choosing the r best ranked features or to determine a threshold of the score above which a feature is chosen.

In the following, two common relevance criteria will be introduced. Assumed that the data set with d features is denoted as $X = \{x_1, \ldots, x_{|X|}\} \subset \mathbb{R}^d$. Define the mean μ and variance V of dimension j , $j = 1, \ldots, d$ as

$$\mu_j(X) = \frac{1}{|X|} \sum_{i=1}^{|X|} x_{ij}$$

$$V_j(X) = \frac{1}{|X|} \sum_{i=1}^{|X|} (x_{ij} - \mu_j(X))^2$$

The subsets $X^+ \subset X$ and $X^- \subset X$ denote all points with positive or negative class labels:

$$X^+ := \{x_i \in X \mid y_i = 1\} \quad \text{and} \quad X^- := \{x_i \in X \mid y_i = -1\}$$

- *Correlation Coefficient*
  (Pearson's) correlation coefficient is a very well-known linear criterion. It determines for a single feature j how strong it is linearly correlated with the labels Y.

$$r_j = \frac{\sum_{i=1}^{|X|} (x_{ij} - \mu_j(X))\,(y_{ij} - \mu(Y))}{\sqrt{V_j(X)}\sqrt{V(Y)}} \tag{2.2}$$

  The numerator is simply the covariance between feature j and the labels, while the denominator only serves for normalization so that r is invariant for scaling of the feature and the labels. Its values are in the range of $[-1, 1]$. Negative values close to -1 stand for a strong negative linear dependency, values close to 0 for linear independence, and values close to 1 for strong linear dependency. The correlation coefficient is frequently combined with a t-test that delivers the probability of the hypothesis of independence between two variables. In this case r^2 is an estimate for the variance between the two variables that can be explained by a linear relationship. Unfortunately correlation coefficients do not follow the normal distribution.

- *Fisher Criterion (FC)* The Fisher Criterion determines how strongly a feature is correlated with the labels. It can be seen as the correlation coefficient corrected for the non-normality. It reduces skew and makes the sampling

distribution more normal as sample size increases [Bis95]. It considers both, the within-class variances and the between-classes distance. The score r_j of feature j is given by:

$$r_j(X) = \frac{(\mu_j(X^+) - \mu_j(X^-))^2}{V_j(X^+) + V_j(X^-)} \tag{2.3}$$

The ranking of all features is easily obtained by sorting the features according to their score.

2.5.2 Wrapper Methods

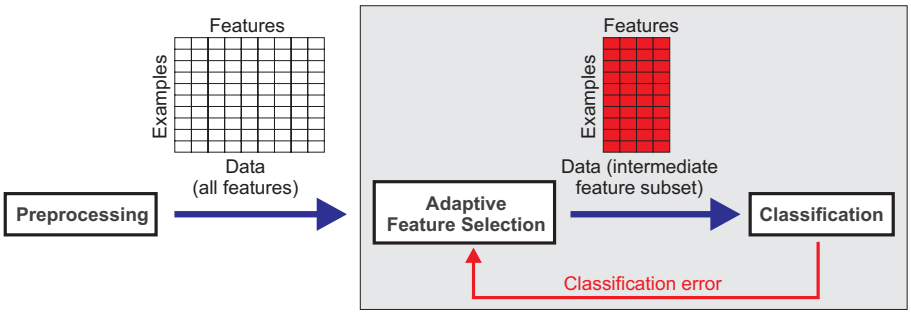


Figure 2.18: *Scheme of the wrapper method for feature selection. This feature selection process iteratively tunes a feature subset to a specific classifier. The change of the feature subset in each iteration depends on the estimated classification error.*

In a wrapper approach, the selected feature subset is dependent on the result of the classification error estimation (e.g. by cross-validation). Subsets leading to small classification errors estimates are preferred over subsets resulting in higher error. Thus the feature selection and the classifier are not independent of each other any more. The error rate of the classifier has to be estimated for every iteration of the search, which can be very costly. Popular search strategies are:

- Sequential forward selection (SFS)
  This strategy iteratively adds one feature to the current subset until the classification error estimate does not improve any further or until other criteria are met. During an iteration, each of the remaining features (one at a time) is added to the current feature subset and the error is estimated. Then the feature yielding the biggest improvement is permanently added to the feature subset and the next iteration starts. The iteration starts with one feature only, which is chosen according to the lowest error. SFS can need a maximum of $\frac{d(d-1)}{2}$ error estimations.

- Sequential backward elimination (SBE)
  This strategy is a reversed FS. Starting with all features, iteratively one feature of the current subset is removed until some stopping criterion is met. As the estimation of training errors is dependent on the number of features used, SBE is usually slower than SFS.

- Forward floating search (SFFS)
  SFFS is a combination of SFS and SBE. Starting like the SFS strategy, a feature subset is built up until no further improvement is possible. Instead of stopping at this (possibly local) optimum, SFFS allows for some backward elimination steps before the search is continued by SFS in another subspace of the problem. As a worst case SFFS can lead to a full search testing $2^d - 1$ subsets.

- Genetic Algorithm (GA)
  Genetic algorithms are based on evolutionary principles, where feature subsets are coded in the form of simple sequences which are considered the genome of the individuals of a population. The population changes by reproduction of its individuals. For reproduction, operators like mutation and crossing over are applied. The fitness of individuals is represented by the classification performance of the corresponding feature subset and determines the chance of reproduction. Over several generations the fitness of the population and its individuals improves. When a stopping criterion is met, the feature subset represented by the fittest individual is selected. GAs are optimization strategies that do not assume a continuously differentiable search space. In a population usually feature subsets of varying numbers of features are present that initially cover the search space randomly. GAs usually need more error estimations than e.g. the SFS strategy.

3 State-of-the-Art BCI Techniques

During the first international meeting on Brain-Computer Interface technology[WBH+01] that took place in June of 1999 at the Rensselaerville Institute near Albany, New York, BCI researchers from 22 research groups agreed upon the following definition:

> **A brain computer interface is a communication system that does not depend on the brain's normal output pathways of peripheral nerves and muscles.**

3.1 BCI Systems in a Nutshell

Following the above definition, a BCI system is supposed to work for completely paralyzed patients that are restricted to the use of brain activity signals and cannot control their muscles. This can be motivated as follows: Every cognitive activity (conscious and unconscious), every thought is reflected by a spatio-temporal electric pattern in the brain. The idea of a BCI system is to exploit those patterns in order to provide a means of communication for patients that do not have enough control over their motor system (muscles) to communicate in a normal way.

A repetition of the cognitive activity does not lead to exactly the same pattern but will vary to a certain degree, so that typically repeated and averaged recordings are necessary to reveal the pattern. The activity pattern which is characteristic for a well-trained and simple cognitive task itself can be quite stable but is nevertheless superimposed by background activity originating from e.g. spontaneous activity of neurons or from unconscious processes.

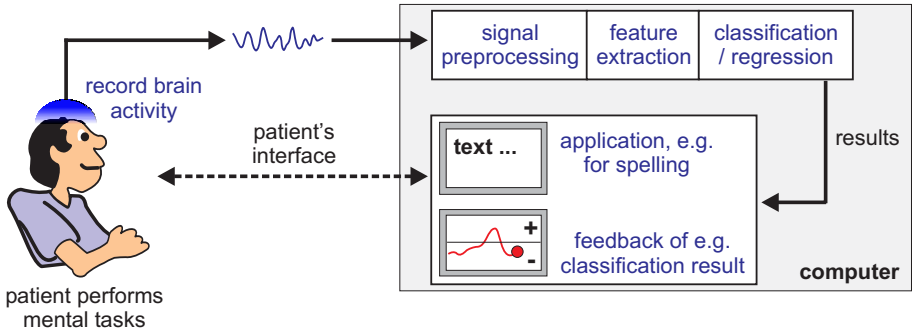


Figure 3.1: *Scheme of a Brain-Computer Interface (BCI) system. While a patient performs cognitive tasks that represent a volitional intention, the brain activity is recorded and processed by a computer system. Algorithms trained to recognize the intension from the recorded signals can provide simple feedback for the patient or even control a BCI application.*

BCI systems comprise a patient or subject, further a recording system that picks up certain signals that represent correlates of the patients brain activity, and a computer that processes the signals and interprets them as the patient's volitional intention. BCI systems usually don't work with the averaged patterns from repeated trials but process and interpret the signals of single trials, which is much more challenging. Additionally a BCI system can offer feedback information about the detected intensions in visual, auditory or tactile form and provide a control interface for an assisting person. The volitional intentions recognized by the BCI system are often converted for specific user interfaces and applications (see section 3.2).

For a BCI system to work, a learning process has to take place that is determined by two coupled components: the computer system must learn how to interpret brain signals (solving a classification or regression task) and the patient must learn how to generate robust and steady signal classes. This coupled system can be very fragile, especially as the daily condition of a patient is subject to changes. Section 3.3 will deal with this issue. The learning process of both the computer system and the patient, is based on several experimental sessions (sometimes also called training sessions) where the two system components can adapt to each other.

A typical BCI experiment can be roughly divided into four steps as depicted in figure 3.2. First, the recording system has to be set up for the user. For most BCI systems this step consists of the montage of an EEG cap or single EEG

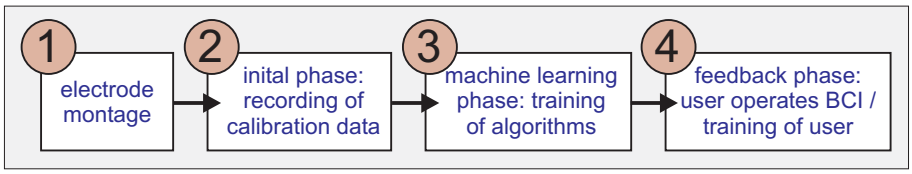


Figure 3.2: *Four phases of a typical BCI experiment.*

electrodes on the scalp of the user. The impedance of the electrodes to the scalp must be fine-tuned in order to gain a good signal-to noise ratio. This first step typically requires between 10 minutes and 40 minutes depending on the number of electrodes.

The second step is often called calibration phase. Usually, it is subdivided into several runs of about 10 minutes duration and breaks in between. While a user is performing mental tasks (e.g. motor imagery), brain signal data is recorded that later on (during step three) is used to set up the signal processing and classifiers correctly. The collected data is referred to synonymously as either *training data* (set) or *calibration data* (set). In this phase neither the user nor the computer learns. Some BCI systems do not use any calibration data at all. If calibration data is collected, this phase takes approximately 30 minutes.

The third step comprises the fine tuning of signal processing methods and the training of machine learning algorithms, classifiers etc. The user can relax during this step, which usually takes less than 15 minutes.

The fourth step is usually called online phase or feedback phase. It comprises the actual application of the BCI system for spelling, control or other purposes (see next section). In this phase, a user learns how to deal with the interface and can adapt to the BCI system. Depending on the motivation and concentration ability of the user, this last phase can take up to several hours. For patients, this phase is often subdivided into several so-called runs of approx. 10 minutes duration each.

The outlined four steps are not found in every BCI system. Some systems do not adapt to the user on a daily basis. They have a fixed and typically very simplified signal processing system so that steps 2 and 3 are omitted. Instead they solely rely on the training of users in step 4 during repeated sessions.

During a so-called adaptive BCI, a variant of the system described above, the algorithms can constantly be adapted to the user's brain signal during the feedback phase. In principal this allows to compensate for non-stationarity in the brain signal but bears the risk of overfitting and oscillations between the two learners (user and computer).

3.2 Application Fields for BCI Systems

Trained BCI systems have the goal to establish a communication channel between a severely handicapped patient and a computer based assistive device. Such devices can be grouped according to the applications that are interfaced by the BCI during the feedback phase. Some examples of applications are:

- *Personal communication*:
  This group of applications comprises the spelling of text, reading text out aloud, composing and sending emails and letters. If a patient is severely paralyzed, the only way of communication without a BCI (if any) is with the help of a caretaker. Introducing BCI-based devices enables the patient to send confidential messages which provides a great amount of independence and self-determination.

- *Environmental control*:
  Opening the front door to visitors, turning lights on or off, controlling shades, regulating the room temperature, changing the backrest position are examples for this group of applications. Their use increases the independence of the patient and decreases the workload of caretakers.

- *Leisure and information*:
  Applications like the control of a reading machine which is able to turn pages provides variation. Other application examples belonging to this group are the remote control of a TV, of a web browser and all web-related applications like chess over the internet.

- *Rehabilitation training*:
  The rehabilitation of stroke patients can be supported by BCI training. As damaged cortical areas cannot accomplish their function any more, paralysis of body parts can be the result. If neighboring cortical areas can be

recruited to take over the missing functionality at least partially, the paralysis will improve. This cortical plasticity can be sped up by BCI training.

- *Control of paralyzed limbs, bowel and bladder*:
  Using a BCI system can lead to the control of a limb orthosis or even allow for direct control of body parts via functional electrical stimulation (e.g. for bowel and bladder control or for paralysis due to different types of palsy).

- *Mobility*:
  Mobility is of inherent interest for a paralyzed patient. A straight-forward application of BCI is the control of electrical wheelchairs. If classifiable signals can be reliably translated into control sequences for the wheelchair, the paralyzed patient regains a certain degree of mobility.

Other BCI approaches want to provide extra information channels in addition to the muscular activity of healthy persons, e.g. for computer games or warplane control. As it is questionable if these BCI systems follow the definition given in section 3, they are not in the scope of this work.

Furthermore, BCI systems have to be differentiated from various forms of the more general (bio-)feedback systems. Like BCIs, they are used to provide conscious access to (unconscious) body functions. The recorded physiological quantities can be taken from various locations (e.g. blood pressure or muscle tone) but do not necessarily have to be correlates of voluntary neural activity.

3.3 Working with Patients

BCI systems are developed specifically for severely handicapped people. There are several reasons why this task is much more difficult than the development of a system for healthy subjects.

First of all, the brain signals derived from patients cannot be expected to be normal and the knowledge of typical brain signal responses and frequency characteristics from neurophysiology literature might be of limited use only. The reasons for this abnormality can be (a combination of) strong medication, long term neural changes caused by degenerative diseases like amyotrophic lateral sclerosis (ALS, see below) or damages caused by stroke or cerebral palsy. Due to strong inter-individual variation a BCI system has to be tailored to the specific brain signal characteristics, the needs and the abilities of every patient.

Second, technical problems can arise already during the signal recording. When patients are artificially ventilated or other electrical medical devices are in constant use, artifacts caused by these devices frequently appear. Although paralyzed patients might not have voluntary motor control, they possibly show spastic movements that can lead to additional signal artifacts in EEG and MEG.

Third, the initial phase of recording calibration data as well as the user training phase are very difficult. Both phases are preceded by the application of a recording technique, e.g. the montage of an EEG cap, and both phases require the patient's concentrated cooperation, sometimes for twenty or more sessions, each having one or more hours of duration. It is obvious that the reduction of the montage time and the user training time is an important goal for the improvement of a BCI. Patients are usually immobile and the experiments and training sessions have to be performed at the patient's home. In addition, they are often under medication, might have limited attentiveness, can suffer from depressions etc. For these reasons the recording sessions often have to be restricted to short time intervals.

Current BCI systems that work with patients try to get them into BCI training programs at an early stage, before the patients are completely locked in. It is believed that early training in the LIS state is one key issue for successful communication over BCI once the patient is in the CLIS state.

Amyotrophic Lateral Sclerosis

Patients suffering from Amyotrophic Lateral Sclerosis (ALS) are the main target group of the Tübingen BCI project. ALS is a progressive and fatal multisystem disorder affecting predominantly the motor system and is one of the main diseases leading to the (completely) locked-in syndrome.

The overview paper of Sperfeld et al. [SKL04] characterizes ALS as a degeneration of the pyramidal cells of the motor cortex and of the α-motorneurons of the brain stem and the anterior horn of the spinal cord. In addition, the autonomous nervous system can be mildly affected. In 2-5 percent of the patients, frontal dementia is reported. Worldwide the prevalence of ALS is reported as high as 1.2 to 2.6 per 10^5 individuals with slightly increased rates in Europe. So far (2005) there is no clinical trial known that satisfactorily slows down the progression of ALS or even assures a patient's survival. The progressive disease predominantly affects men, the average duration is between three and five years [SKL04].

Recent studies revealed that the quality of life of ALS patients is strongly dependent on their degree of self-determination and that maintaining communication ability is an important key to self-determination for ALS patients [NWB04].

3.4 BCI Paradigms

The various historic and present BCI systems differ in many aspects. This section introduces the most relevant of them as a preparation for the analysis of existing BCI systems in section 3.6.

3.4.1 Subjects, Patients and their (Dis)Abilities

BCI experiments are often carried out with healthy subjects in order to refine experimental methods. The goal of most BCI systems, though, is to transfer the developed methods to patients. The degree of disability of patients varies in a large spectrum from slightly reduced motor abilities (e.g. after a stroke or in the first stages of ALS) to the completely locked-in state (CLIS) that paralyzed ALS patients can reach in the last stage of the disease. As pointed out in section 3.3 working with severely handicapped persons slows down experiments and enlarges the training times necessary to validate a new BCI paradigm. Working with healthy subjects on the other extreme allows quicker experimental and methodical advances but produces results that are possibly not or only partially transferable to patients.

3.4.2 Learning in a BCI System

The overall communication rate a BCI System can provide depends on the interaction of two coupled aspects: the ability of the subject to generate distinguishable signals and the ability of a computer system to classify these signals correctly. On one side of the spectrum, the patient has to learn (e.g. by conditioning) the control of certain (simple) aspects of his brain signals while the computer system has no need to learn or adapt. This pure patient learning is very time-consuming and can easily take the patient weeks of intensive training. As pure patient learning has low requirements for the computing power of the BCI system - usually such a system can be reduced to the recording and displaying of simple signals - it has been used since the beginning of BCI research. A typical example is the slow cortical potentials training [BGH+99].
On the other end of the spectrum is the attempt to capture the BCI problem purely with learning algorithms. This paradigm shifts the responsibility from the user to the machine. Instead of training a user to generate certain distinguishable signals, the computer is trained to distinguish rather natural mental states that are expected to be reflected by at least slightly differing classes of brain signals.
Most of today's existing BCI systems use mixed strategies. Such strategies comprise initial training of (only) the algorithms, then a training phase for the patient with periodic re-training of the algorithms. So far no profound results are available about the difficult interaction of the two components, especially not for longer time intervals.

3.4.3 Experimental Tasks and Exploited Signals

The experimental settings for BCI systems depend strongly on the kind of expected signals that shall be exploited. For systems based on EEG, ECoG or MEG, the following typical signal types can be used for communication [GHEV04] (together with possible experimental tasks):

- Slow Cortical Potentials (SCPs, endogenous)
  SCPs are DC shifts (positivation or negativation) that last from one to several seconds. They can be observed e.g. during relaxation movement or in preparation state. In the latter case they are also called *Lateralized Readiness Potential*, *LRP* or *Bereitschaftspotential*. Subjects can learn to control their SCPs by reinforcement training. Birbaumer et al. proposed visual or auditory feedback for SCP learning [BGH+99].

- Oscillatory signals / Event-Related Desynchronization (ERD, endogenous)
  Even without external stimulation, the cortical networks establish oscillations of the brain signals. Typical rhythms are shown in table 3.1. They are not identical for every subject but can usually be recognized by visual inspection of repeated trials. A well-known ERD is the so-called μ-rhythm, which is located in the alpha band at approximately 8-12 Hz, often accompanied by an ERD in 18-26 Hz. These rhythms are strongest over the motor cortex. If motor areas are involved in the preparation, the execution or the imagination of motor actions, the amplitude of the μ-rhythm decreases. This mechanism is exploited for BCI systems based on various motor imagery experiments like imagined hand, foot, or tongue movement.

- Event Related Potentials (ERPs, exogenous)
  ERPs are short but relatively stable transient patterns that appear involuntarily after external stimulation. Subjects can be trained to influence the amplitude of the ERPs for communication.

Band	Approx. Freq. Range [Hz]	Comment
δ	0.1-4	Increased during deep sleep
θ	4-8	Increased during drowsiness, light sleep and hypnagogic states
α	8-12	Increased during state of consciousness and relaxation (contains μ sub band)
β	12-30	Decreased during motor activities
γ	>30	Increased during higher-level cognitive activity (temporal binding of networks?)

Table 3.1: *Typical EEG rhythms and associated frequency bands.*

- Steady State Evoked Responses (SSER, exogenous)
  SSER are oscillatory patterns of the EEG that are driven by external stimuli of the same frequency. Blinking screens or beeping tones are used to produce visually evoked SSER (SSVER) and auditorily evoked SSERs. For BCI experiments, subjects can concentrate on one of several provided stimuli. The oscillatory pattern of the stimulus that is in the focus of attention can best be observed in the EEG of the subject [DSW00].

For healthy subjects most of the above signals can be exploited for more or less good BCI control, as has been shown in numerous BCI publications. Handicapped patients, however, might not have the cortical prerequisites (e.g. after stroke or during ALS) to fulfill standard paradigms. In these cases new experimental tasks (e.g. new imagery tasks) have to be tested individually with the patient to find a suitable setting. As this involves finding classifiable features in the signals the task of feature selection becomes most important when working with impaired BCI users.

3.4.4 Binary vs. Multi-class

Most BCI systems are based on EEG and rely on the classification of two brain states. As the duration of one trial cannot be shortened arbitrarily (most systems need at least 2 seconds of recording time per trial) the overall communication rate of a BCI can be increased either by better classification rates or by expanding the number of classes. Unfortunately, these two dimensions are not independent: In EEG setups for example, two-class settings can reach classification performance of more than 80 percent, but classification performance decreases to much lower rates when a third or more classes are added. This is usually not a problem of the classification method (most can deal with multi-class problems quite well), but a problem of signal separability: the more classes are used, the more overlap is found between the class distributions in feature space. First studies ([KVP05]) have shown that for EEG recordings and existing imagery tasks the best information transfer rates can be obtained with BCI systems with a maximum of 3-4 classes, depending on the performance of the subject and the chosen imagery tasks. For some subjects, the binary class setting was still optimal ([KVP05]). Theoretical analysis of the information transfer rate has shown that the most important factor is still a high classification performance and that in many cases a two-class problem has to be preferred if otherwise the classification performance decreases.

3.4.5 Trial Mode

A synchronous BCI sets fixed time windows (indicated e.g. by sounds or visual markers) during which the user has to produce the mental imagery (see the experiments in section 5). In contrast, an asynchronous setting tries automatically (based on constant signal analysis) to determine the time window during which a mental imagery was performed. For driving e.g. a two-class system (with two mental imagery tasks), a BCI in asynchronous mode has to detect three different brain states: relaxation and the two actual imagery states. A synchronous BCI only needs to distinguish the two imagery states as the relaxation periods are fixed by the trial scheme. For both modes user control becomes effective after the trial / time window.

3.4.6 Feedback Mode

The result of the classification can be provided for the subject by the combination of visual, acoustic and haptic feedback. The choice should be based on the remaining abilities of the patient and made so that the feedback can be supported by the final application but does not interfere with it. Feedback can be given trial-based after the end of each trial (a mode often combined with synchronous settings) or continuously throughout the trial.

3.4.7 Recording Techniques, Feature Domains and Signal Processing

Recording techniques provide different resolution in time and space as pointed out in section 2.2. The right choice depends on whether the technique should immediately be applied for patients (EEG, ECoG) or whether high signal quality should be obtained for pre-patient experiments (MEG), on the costs of the method etc. BCI systems can be classified according to the used recording technique, according to the features extracted from the raw measurements and according to the signal processing and classification algorithms applied at later stages. All these aspects have been introduced in sections 2.2, 2.3 and 2.4.

3.5 Technical Challenges in BCI

The task of setting up a BCI system leads to a number of technical challenges which are not dependent on the chosen approach, the chosen experimental setting, the mental task etc. but which are of a more general nature. One group of these difficulties emerges from the training of handicapped patients as addressed above in section 3.3. They have been addressed by the research topics of medical psychology groups and will not be discussed here. Another set of difficulties emerges from the signal processing part of BCI systems:

- On the one hand the recorded data is usually *high-dimensional*, but on the other hand only a *few trials* are gained during an experimental session. Following the notion of the *VC dimensionality*, the *shattering* of data points and the *capacity* of a classifier ([SS02, Vap95]) this situation is difficult for the application of most classification algorithms, as it hampers the generalization ability.
- The signal processing algorithms have to be able (at least initially) to deal with *huge data amounts* - even a simple experimental session easily produces several hundred megabytes of raw data.
- Signal processing algorithms have to be chosen so that they allow for *fast online treatment* of data during the BCI experiments and during the final use of the system. In return for this, the initial tuning of data preprocessing methods and the classifier training are not as time-critical. This period of optimizing the system (after the initial data has been recorded) may take a few minutes.
- The recorded data is possibly *non-stationary*. The state of the patient and at the same time the characteristics of the recorded brain signal can change from session to session or even within a session. For BCI systems to cope with this non-stationarity the choice of stable features is very important, or the algorithms must be designed to continuously adapt to the patient. The second approach is very difficult and can lead to oscillating behavior on the part of the BCI system.
- To make communication feasible and to provide a motivating system for the patients, high *classification performance* is necessary. Misclassification, although always possible for reasons of non-stationarity, has to be reduced. For the example of a language support programme, Perelmouter et al. [PB00] reported a critical level at approximately 30 percent error. For every-day work with a patient, 20 percent of error or less are necessary to keep the communication process feasible. Furthermore, correcting steps that have to be performed due to misclassifications discourage the patient and lead to a loss of confidence into the system or into one's abilities.
- Every subject and *every patient is unique* with respect to his or her abilities and (except for very simple cognitive tasks), to varying functional organization of the cortex during ontogenesis and due to learning. Therefore generalizations about the best features are typically suboptimal, even for well-known paradigms.
- The identical *repositioning* of the recording apparatus is not easy for EEG and MEG. Due to a lack of experiments with humans, it is not known for ECoG how much variation in the grid position must be expected over a longer period of time.
- Due to limited time and *limited attention* of a patient, the initial screening process and the optimization of the classifier should be quick (within a few minutes) and reliable. These requirements show that a stable work flow (with as little interaction as possible) is desirable for the screening process.
- For some imagination tasks, the applied recording technique might not be suitable to pick up characteristic (classifiable) signals.
- Medical staff and psychologists are very interested in understanding the meaning of classifiable patterns. Thus, a desired property of a BCI is the possibility of a simple *interpretation* and visualization of characteristics of the

trained signal processing and classification algorithms. This becomes more difficult with larger dimensionality of the data.

- In order to *identify additional classes* that can be used for the patient's communication, *new experimental paradigms* (e.g. new imagination tasks) have to be tested. For very simple paradigms the neuropsychology literature can recommend features and locations. But for most new paradigms it is usually not known which signal features lead to good classification results. This disadvantage prolongs necessary screening processes.

Most of these difficulties are rather complementary and difficult to solve directly. The best non-invasive recording technique at present, for example, seems to be MEG, but this technique is not suitable for LIS/CLIS patients due to financial and practical reasons. Another example is that the problem of high dimensionality is best compensated by recording more trials during an experiment which unfortunately collides with the limited concentration ability of patients and the amount of data to handle. The need for better classification rates, on the contrary, implicates the need for bigger training sets. If the recording procedure is initially restricted to few positions suggested by literature, this would save screening and processing time. The drawback of this reduction would be that the full classification performance possibly cannot be gained. This was shown by [LSH+04] for the example of EEG recordings for a simple paradigm that was considered well-known. It can be expected that a reduction of recording sites *a priori* leads to even stronger performance decreases if less well-known paradigms are tested.

3.6 Analysis of Existing BCI Systems

After the general analysis of BCI settings in the previous section, the BCI systems of four BCI groups will now be analyzed.
Three of the groups were among the first to address aspects of BCI research [WMNF91, RBEL84, RWP97, PPF94]. Their BCI systems have been developed over a relatively long time of more than 10 years. Most importantly, the experience of applying BCI systems to severely handicapped patients has entered into these three systems. The fourth group (Berlin BCI Group, [BCM02]) has not been dealing with BCI systems for a comparably long time but has been chosen as it represents a novel approach to BCI research - the intense use of machine learning techniques. All of the four groups have made important contributions to the BCI research during the last years and have therefore been chosen to represent the state of the art in BCI research.

As the research activities of the four groups cannot be completely covered in this thesis, the following structure will be applied: after characterizing the main focus of the group and the most important BCI applications, a small number of outstanding experiments or publications of the group will be presented in a condensed form. Thereby special emphasis is put on aspects of feature selection.

3.6.1 Graz BCI

The Graz BCI has gained much attention due to the results of the research group in rehabilitation training and the control of neuroprosthetic devices. In addition, the group reported on a variety of experiments with healthy subjects. Most experiments are based on EEG recordings except for a small number of ECoG recordings of epileptic patients.

In [MPSPR05] the group presents the results of applying the Graz BCI system to control a neuroprosthesis (freehand system) for a young patient suffering from a traumatic spinal cord injury. Shoulder movements and positioning of the hands remained under the patient's control, but not hand movements. The Graz BCI system was connected to the freehand system at the right hand of the patient. Within three days of training it could control a simple hand movement (lateral grasp) in a synchronized setting. The BCI system used imagined movements of the feet and the left hand. Two bipolar electrodes were used, whose position was chosen by hand close to the EEG channels Cz and C4. An algorithmic channel selection was not performed. For classification, a linear discriminant analysis (LDA) was trained on features representing the band power of two frequency bands (12-14 Hz and 18-22 Hz). The frequencies were chosen after visual inspection and single-feature linear correlation analysis. This way of choosing features corresponds to a filter approach that relies on prior knowledge about the task.

In publication [LSK+05] the authors describe view direction control in a virtual environment. Four healthy subjects performed the experiment. One trial lasted 8 seconds. Again the experimenters chose two bipolar channels (around C3 and C4) by hand for the experiment. The Fisher Linear Discriminant Analysis (FLDA) was trained on the band power of two bands (10-12 Hz and 16-24 Hz). The publication [PNM+03] gives details about the used features: the optimal time window for classification was chosen by cross-validation among ten possible time windows.

Publication [PNM+03] is a rather broad overview over the state of the art of the Graz BCI system. It deals with four different clinical applications (virtual keyboard, control of a hand orthosis, BCI training via telesupport, and a basket

paradigm). Again the described approaches use linear classification and either one or two bipolar electrodes around CZ. The overview does not state any algorithmic solution to determine optimal channels but again shows that this selection does rely on prior knowledge. As frequency features, the paper reports band power features (chosen *a priori* or by single correlation analysis) derived by FFT or AR models.

Feature selection by genetic algorithms has been reported by two publications of the Graz group for EEG and for frequency selection in single channel ECoG [GHLP04] data. Genetic algorithms in general are useful to solve the feature selection problem in wrapper approaches, but show a very bad performance so that they are currently only feasible for the offline analysis of BCI data.

It can be summarized that the Graz BCI currently either does not perform feature selection during online experiments or only a filter approach that is based on prior knowledge or simple correlation analysis of single features. An advantage of this approach is that the experimenter does not have to spend much time after an initial calibration data recording before he can proceed to the feedback phase. This is of course advantageous under time pressure. A clear disadvantage of these methods is that the features are not properly adapted to the individual subjects - which might lead to a loss of classification performance. The authors of [PNM$^+$03], [LSK$^+$05] themselves speculate that an improved performance could have been possible if the electrode positions and the used frequency bands had been optimized. It is not stated why this has not been done.

3.6.2 Wadsworth BCI

The Wadsworth BCI system is designed for severely motor-impaired patients. Besides this focus, the group performs studies with healthy subjects. The signal processing of the Wadsworth BCI focuses primarily on EEG rhythms recorded over the sensorimotor cortex [WMVS03], but latest publications also use ECoG signals in a study with epileptic patients[LSW$^+$04].

Typically, motor imagery tasks like hand, foot or tongue movement imagery are used. When using EEG signals, the group records 64 channels for offline artifact analysis but only uses two channels (C3 and C4, chosen by prior knowledge) for calculating motor related features (e.g. intensity of 8-12 Hz and 12-24 Hz band, chosen by prior knowledge) [WMVS03]. Classifiable channels or frequency components are always determined based on the correlation of single features with imagery classes (r^2-values). For classification, these features are linearly weighted. Most publications report fixed weights, not subject specific learned weights.

Recently, healthy and motor-impaired patients were trained with EEG over tens of sessions in a long term study. In their publication [WM04] the group reported that an adaptive strategy was applied. Therefore the features were re-weighting after each session. Over several weeks of training this has led to two-dimensional control by motor imagery. The authors showed that the re-adaptation of the BCI system to the daily changing signals of the subjects was necessary for this study.

In recent research, the group reports ALS patients that use BCI communication systems based on EEG signals in everyday life (personal communication, publication in preparation).

Experiments with ECoG [LSW$^+$04] showed that one-dimensional online control can be achieved by very short user training in the range of 1-2 hours. Therefore up to four frequency bands from one or two ECoG channels were chosen as features. The authors recognized that task related, classifiable information was contained in the gamma band above 40 Hz by correlating all possible bands (bandwidth 2 Hz) to the class labels.

The Wadsworth BCI group recognizes that further improvements of classification rates and the expansion of one-dimensional control to two- or multi-dimensional control might be possible only if the adaptive algorithms are expanded by additional EEG features [WM04]. As examples, the authors mention additional recording positions, additional frequency bands and time domain EEG features. This expansion will increase the need for fast individual feature selection in online experiments that take the interaction of several features into account.

3.6.3 Berlin BCI

The research of the Berlin BCI group differs from that of the other analyzed groups as their primary goal is to develop a BCI for healthy users. Nevertheless their methods could generally be used for patient BCIs as well. The Berlin group rather develops a BCI system based on paradigms of motor imagery or motor execution for healthy users. Emphasis is put on reaction-time-critical applications like computer games [KBCM07], [KDB$^+$04], high-speed spelling [BDK$^+$06] or mental state monitoring [MTD$^+$07, KDB$^+$07]. This system was included in the analysis of existing BCIs as it stands for the intensive use of machine learning algorithms. In the context of patient BCI this means that the training effort is shifted from the user to the machine as far as possible [BCM02, MSK$^+$06, BDK$^+$07].

The publication [BCM02] presents earlier work of the group where the BCI was used to detect upcoming executed finger movements of healthy subjects in a self-paced setting. Blankertz et al. report high classification rates (>96

percent) for a 2-class setting and 27 EEG electrodes. Features used were capable of capturing the Bereitschaftspotential or Lateralized Readiness Potential (LRP), a relatively slow negative EEG shift over contralateral motor areas. This choice of features reflects the use of prior knowledge. For classification, the authors applied several methods, but found that Fisher linear discriminant analysis (FDA) and linear SVM were able to classify upcoming movements with the aforementioned high accuracy. By regularization with l_1-norm a sparsified variant of FDA used only 68 out of 405 available features which can be interpreted as an implicit form of feature selection. For this reason, the FDA has been included in a comparison with the method developed in this thesis (see section 6.1).

Later work of the group ([KBCM07], [KDB+04]) reports results from offline and online experiments with healthy subjects. In an EEG setting with up to 128 channels the group used either LRP features for a fast reaction setting (finger tapping) or oscillatory features (event-related potentials, ERP) in a motor imagery setting. ERP features were captured in the well-known μ- or β-bands by spatial filters like common spatial patterns (CSP) [DBC03] or variations thereof [LBCM05, KDB+04, DBK+06b, TDN+06, KSBM07]. The latter is a label-dependent method that works on the covariance matrices of the imagery classes. It finds projections that maximize the variance for one class while minimizing it for the other class. The optimal CSP components for classification with an l_2-norm regularized FDA ([KBCM07]) or LDA [DBCM04] show high classification performance but have to be determined by hand after visual inspection. For the correct estimation of CSP filters, which discriminate the imagery classes, a relatively dense recording of the scalp surface is necessary and only a limited number of sensors can be omitted. This would be a problem for the every-day application for patients.

Another recent publication reports an extended approach, where the slow LRP features are combined with autoregressive features (AR) and/or ERP/CSP [DBCM04]. It has been shown that those feature classes contain different information about the movement task and that a combination of features helps to achieve better classification rates. Feature selection was not performed in this work, although the total number of features was increased by these combination approaches.

The Berlin BCI does not use the advantages of explicit feature selection, except for the intrinsic form realized by l_1-regularization in FDA or LPM as in [BCM02], or feature selection by hand according to domain knowledge.

3.6.4 Tübingen BCI Group

In contrast to most other BCI systems, the Tübingen BCI is centered around (mainly ALS) patients [HLS+06, HNK+07]. It has a comparably long tradition. The group concentrated very early on conditioning slow cortical potentials (SCP) for use in BCI systems[RBEL84] which until now is in the focus of research [BGH+99], [KNK+01], [BHKN03], [KNW+04]. Movement imagery tasks and typical event related (de-)synchronization (ERS/ERD) are used in the later publications of the group [KNM+05]. The work addressed in this thesis contributed to that newer kind of BCI approach. As the results of this newer approach are dealt with in section 6, this analysis of the Tübingen BCI system is restricted to earlier work based on SCP paradigms and on ERD/ERS paradigms published in [KNM+05].

In older as well as in recent publications, the group trained disabled patients based on operantly conditioned changes of slow cortical potentials [BGH+99], [KNK+01], [KNW+04]. For class decisions, a simple threshold decider was used. The SCP feature was derived from very few EEG channels around Cz chosen by prior knowledge. Algorithmic feature selection of any kind was not performed. No learning algorithms were applied for the classification task. Due to the huge amount of user training sessions of up to several hundred, the group could establish slow communication via a BCI spelling device for most of the patients that suffered from ALS for several years but were not yet in a completely locked-in state.

A recent publication reports on the long-term training (20 sessions) of four severely disabled ALS patients [KNM+05]. The EEG recordings were performed using 16 EEG channels. For control of a vertical cursor the band power of μ-rhythms in the α- and β-band were fed into a linear discriminator. The used frequency bands were fixed, and chosen on a subject-specific basis according to prior knowledge (for the initial training of the patient) or according to the r^2 values from linear correlation analysis (for later sessions).

The traditional version of the Tübingen BCI uses no feature selection, or at most very simple correlation measures for choosing frequency bands. EEG channel selection is not performed automatically.

Recent publications also evaluate the use of spatial filter methods like CSP [HLT+07] that take advantage of a spatially densely sampled scalp and propose new variants which explore the trade-off between spatial sampling density and performance [FHS06].

Although showing very accurate classification results, these methods have not yet been used for the everyday application with patients.

3.7 Discussion

All of the four groups and most of their publications report on well-known motor imagery or executed motor paradigms with EEG. A small number of publications of the Graz group reports on ECoG experiments in addition. Most experiments of this group are accomplished with only very few (two bipolar) EEG channels. Contrary to this, the Berlin BCI group relies on a large number of recording channels for their methods which is contradictory to the demands of daily patient work. In order to determine the features of channels (e.g. components of CSP patterns or band power) most publications rely on *a priori* knowledge and mention hand-optimization steps. This knowledge is centered around the choice of the EEG channels C3 and C4 and the typical motor related frequency bands. In several publications of the Tübingen BCI, the features are even completely predetermined. In most other publications, a small subset of features is chosen by correlation values of single features to the class labels (r^2-values). Features that are only effective in combination with other features are usually not considered.
A few exceptions are observable for offline data analysis publications, but not for online feedback experiments. The Graz BCI group for example has applied genetic algorithms to determine a feature subset from one single ECoG channel. The choice of channel itself was not part of the selection process. Another positive example is the use of l_1-regularization in FDA by the Berlin BCI group that has implemented an implicit feature selection process [BCM02]- apart from this, the group reports a hand-optimized selection process for CSP features.

It can be concluded that feature selection is disregarded in the examined BCI systems and that its potential advantages (introduced in Section 4.1) are not used. This is astonishing, as several publications indicate that the combination of more non-standard features might boost the BCI performance in future research, especially in combination with the ECoG recording technique that provides an increased frequency bandwidth to choose features from.
Furthermore, the need of reduction of the number of EEG channels for practical reasons in patient work is insufficiently addressed by the existing BCI systems - instead of optimizing the choice of channels for individual subjects and paradigms, either standard EEG positions are chosen, or this need is not met at all.

4 Exploiting Individual Feature Selection for Fully Automated BCI Training

What is it that makes feature selection for BCI an interesting field of research? The answer is given by a number of difficulties that arise in BCI and that are all related to the feature selection problem. These difficulties are introduced in a number of problem statements of Section 4.1. They lead to the insight that a generally applicable solution of the feature selection problem is needed. In Section 4.2 this thesis proposes such a solution: an (algorithmic and automated) individual feature selection (IFS) , as the core of a fully automated BCI training concept. The characteristics of the new IFS during the machine learning and the feedback phase are described in Sections 4.2.1 and 4.2.2. The core of the IFS is an embedded method that is introduced in Section 4.2.3. Section 4.3 introduces the primary performance goals of feature selection by an example of BCI channel selection and discusses additional possible quality metrics.

4.1 Problem Statements

Based on the more general difficulties that the field of BCI bears (cf. Section 3.5), those that are related to feature selection are highlighted in further detail. They show that many reasons exist to invest time and effort in individual feature selection before the final classifier is trained for the feedback phase of an experiment. The following reasons are not orthogonal and several of them do address the same underlying point. Nevertheless they are listed from a rather practical point of view:

- *Many features*
  Typical BCI classification problems comprise numerous input features. In practical applications, even for a simple task hundred or more features are a quite realistic setting as the number of recorded channels multiplies with the number of features extracted per channel (see Section 2.3) and with the number of time intervals within a trial. Newer signal recording techniques like ECoG or invasive ones tend to provide even more recording channels and capture signals of a broader spectrum. This often leads to thousands or tens of thousands of initial feature dimensions the BCI system can potentially profit from, in contrast to the small number of data points.

- *Irrelevant and misleading features*
  Many features do not contain relevant information for the classification problem. Some of these input features that contain noise might pretend to contain relevant information although they only correlate with the task labels of the training set by chance. Such features will further be called artifacts. A classifier trained on these artifacts might overfit to these false regularities and fail on new recordings.

- *Burden on the classifier: capacity and efficiency*
  During its training, a classifier usually has to invest in parameters (e.g. weights) to handle the input dimensions. This not only slows down the training- and classification process but also enlarges the complexity and capacity needed to reach a good separation on the training data. If noisy features can be removed, the capacity is not unnecessarily increased. This, to some extent, prevents a classifier from overfitting to the training data.

- *Too few training vectors*
  BCI experiments are usually performed under difficult conditions (see Section 3 in general and Section 3.5 for details). One of them is that patients are not able to endure very long recording sessions and even healthy subjects cannot maintain concentration for too long a period of time. These circumstances requires that a BCI system and especially its classification algorithm deal with a small number of example training vectors (typically only a few hundred trials) before the system can be used to classify unseen data.

- *Individual differences between patients and recording sessions*
  Even if the imagined task and the recording technique are kept constant, good features cannot usually be transfered easily from one person to another due to various reasons: recording positions cannot be controlled in enough detail, local brain structures are not organized exactly the same way, some users' lack of typical rhythmical activity (e.g. μ-rhythm) and imagined tasks might be performed slightly differently. This means that an individually optimal set of features has to be determined for each subject and (on a limited scale) also for each experimental session of the same subject.

- *Benefit: Interpretability*
  For some BCI applications, the interpretability of a trained classifier is as important as good classification results. Especially in EEG and MEG, the movement of upper skeleton muscles leads to high electrical activity that is picked up by the sensors. In experiments where healthy subjects have to learn a task and are provided with feedback, the use of a specific muscle movement can accidentally be conditioned to a task. To detect such possibly involuntary cheating, feature selection can be very helpful. Ideally in this case the feature selection process ranks peripheral recording positions close to neck or jaw muscles as very important and on the other hand excludes channels that should be relevant from a physiological point of view. Some classification algorithms of the ART family [Gro87] or decision trees [BFOS84] directly allow the deduction of a set of rules that describe the classification in a form that is readable for humans. In this case fewer features are favorable.

- *No prior knowledge for new paradigms and recording methods*
  For new experimental paradigms, prior knowledge about the importance of features is often missing and standard textbook features of the healthy EEG might even be misleading. This is the case for new imagery tasks (if for example the number of classes has to be increased), or for tasks that have been individually tailored to the limited abilities of a patient (compare Section 3.3). A similar lack of prior knowledge is encountered when BCI research is extended to ECoG or other implanted recording techniques.

- *Individual brains and variance in recordings*
  Due to individual differences, brain structure and brain signal characteristics do vary between subjects. For example, the hand cortex position, the skull shape or frequency bands that are related to movement imagery can be subject to slight variation. Additional variance is introduced by EEG due to the recording equipment: In practice, the exact positioning of EEG recording caps is difficult and the re-application at consecutive recording sessions does not result in identical recording positions. Thus, to get the best out of the BCI recordings, it is in practice useful to repeat the search for the optimal channel subset.

- *Time-consuming application of EEG*
  An EEG recording starts with the montage of an EEG cap. For a cap of 128 electrodes this takes approx. 30 minutes. It is of vital interest to reduce this number of electrodes (and therewith the number of features) to e.g. less than ten electrodes which can be applied more quickly, especially if a BCI system must be used by a patient on a daily basis. Therefore a generally good subset of electrodes should be chosen specifically for that patient and for the BCI task used. Of course the goal of a reduced but fixed electrode set is in conflict with the optimal daily performance[1]. Nevertheless patient training without simplified settings may not be feasible at all (see section 3.3).

For all of the above reasons it is a good idea to reduce the number of features. Furthermore, the analysis of existing BCI systems in section 3.7 has shown that previous BCI approaches have not provided algorithmic feature selection solutions to deal with the above points. Up to now, these systems strongly rely on very few experimental paradigms which in principle restricts the information transfer rate, and on expert knowledge about useful features that can be expected for these paradigms. In order to overcome this limitation, this thesis proposes a new, machine learning based concept that takes automated feature selection (individually for every user) in the focus of the signal processing flow.

But it should not be concealed that the feature selection process itself bears the risk of overfitting. Jensen and Cohen [JC00] argue that over-searching the training data for optimal feature subsets is a problem, especially if many features and simultaneously only a small training set is present. In this case the feature selection process can come up with a subset of noisy features that is not truly relevant for the classification of mental tasks but is only correlated to the labels by coincidence. This thesis avoids that pitfall by careful validation of both the classification results and the feature selection results (see Section 6.1).

4.2 Signal Processing Concept

This thesis proposes a new concept of signal processing in BCI systems. The scheme of this new concept is depicted in Figure 4.1. A first phase for recording of calibration data is omitted for the sake of clarity as it does not differ from any other BCI systems that use some form of algorithmic adaptation to the user. Similarly to other concepts, a BCI session can be subdivided into a machine learning phase (see Section 4.2.1) and a feedback phase (see Section 4.2.2) that are depicted in the upper and lower block in Figure 4.1. Different from existing approaches, the proposed new design includes a feature selection step. It is performed during the machine learning phase. The IFS interacts in embedded form

[1] In addition to the problem of reproducing electrode positions accurately, they might change slightly from day to day due to changing states of the patient.

(see Section 4.2.3) with the classification step, which is executed as a linear SVM. As channel selection has already been identified as a hierarchical and thus more complicated case of IFS in BCI, it will be used in the following sections of this thesis whenever a concrete example of feature selection is needed.

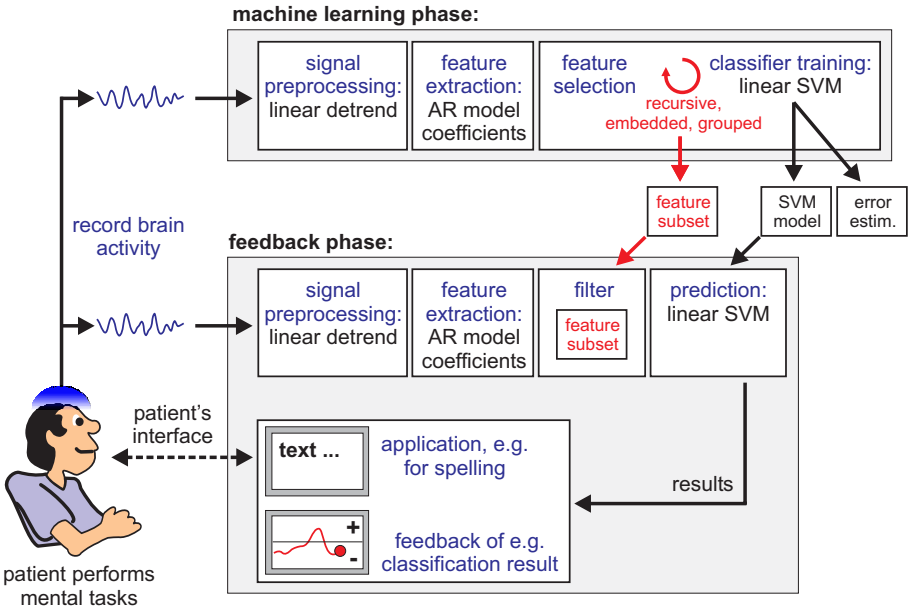


Figure 4.1: *Scheme of the new automated BCI concept. During a machine learning phase, the individual feature selection (IFS) algorithm iteratively reveals a feature subset that allows good classification. The discovered feature subset can then be used during the feedback phase.*

Chapter 5 introduces a number of BCI experiments that have been realized with this new concept. They all use motor imagery tasks (see Section 3.4.3) and exploit changes in the amplitude of EEG rhythms. Their results are given in Chapter 6 and discussed step by step according to performance and quality metrics of Section 4.3.

4.2.1 Machine Learning Phase

The machine learning phase is preceded by the collection of calibration data. This data comprises trials that are representative of the user's brain activity during certain mental tasks. The data is labeled according to the class of performed tasks. It is the basis for the training of the machine learning algorithms. The trial collection is usually performed in so-called *runs* of approximately 50 trials, which are separated by breaks of a few minutes duration. The subject can relax during the breaks and regain concentration while the algorithms are trained.

The data is processed in the following way: First, the continuous time series signal of each channel is windowed into segments of a few seconds, during which a single mental task has been performed. This segment is later referred to as *classification interval* of a trial. Second, the linear trend is removed. Linear detrending is approximately equivalent to some simple form of high-pass filtering of the time series - an important preparation al step: removing the slowest signal components improves the fit of an autoregressive model (AR) to the remaining strongest frequency components according to the literature about PSD estimation with autoregressive models [Kay88]. The detrending step is computationally very simple as it is realized by fitting a linear regression in a least-squares sense to the time series and subtracting the resulting linear function values from the time series signal.

After this preprocessing, some potentially useful features are extracted from the detrended time series. In case of the proposed model, these features are capable of describing the frequency content of the signal. They are gained by fitting a separate autoregressive model (see section 2.3.3) to each of the time series. This is done separately for the time window of each channel of every trial. The feature extraction step delivers m AR coefficients per channel and trial. This step can strongly decrease the dimensionality as the model order m is usually chosen much smaller than the number of samples contained in a window.

In order to form a training point (x, y) consisting of the feature vector x and the class label y of a trial, the AR coefficients of all channels have to be combined to form x and represent the trial. A simple solution is the channel-wise concatenation of the AR coefficients. Although this combination method does not take the hierarchical structure of the data into account, it is a very suitable data representation for the consequent computations.

After a sufficient number of trials has been collected and preprocessed (usually about 100 trials per class), and as soon as the trials are represented in the form of vectors of concatenated AR coefficients, they can be fed into an embedded method for feature selection and classification. As the choice of an embedded method is central for this thesis, two competing variants are explained in detail in Section 4.2.4 and 4.2.5.

The result of the embedded method is threefold. It delivers a ranking of the available features, arranged in an order according to their importance for the classification task, and a ready-to-use SVM model that has been trained on the r best ranked features that is capable of classifying new trials. When performed within a cross-validation loop, it delivers an estimate for the classification error of that model on new, unseen trials. The trained SVM model and the list of r best ranked features will be transfered into the consequent feedback phase for online use.

It should be mentioned once more at this point that the notion of *features* in the context of this thesis is ambiguous. As the training vectors consist of many AR coefficients, each single dimension of the vector can be considered a feature. If the feature selection algorithms work on the basis of this notion, the ranking will deliver an ordered list of AR coefficients, ranked according to their importance for good classification results. Another notion of the term *feature* in the BCI context is the abstraction to channels as features. In this case, all AR coefficients belonging to one channel must be treated in a congeneric way and a ranking would result in a list of channels. Throughout this thesis, the term *feature* will be used for either *channel* or *AR coefficient*. The distinction will be stated explicitly in cases where it is important.

4.2.2 Feedback Phase

The feedback phase is the productive part of a BCI experiment. The user is supplied with an application interface (for examples of BCI applications see section 3.2) and is trained to use the interface. However, the goal is to shift the burden of learning as much as possible away from the user and onto the machine. The role of feedback is therefore necessarily reduced.

The recording of brain signals is done identical to the preceding steps, with one exception: the duration and timing of a trial can either be fixed or it can be coupled to the feedback application. In the latter case, the time windows used for the next steps can have a different length. An important condition for short time windows is that their length is still suitable for the estimation of the AR coefficients. The AR estimation procedure is identical to the corresponding steps in the machine learning phase.

From this point on the feedback phase is rather simple. It does not require SVM training or the search for a good feature subset. Instead, the feature subset and the classifier determined by the embedded method of the machine learning phase are used. A recorded trial can directly be processed and classified. The SVM indicates an estimate for the class label (-1 or 1) of the trial, which codes for the mental imagery class the user has produced willingly. This binary code is interpreted as a control signal for the user's application interface. Section 5.2.1 will show results gained in the feedback phase of an experiment with a spelling application.

4.2.3 Embedded Feature Selection Methods for BCI

The term *embedded* refers to the relation between the features selection method and the classification method. Following the overview paper of Pudil et al. [PNK94] the most well-known feature selection approaches are filter and wrapper methods. They have been introduced in sections 2.5.1 and 2.5.2. In filter methods, feature selection does not interact with classification but is implemented as a pre-filter step to classification. In wrapper methods, the optimization of the feature subset and the classifier is an iterative process. Both methods might be optimal for some problem settings but bear obvious risks.

The biggest problem of filter methods is that they expect features to be independent from each other and non-redundant. If this is the case, filter methods can even be the best performing choice. A clear advantage for filter methods is, that prior knowledge can be easily incorporated.

Wrapper approaches in general are only tractable for smaller feature sets as they involve many error estimation steps. The numerous estimations are not only a computational problem but can additionally lead to an overfitting of the feature selection to the specific training set, so that their evaluation must be done very cautiously.

In an embedded method the feature selection algorithm and the classification algorithm are fused together into one module. This tight coupling leads to a good feature subset while the classifier is trained. The coupling itself can be realized in different ways. well-known examples for embedded methods are neural networks if they finalize the training with the pruning of input neurons, or decision trees (e.g. C4.5). While a tree grows, a feature has to be determined for every inner node that is decisive enough to improve overall classification. Decision trees often end up with a solution that does not use all features in the inner nodes. Thus they realize an implicit feature selection embedded in the classification algorithm. It can be seen that feature selection can sometimes be expressed in the form of a sparse solution, provided by some classification algorithm. All embedded methods have in common that the feature selection exploits some property

of the classifier during its training process. In the following, two lately developed embedded methods are introduced. It is furthermore explained in detail how they must be extended for the use with hierarchical BCI data. Both methods are based on the SVM method for classification (compare Section 2.4.3) and make explicit use of properties of the SVM to select or remove features. Starting with all available features, they work in iterations, removing the least relevant feature during each iteration according to a measure. If the removal of features is tracked from the beginning to the end of the algorithm, this finally leads to a complete ranking of the features.

4.2.4 Zero-Norm Optimization (l_0-Opt)

Weston et al. [WEST03] recently suggested to minimize the zero-norm[2] $\|w\|_0 :=$ cardinality$(\{w_j : w_j \neq 0\})$ instead of minimizing the l_1-norm or l_2-norm as in standard SVMs (cp. equation (2.1)):

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \|w\|_0 + C\|\xi\|_0 \\ \text{s.t.} \quad & y_i(w \cdot x^{(i)} + b) \geq 1 - \xi_i \ (i = 1, ..., n). \end{aligned} \tag{4.1}$$

The solution of this optimization problem is expected to be much more sparse than the solution of the SVM problem (2.1). Thus feature selection is done implicitly. Unfortunately the optimization problem 4.1 is not convex any more (like the standard SVM formulation) so that an optimal solution cannot easily be found, as local minima might exist. The above problem has even been shown to be NP-hard but the authors developed the following iterative method to approximate the solution:
During every step of the iteration, the features are multiplied with a scaling factor. Once a scaling factor is zero, the corresponding feature is removed. This is repeated until only non-zero entries are left. Unfortunately, the iteration can stop with a solution w^* that has less than the desired number of zero entries. In this case, the remaining features $\{j\}$ can be ranked by another method, e.g. according to corresponding values w_j^* as it is done during each iteration step of the RFE method (see section 4.2.5).

For channel selection, where each channel hierarchically contains several frequency features (e.g. AR-coefficients like in Figure 4.1 or Welch PSD estimates for different frequency bands etc.), the Zero-Norm optimization method is adapted in the following way: the scaling factors of the features corresponding to a channel are substituted by their mean value. Thus all features of one channel are either removed completely (the channel is removed) or all features of the channel remain (the channel remains).

As in the case of SVM, the parameter C has to be estimated from the training data. This can be done by searching a limited number of values in a cross-validation setting.

4.2.5 Recursive Feature Elimination and Recursive Channel Elimination (RFE / RCE)

The Recursive Feature Elimination method was proposed by Guyon et al. [GWBV00] and is based on the concept of margin maximization for an SVM. The importance of a dimension is determined by its influence on the margin of a trained SVM. To recapitulate section 2.4.3, the margin is the distance of the support vectors of an SVM solution to the separating hyperplane. Let W be the inverse of the margin γ, X and Y form the training set and C be the regularization parameter:

$$W(X, Y, C) := \frac{1}{\gamma(X, Y, C)} = \|w\|_2$$

Like the Zero-Norm optimization, the RFE is an iterative method that starts with all features. At each iteration one SVM is trained and the features $\hat{j}$ which minimize $|W(X, Y, C) - W(X^{-j}, Y^{-j}, C)|$ are removed (typically this is one feature only). This procedure is equivalent to removing the dimension(s) $\hat{j}$ that correspond to the smallest $|w_j|$.

For the special application of channel selection, the RFE algorithm has to be adapted as well. Let $F_k \subset \{1, ..., d\}$ denote the features indices of channel k. Then for each channel k the score s_k is defined as:

$$s_k := \frac{1}{|F_k|} \sum_{l \in F_k} |w_l|$$

At every step of the iteration the channel with the lowest score is determined and all of its features (e.g. AR coefficients) are removed. Again, the regularization parameter C has to be estimated from the training data.

For the remainder of this thesis, the adapted feature selection methods will be referred to as channel selection methods. Furthermore the adapted RFE algorithm will be referred to as *Recursive Channel Elimination (RCE)*.

[2] The zero-norm or l_0-norm of a vector v is equal to number of nonzero entries of v.

4.2.6 Implementation

For the recording of raw brain signals and for the application interfaces, the BCI software package Thought Translation Device (TTD) developed by Thilo Hinterberger [BKG+00] was linked together with Matlab (a commercial software package developed by *The MathWorks Inc.*) and C++ routines. The TTD in its original form (see the description of the Tübingen BCI system in section 3.6) realizes a basic concept of a data driven signal processing pipeline with simple threshold classification. It did not provide any machine learning algorithms for data analysis or even feature selections, but implements several simple feedback applications, among them a spelling interface.

In order to implement the new signal processing concept IFS that contains an embedded channel selection method, a new module was added into the TTD data pipeline that interfaced a Matlab environment via TCP socket communication. Receiving a stream of data packets, each one containing e.g. 30 sample points from each channel, the Matlab software implemented both the machine learning phase and the feedback phase of the new signal processing concept as proposed in this thesis. During both phases, the transmission of a trial was done via the same modules. The Matlab software had to collect and concatenate data packets until the most important part of a trial, the so-called *classification interval*, had completely been received. Then the signal processing chain was activated for this trial. As described in Section 4.2, the detrending and the fitting of AR models was necessary in both, the machine learning and the feedback phase. The linear detrending could be realized directly by Matlab functions. The signal processing package for Matlab, which has the possibility to fit autoregressive models, is optionally available but cannot be found in every installation. Thus the special case of the forward-backward linear prediction (see Section 2.3.3) was implemented in two variants. The first variant was in the form of a Matlab function, the second one as a C++ function called from Matlab via the Matlab MEX interface.

The *Spider toolbox* for Matlab developed by Jason Weston et al. [WEBS05] provided a framework for the implementation of the embedded feature selection methods (l_0-Opt and RCE) as well as for the filter method based on the Fisher Criterion (see section 2.5.1). For the special case of BCI channel selection, the algorithms were modified according to sections 4.2.4 and 4.2.5. The hierarchical grouping of features is defined by a vector that codes the group membership for each input dimension. To use the Fisher Criterion as a channel ranking criterion, a similar algorithmic modification had to be performed.

After the TTD has recorded a trial and transmitted its raw signals to Matlab, it waits for a result code. During the feedback phase, this code was generated by an SVM trained individually on the data of a subject. The result code consists of the estimated class labels (either -1 or 1)[3]. The returned result code is used by the TTD to generate the user feedback. The feedback can either simply visualize the result code or use the result code to control an interface, e.g. for spelling.

4.3 Performance and Quality Metrics for Feature Selection Solutions

Different performance and quality metrics will be investigated in the following. The first one can be treated sufficiently on the spot in Subsection 4.3.1, while the remaining ones are introduced in Subsections 4.3.2, 4.3.3 and 4.3.4 but have to be checked based on results of real EEG data and online experiments in Section 6.

4.3.1 Flexibility and Computation Time

The first metric is motivated by certain requirements that (compared to other learning problems) are specific for a BCI setting. As brain data is sampled from several channels which itself may contain a set of features (e.g. frequency features) the feature space can be seen as hierarchically organized. To perform feature selection on a high level of the hierarchy, like in the case of channel selection, a suitable selection method must be flexible enough to treat all child features in a suitable and uniform way. By the algorithmic extensions of RFE the Zero-Norm method that have been proposed in the previous section, this requirement is perfectly met. Each single AR coefficient can be treated as a feature as well as a group of AR coefficients can be treated as a single feature - if the group reflects the characteristic of a channel and if channels are the atomic objects of interest.

A second requirement is speed. Methods are only suitable for BCI if both their training time (during the machine learning phase) and their online use is possible to terminate quickly enough within given time frames. For the machine learning phase, a few minutes of computing time is acceptable for a user as this duration is comparable to other breaks of a session. Only after a good feature subset has been determined, the experiment can proceed to the feedback phase.

For the feedback phase, the classifier must be able to respond to a freshly recorded trial within a split second. The exact requirements of *real time* naturally depend on the kind of feedback given. A classification on the basis of whole

[3] During the machine learning phase, when more trials have to be collected before the algorithms can be trained, a dummy value is returned to the TTD.

trials can take up to 0.5 seconds. If continuous graphical feedback is constantly given during trials, at least 25 updates per second of the classification output must be possible. This rate corresponds to a maximum of 40 ms of calculation time and gives the impression of fluent movements. During this time frame, the calculation and the reduction of all available features to the chosen subset must be performed, including other typical preprocessing steps and the final classification.

In the current paradigm there is no limitation imposed by continuous feedback, but nonetheless the system must deliver a classification result at the end of each trial without any significant delay.

When the used imagery task is a well-known BCI paradigm, literature suggests features that are reasonably useful for healthy subjects. In such settings, the search space for the fine-tuning of a good subset of features is rather limited, and speed requirements are not very difficult to meet. The problem gets more interesting if the positions of relevant cortex areas and of relevant frequency bands are only partially or not at all known. This is the case when new experimental tasks or paradigms are tested - a situation that often arises when working with impaired patients that have problems following a standard paradigm. But even under these aggravated conditions, a good feature subset has to be determined by a flexible method within the machine learning phase.

The training time of machine learning algorithms is dependent on the number of training examples (in this context, the number of example trials collected during the calibration recording) and the number of dimensions of the input space. To check the computation time of the proposed new signal processing scheme including the embedded feature selection, BCI experiments have been performed, partially including extensive feedback phases under authentic conditions (see section 5). The proposed methods all performed quick enough during offline machine learning phases and also during online training as a full feature ranking was available within less than 5 minutes. Concerning processing times during the online feedback phase, the methods under investigation were equally well-suited, as only linear classification methods were used throughout this work. Feedback to subjects was given on single trial basis and none of the subjects reported the perception of time delays. For those reasons, both the flexibility and the computation time criterion are considered fulfilled and will not be reviewed in the same intensity as the following criteria.

4.3.2 Classification Performance and Size of the Feature Set

Classification performance is a limiting factor for the communication speed (reported for example in bits per minute or characters written per minute) that can be achieved via BCI. On the other hand, a small feature subset is possibly easier to interpret for a neurophysiologist, and if a channel selection method can even save time during the electrode montage this is very valuable in the every-day work with patients.

Thus finding the smallest possible feature subset with the best possible classification performance is a typical optimization setting. In reality these are conflicting goals that need a trade-off. For this reason, the possible outcome of a feature ranking tools like the Zero-Norm optimization or the RCE method will be examined briefly.

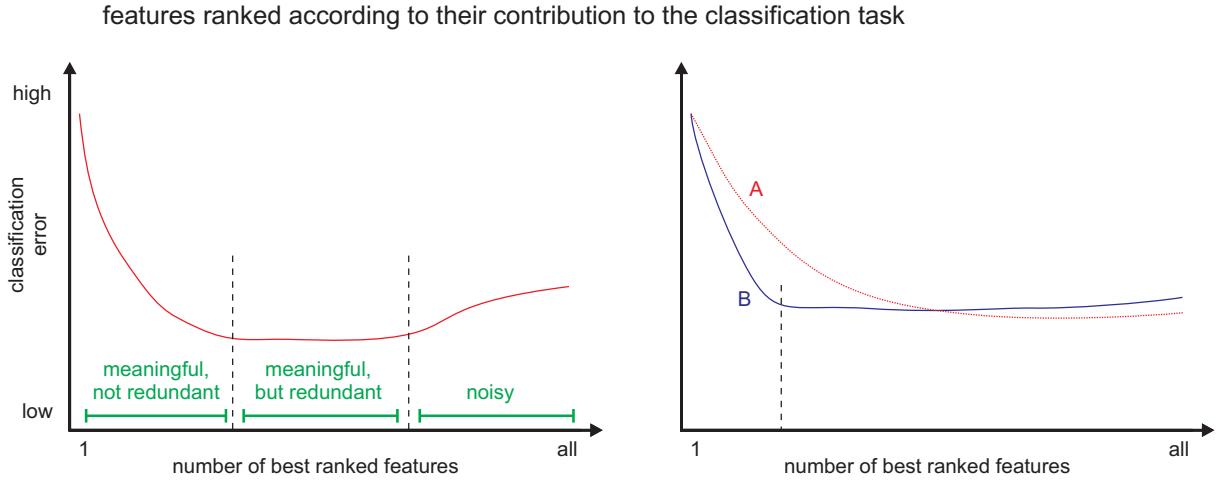


Figure 4.2: *The left plot shows a typical scheme of the classification error plotted against the size of the feature subset used for classification. The right plot shows a curve shape that reveals a good trade-off between feature set size and error (curve B) and one that does not show such a distinct point. For both figures, features are assumed to be ranked according to their contribution to the classification task.*

If such methods only deliver a ranking of features, it is not yet clear how to interpret that ranking. The additional information necessary is how well the classification task can be solved based on e.g. the 5 or 10 best ranked features. If an error estimate is available for every rank position as depicted in the left plot of Figure 4.2, this is of great value. The figure shows the idealized classification error curve vs. the size of the optimized feature subset, ranked according to class discrimination contribution. Noisy or completely uninformative features do not contribute to the class discrimination. If a limited number of data points is available, their removal can even lead to smaller classification error. Removing useful, but redundant features does not change the classification error. In contrast, the deletion of the remaining features leads to increased errors. They are sometimes characterized as *necessary but not redundant*. The shape of error curves of real problems of course varies as the number of noisy, redundant or necessary features is problem-specific. For some problems, the error curve has a clear minimum, for others, the error curve increases slowly but steadily towards less but higher ranked features. The shape of the curve can be an important factor for the comparison of feature selection methods as well as the absolute minimum of the curve. The shape might reveal how difficult it is to find a feature set that is a good compromise between size and performance (see the differences in curve A and B in the right plot of Figure 4.2). These different characteristics will be taken into consideration in the analysis in Section 6. A number of possible strategies will be presented there that propose a reasonable trade-off between feature subset size and error.

A signal processing system that incorporates feature selection must of course be able to deliver classification rates on known paradigms that are competitive to those of known methods before it can confidently be used for unknown paradigms. This argument leads to the choice of movement imagery as a group of fairly well-known experimental paradigms. The results of the proposed new IFS concept are to be checked against prior knowledge during the analysis of experiments in Section 6.

4.3.3 Plausibility and Interpretability

BCI systems are utilized in a medical environment. For this area of application it is very natural that BCI methods will only be accepted by medical doctors, if their solutions are plausible, that means, if they are in compliance with prior knowledge (in case such prior knowledge is available at all) and if the methods provide the possibility of insight, e.g. by delivering an importance ranking of features. In particular, if IFS should eventually be used to propose implantation sites (e.g. for ECoG electrodes), then this is possible only if the surgeon and the neurologist can agree to the interpretation provided by an algorithm.

Feature selection methods can (with certain limits) be used without having expert knowledge about the physiological processes that happen during a task. As the new signal processing approach proposed in this thesis is not based on prior knowledge, only very minor assumptions are made about the tasks and about the anatomical brain characteristics. The reason for choosing this approach is found in the difficult task of working with patients. On this field, expert knowledge gained from healthy subjects might be of limited use only. This is especially the case for outlier brains of handicapped patients that show brain damages and individual cognitive avoidance strategies and workarounds. An automated and individualized algorithmic approach on the other hand is in principle not more likely to fail or succeed with patient data than with data from healthy subjects. If an algorithmic feature selection approach succeeds and comes up with text book features for a well-known task, and if this result is just based on the existing data and the empirical method, then the transfer of that method to handicapped patients and new experimental paradigms can be tackled.

Apart from medical motivations, insight into a given BCI solution can in general be useful to validate the technical quality of an experiment. A selected channel set should of course contain channels close to the motor cortex, if motor imagery was used as a paradigm. If chosen channels contain muscle artifacts or are far away from where they could be expected, then the experiment conditions (electrode montage, sources of distraction etc.) might have to be improved.

The BCI experiments performed for this thesis used standard motor imagery paradigms. For this reason they are very well-suited to evaluate the interpretability and plausibility of channel selection results (see Section 6).

4.3.4 Stability and Transferability

Stability of an algorithm can be measured in many respects. A rather simple one is whether it can be transfered from the original EEG-BCI setting it was developed for to other settings, where new types of brain signal like MEG or ECoG are used. The answer to this will be derived by testing the new signal processing method in a first step on EEG data (see experiment 5.1 and then validating the success of the method in a second step on the data types MEG (in Section 6.1.5) and ECoG (Section 6.1.3).

Another aspect of robustness deals with retraining situations: It is not clear how often feature subsets derived by IFS must be optimized during a period of several sessions, in how far good feature subsets are subject to change, whether a user can adapt to a certain feature subset during a longer user training program and whether good feature subsets can be transfered from one subject to other subjects and other sessions. As the answer to most of these aspects requires

very extensive work with many patients and time-consuming user training programs, this thesis will only deal with the transfer across subjects in Section 6.3.

4.3.5 Online Performance

The hardest test environment for the new IFS method is an experimental session, where feedback is given online to the subjects and where the subjects are to fulfill a communication task with the help of a BCI. Online experiments are restricted by extremely tight schedules such that the exploration of possible feature sets and the training of a classifier must be done quickly. The results of such a test are reported in Section 6.5.

4.4 Discussion

In this section, the motivation for the specific design of the IFS chosen in this thesis shall be clarified: the choice of SVM for classification, AR models for capturing frequency characteristics of brain signals, and - most importantly - embedded feature selection as a core concept.

SVM

For the classification step the SVM algorithm was chosen, although it is generally not possible to rank the huge variety of existing classification methods in an absolute way. In such a situation, it is useful to examine general characteristics in order to make a choice.
First of all, the SVM algorithm has shown to perform strongly in a large range of benchmarks and real-world problems [DS02], including BCI [BCM02] classification problems. In addition to its good performance, SVM is comparably easy to handle: The ability to deal with high-dimensional data (as in EEG) is a specific strength of the SVM. Compared to other classification algorithms, the it is numerically and computationally capable of dealing with an unbalanced situation: the presence of many feature dimensions combined with relatively few examples [4]. It introduces only a small number of hyperparameters - usually one or two real valued parameters - that have to be set by the user. The linear SVM formulation used in this thesis contains only one parameter, which weights the regularization term in order to prevent overfitting. If prior knowledge is available for a problem and if this knowledge can be expressed by a distance measure, it can easily be utilized for the training of an SVM.
Furthermore, very quick implementations exist for the SVM (e.g. the libSVM package in C) and several Matlab toolboxes are available. Last but not least, the SVM can be utilized with the proposed embedded methods for individual feature selection, the l0-norm method and RFE.
Last but not least, SVM has an excellent theoretical foundation and theoretical bounds. It formulates the classification task as a convex optimization problem that does not have local optima such that its global solution is guaranteed to be found (provided the problem is numerically stable).

Of course the SVM algorithm is not always superior to all other classification algorithms, but due to the above benefits it is very suitable as a robust method to tackle a classification problem. For this reason newly discovered algorithms are often benchmarked against SVM. Last but not least SVM provides theoretic bounds on the quality of the solution, an ability that in principle could be exploited for feature selection. As this thesis emphasizes on feature selection, and the classification error of different feature selection strategies needs to be estimated and compared, the author committed to SVM as the principal classification tool.
These characteristics were considered a good reason to choose SVM as the classifier within the proposed new signal processing concept.

AR

Another point to clarify is the choice of autoregressive models instead of FFT-based methods for estimating the frequency characteristics of brain signals. In principle, FFT-based methods can capture the frequency information, but again, there are some tendencies that argue in favor of AR. First of all, the FFT-based methods need more fine adjustment and have more free parameters to set. Necessary parameters are for example the number of overlapping windows for the Welch method (see 2.3.3), the granularity of the frequency bands that are to be extracted from the power spectrum etc. The AR models on the other hand only need the model order as a parameter. It decides the smoothness, the compactness and the granularity of the PSD estimate. It can be set in a cross-validation procedure without problems. AR summarizes spectral peaks more compactly, which is helpful for feature selection. Furthermore, the PSD estimates gained by AR are more reliable than those gained by FFT for very short time series signals [Kay88].

[4] This ability is owed to the dual formulation of the SVM that exchanges the number of dimensions and number of constraints.

The proposed concept of an embedded feature selection algorithm is the core of the IFS concept of this thesis. Different variants of embedded concepts have to be tested against competing and well-established feature selection methods like the commonly used filter approach in section 6.1. A direct comparison between the faster filter methods and embedded methods shows that most filter methods make strong assumptions about the distribution of the data. If the distributions are known, filter methods can perform very well, but in general the distributions are not available for data gained with new BCI paradigms.
Independently from the results, there are good reasons to try embedded approaches: the feature selection step has access to interior measures of the classification algorithm (e.g. to the margin width) and can exploit this knowledge for the selection of features. As the classification method and the features selection method optimize at least partially the same objective function (large margin!) the embedded approach seems natural.
Compared to wrapper approaches, which do not have access to interior variables and often need a very large number of iterations, embedded methods are less prone to overfitting. Prior work of the author has revealed strong overfitting tendencies of a wrapper approach with SVM and genetic algorithms [SBR+03]. For this reason, a wrapper approach was discarded from this study.

IFS with Embedded Feature Selection

The performance and quality metrics introduced in Section 4.3 are the basis for judging the suitability of the proposed IFS concept of this thesis. They will need to be validated by a number of BCI experiments in Section 5 and 6.

5 IFS Experiments

The following sections 5.1, 5.3 and 5.2 will introduce the BCI experiments performed with three different recording techniques: EEG, ECoG and MEG. The experiments were executed with healthy subjects (EEG and MEG) and with epileptic patients (ECoG).

5.1 EEG Experiments

During this series of BCI experiments([LSH+04]), eight healthy, untrained, right handed, male subjects each performed two types of movement imagination tasks during a single session of approximately three hours duration. The subjects were paid for the experiment. As they did not receive any feedback about their performance during the experiment, this experiment can be categorized according to Figure 3.2 as a calibration data recording phase and a later offline machine learning phase for the analysis of the data.

5.1.1 Experimental Setup and Mental Task

EEG signals were picked up at 39 positions of the skull using silver chloride electrodes. Figure 5.1 shows a schematic plot of the electrode positions. The reference electrodes were positioned close to the ears at positions TP9 and TP10. The two electrodes Fp2 and 1cm lateral of the right eye (EOG) were chosen to capture possible EOG artifacts and eye blinks while two fronto-temporal electrodes (FT9, FT10) and two occipital electrodes (O9, O10) were positioned to detect possible jaw and neck muscle activity during the experiment. Before sampling the data at 256 Hz an analog bandpass filter with cutoff frequencies 0.1 Hz and 40 Hz was applied.

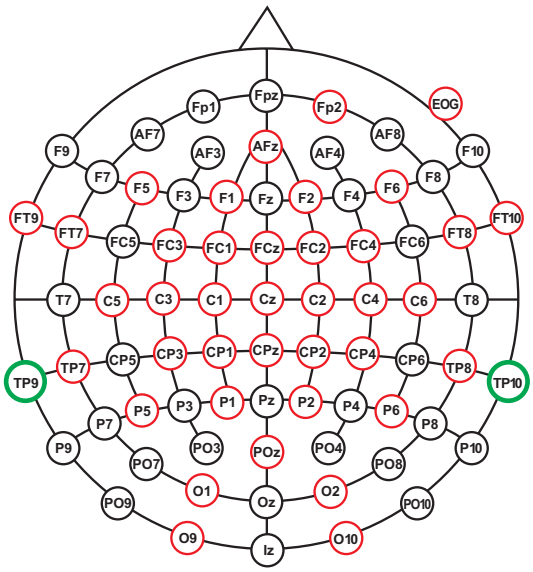


Figure 5.1: *Schematic top view on EEG electrodes on the head. The position of 38 EEG electrodes and one EOG electrode used for data acquisition are marked in red circles. The two referencing electrodes are marked in green circles.*

All subjects were seated at 1 m distance in front of a computer screen like indicated in image 5.2. The screen gave information about the trial structure (e.g. indicated relaxation phases or motor imagery phases), helped the subjects to focus their gaze to the center of the screen and provided class cues of the motor imagery tasks that should be performed. Following the experimental setup of other motor imagery BCI settings (e.g. [PNSL98]) the subjects were asked to imagine either a left or a right hand movement during each trial. During a single session, at least 400 trials were recorded from every subject. The total length of each trial was 9 seconds. Additional intertrial intervals for relaxation varied randomly between 2 and 4 seconds. They were added for two reasons: First, as the experimental task was quite

Figure 5.2: *Subject with applied EEG cap that provides up to 128 electrodes. The screen displays a cue for a movement imagery task (here: imagined tongue movement).*

boring for the subjects, the random intervals keep the subjects alert, awake and prevent the subjects from falling into a mode of rather unconscious automatic behavior. Second, although designed carefully, an experimental setting can easily be disturbed by a rhythmic source of noise that contaminates the EEG recordings in a structured way and might thus have influence onto the classifiability of tasks. The randomization breaks such regularities and prevents most structured disturbances.

Figure 5.3 depicts the trial structure. Each trial started with a blank screen. A small fixation cross was displayed in the center of the screen from second 2 to 9. A cue in the form of a small arrow pointing to the right or left side was visible for half a second starting with second 3. This cue indicated the kind of motor imagery the subject should perform. The sequence of cues was block randomized for the same reasons as described above. In order to avoid the processing of event related potentials in later processing stages only data from seconds 4 to 9 of each trial was considered for further analysis (cp. to classification interval in Figure 5.3). Feedback was not provided at any time.

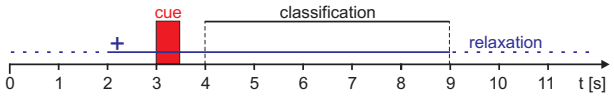


Figure 5.3: *The structure of the trials used for the motor imagery experiments with EEG recordings. The random intertrial intervals for relaxation varied from 2 to 4 seconds. The data within the classification interval is used for training the BCI system (during the machine learning phase) and for classifying unknown trials (during the feedback phase).*

5.1.2 Pre-Analysis

Pfurtscheller and da Silva have reported [PdS99] that movement related desynchronization of the μ-rhythm (8-12 Hz) is not equally strong in subjects and might even fail for various reasons (e.g. because of too short intertrial intervals that prevent a proper resynchronization). For this reason a pre-analysis was performed in order to identify and exclude subjects that did not show significant μ-activity at all. For seven of the eight subjects the μ-band was only slightly differing from the 8-12 Hz α-range usually given in the EEG literature. Only one subject showed scarcely any activity in this frequency range but instead a recognizable movement related desynchronization in around 16-20 Hz band in the β-range.

Restricted to only the 17 EEG channels that were located over or close to the motor cortex the spectral energy of the μ-band was estimated using the Welch method (see section 2.3.3). For each subject, this resulted in one parameter per trial and channel and explicitly incorporated prior knowledge about the task. The eight data sets consisting of the Welch-features were classified with linear SVMs including individual model selection for each subject. Generalization errors were estimated by 10-fold cross-validation. As for three of the eight subjects the pre-analysis showed very poor error rates close to chance level their data sets were excluded from further offline analysis.

5.1.3 Data Preprocessing

As indicated in Figure 5.3 a classification interval of 5 seconds duration was used for further offline analysis of the data. The time window contained 1280 sample points per channel. For the remaining five subjects the 1280 sample points per channel were not used directly for classification, instead rather higher-level features were extracted from these time series. Therefore, an autoregressive (AR) model (see section 2.3.3) was fitted to the time series. In order to chose a proper model order, different model orders were compared in the following way: For a given order an AR-model was fitted to each EEG sequence. For a proper model selection with 10-fold cross-validation (CV) an SVM was trained on the coefficients and the classification error was estimated. Model order 3 resulted in the best mean CV error. The AR models were calculated for all 39 channels using forward backward linear prediction [Hay95]. The three resulting coefficients per channel and trial formed the new representation of the data.

The resulting AR features are not directly linked to the μ-rhythm but describe the power spectrum in general. Thus the extraction of the AR features does not explicitly incorporate prior knowledge, although autoregressive models have been used for motor related tasks (e.g. [PNSL98]).

5.1.4 Notation

Let n denote the number of training vectors (trials) of the data sets ($n = 400$ for all five data sets) and let d denote the data dimension ($d = 3 \cdot 39 = 117$ for all five data sets). The training data for a classifier is denoted as $X = (x^{(1)}, ..., x^{(n)}) \in \mathbb{R}^{n \times d}$ with labels $Y = (y_1, ..., y_n) \in \{-1, 1\}^n$. For the task $y = -1$ denotes imagined left hand movement, $y = 1$ denotes imagined right hand movement. The terms *dimension* and *feature* are used synonymously. For $l \in \mathbb{N}, l > 1$ the feature set $M^{-j} \subset \mathbb{R}^{l-1}$ is obtained from a feature set $M \subset \mathbb{R}^l$ by removing the dimension j from every point $m \in M$ by canonical projection.

5.2 MEG Experiments

5.2.1 Experimental Setup and Mental Task

Ten healthy subjects ($\mathcal{A}$,$\mathcal{B}$,...,$\mathcal{J}$) participated in this feedback experiment [LSH$^+$05]. Each subject went through up to three experimental stages during a single session of approximately three hours duration. All MEG recordings were sampled at 625 Hz from 150 channels[1], whose position is depicted in Figure 5.4. The subjects were paid for their participation.

Stage 1

This first of altogether three stages realized the recording of calibration data and the machine learning phase according to the IFS proposed in figure 4.1. Within 60 to 90 minutes 200 trials of motor imagery tasks were recorded for each subject. The subjects were seated relaxed in front of a projection screen with their heads fixated to avoid movements. The tasks (imagined movement of the left little finger vs. the movement of the tongue) and relaxation intervals were indicated by images on a screen situated approximately 1m in front of the subjects. During this first stage, the subjects did not receive any feedback about their performance.

Figure 5.5 shows the trial structure. A trial began with a small fixation cross displayed at the center of the screen. At second one the task cue (an image of a tongue or of a left little finger) was displayed for half a second. The fixation cross appeared again at second 1.5 and disappeared at second five. The latter event marked the beginning of a relaxation interval of two to four seconds duration (randomized). Each subject performed four blocks of fifty trials.

Stage 2

During the second stage, the subjects were asked to perform the same tasks in the same trial structure as during stage 1, but received reinforcement feedback at the end of each trial. This feedback was calculated by a classifier that had been trained on the data of stage 1: After every trial that was classified correctly a smiling pictogram was displayed. Negative feedback was omitted but could be recognized by the subjects by the absence of the smiling pictogram. Depending on their performance the subjects completed two to four blocks of fifty trials.

[1] The MEG machine provided 151 channels, but channel RP31 was out of order and discarded from the very beginning of the analysis.

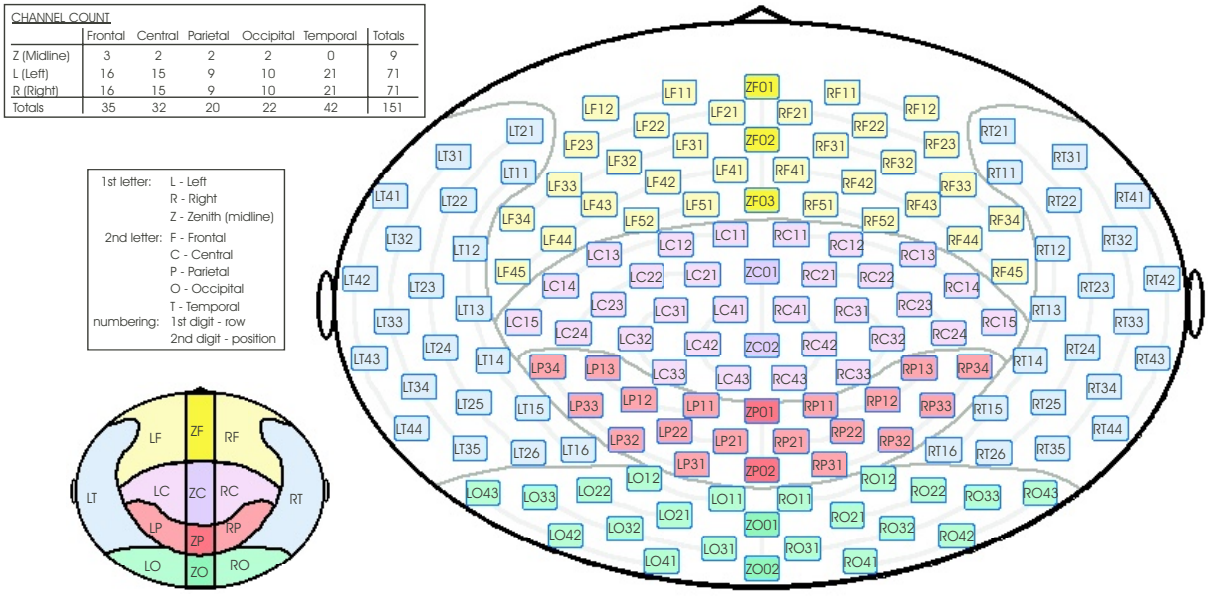


Figure 5.4: *The schematic top view on the head shows the positions of the 151 MEG channels. The colors code for frontal, temporal, parietal, occipital and central areas. As channel RP31 could not be used in the MEG experiments, 150 channels remained in the analysis. (Graphics reproduced with kind permission of VSM MedTech Ltd.)*

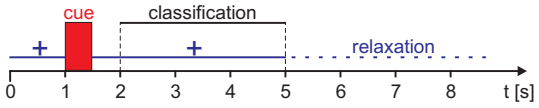


Figure 5.5: *Overview of the trial structure during stage 1 and 2 of the MEG experiment. The time interval used for classifier training started* 0.5 *seconds after the cue had ended. Relaxation intervals of randomized duration separated the trials.*

Stage 3

Those subjects that could obtain classification accuracy higher than 70% in their last block of the second stage could enter stage 3. The task was to use the imagery of little left finger and tongue to control a spelling interface with binary decisions. The goal of stage 3 was to spell a short name. For classification, the data collected during both previous experimental stages were combined to train an SVM as described in Section 5.2.2.

During the copy-spelling task, the target word was present on the screen (see Figure 6.13). At the beginning of the spelling experiment the spelling interface displayed one half of the letters of the alphabet (including some special characters). If the letter to be spelled was among the displayed ones, the subject had to imagine a tongue movement. To communicate that the letter was not displayed, the subject imagined a finger movement (see Figure 6.13). To help the subjects concentrate on the imagination task, the box of correct choice was highlighted (this spelling variant is sometimes referred to as *copy-spelling* and useful for training subjects before proceeding to *free spelling*). In the next step the selected subset of the alphabet was split into two parts again and one of them was displayed. On the last stage of this process the letter had to be confirmed and was displayed on the left part of the screen. The procedure started over again to allow the selection of further letters.
The spelling algorithm additionally allows to delete already selected letters and to reverse earlier binary decisions. Furthermore the splitting algorithm was optimized so that it reflects letter frequencies of the language. For more details please refer to [BGH+99].

5.2.2 Data Preprocessing

From every trial the interval from seconds 2 to 5 (indicated as classification interval in Figure 5.5) was extracted. This resulted in 1875 samples for each of the 150 MEG channels. For each channel and trial the least-square linear

approximation was determined and subtracted from the original time series to remove linear trends. To condense the information contained in each time series, fewer high-level features were extracted: A forward-backward AR model of order 2 was fitted to every (detrended) signal. The choice of the model order was based on a preceeding analysis of different model orders ranging from 2 to 10. For each model order, the classification error on data from all 150 channels was estimated with 20-fold cross-validation. Figure 5.6 shows the errors estimated for the ten subjects and for different model orders. For nine subjects model order 2 yielded the lowest error. The data of one subject resulted in minimal errors using model order 4. Model order 2 was used for all subjects to uniformly represent the data during further processing steps. To represent one trial, a vector of length $150 * 2$ was composed that contained the concatenated AR coefficients of all channels. The label corresponding to such a vector was defined to be -1 if the imagination task was left little finger movement and $+1$ if it was an imagined tongue movement. For every subject $\mathcal{A}$,$\mathcal{B}$,...,$\mathcal{J}$ 200 training points $(x, y) \in \mathbb{R}^{150 \cdot 2} \times \{-1, 1\}$ were used for further analysis.

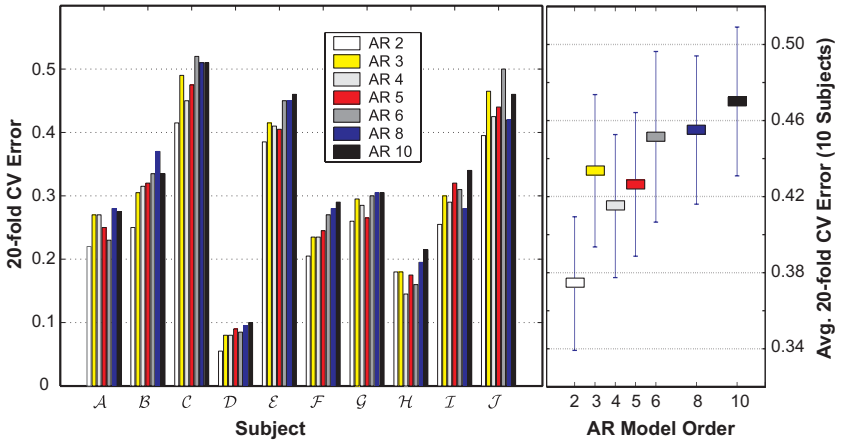


Figure 5.6: *The left plot shows the 20-fold cross-validation error of seven different AR model orders for the ten subjects of the MEG experiment. The estimates were based on data from all 150 channels. The right plot contains the averaged errors of the AR model orders along with standard errors.*

5.3 ECoG Experiments

In this section the method and examples of intra-cranial recordings of three epilepsy patients (referred to as patients I, II, III) with ECoG electrode grids that had been placed on the cortex, are presented. The patients were asked to repeatedly imagine movements of two kinds: either tongue vs. finger movements or left vs. right hand movements, while calibration data was recorded for later offline analysis. One subject (II) was paid for participation in the experiment ([LHW+05]).

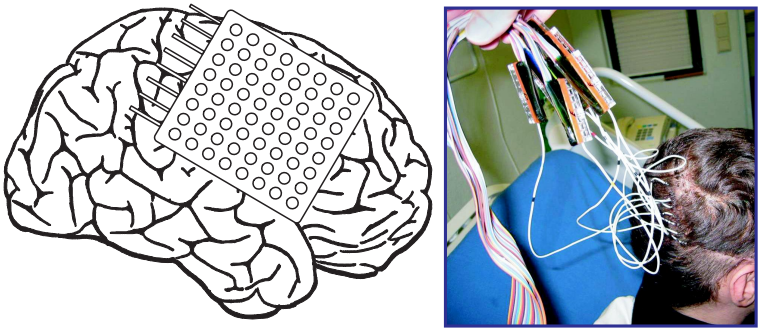


Figure 5.7: *The left picture schematically shows the position of the 8x8 electrode grid of patient II. It was placed on the right hemisphere. As shown in the right picture the electrodes are connected to the amplifier via cables that pass through the skull.*

5.3.1 ECoG and Epilepsy

The presented patients suffer from a focal epilepsy that could not be suppressed satisfactorily by medication. For this reason, the epileptic focus - the part of the brain which is responsible for the seizures - is removed by resection.

Prior to surgery, the epileptic focus has to be localized by means of a long-term recording of the brain activity. This is done by placing surface electrodes onto the gyri or into the sulci of the cortex, then the patient is monitored continuously for a few days. The skull and the meninges over the region of interest are removed for the electrode implantation. The electrodes are organized in grids, which are positioned on the cortex surface, then the incision is sutured. The electrode grids are connected to a recording device via cables. Figure 5.7 shows a schematic drawing of the location of an (8x8)-grid as implanted for patient II and a view of the skull after implantation. Over a period of 5 to 14 days and several epileptic seizures, ECoG is continuously recorded until the epileptic focus could precisely be localized [Eng93]. Prior to surgery the main functionality of those parts of the cortex that are covered by the electrodes has to be identified. This is done by slight electric stimulation via the implanted electrodes while the patient is awake. During the stimulation intervals, the patient can report sensory illusions, twitches in body parts or realize a speech impairment. The results of this stimulation give an estimate of where those cortical areas are located, which are involved in important abilities like hand movements, speech etc.

In the current setup, the patients keep the electrode implants for one to two weeks. After the implantation surgery, several days of recovery and follow-up examinations are needed. Due to the tight time constraints, it is not possible to run long experiments. Furthermore most of the patients reported problems in concentrating for a longer period of time. Therefore only a small amount of data could be collected per subject.

patient	implanted electrodes	motor imagery task	trials
I	64-grid right hem., two 4-strip interhem.	left vs. right hand	200
II	64-grid right hemisphere	little left finger vs. tongue	150
III	20-grid central, four 16-strips frontal	little right finger vs. tongue	100

Table 5.1: *Overview over positions of implanted ECoG electrode grids and strips. All three patients had an electrode grid implanted that partly covered the right or the left motor cortex.*

Experiments and Data Acquisition

The experiments were performed in the department of epileptology of the University of Bonn. The data from three epileptic patients were recorded with a sampling rate of 1000 Hz.

The electrode grids were placed on the cortex under the dura mater and covered the primary and premotor areas as well as the fronto-temporal region either of the right or left hemisphere. The grid-sizes ranged from 20 to 64 electrodes. Furthermore two of the patients had additional electrode strips implanted at other parts of the cortex (cf. Table 5.1). The imagery tasks for each patient were chosen according to two aspects: first of all, the brain areas that were expected to play a role in the imagery task execution should be covered by the electrode grid. Unfortunately this was not perfectly possible for all three patients as the grid position was not determined by the BCI experiment but by the need to treat the epilepsy. Second, the cortical areas that are active during the two mental tasks should be spatially separated. The homunculus scheme of Figure 2.5 was useful to determine a good choice of discrimination tasks for each patient, leading to motor imagery of left hand vs. right hand or little left finger vs. tongue as indicated in Table 5.1.

The patients were seated in a bed facing a monitor and were asked to repeatedly imagine two different movements. At the beginning of each trial, a small fixation cross was displayed in the center of the screen. The 4 second imagination phase started with a cue that was presented in the form of a picture showing either a tongue or a little finger for patients II and III. The cue for patient I was an arrow pointing left or right. There was a short break between the trials. The trial structure is shown in Figure 5.8.

One week after the first recording of the training data of subject II, there was a unique chance to conduct a follow-up experiment with that subject. During this second session 100 trials of new training data were collected. In order to perform an online feedback experiment with the subject, IFS with RCE and SVM training was performed on the combined data of both days. Then the subject was asked to perform two feedback runs, one of them being an online spelling task. The subject was paid for the participation in the spelling task, including a financial incentive for good performance (see Section 6.5.2).

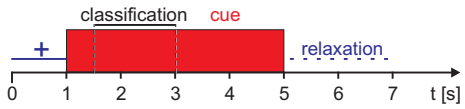


Figure 5.8: *Trial structure for ECoG experiments. The trial duration for the movement imagery tasks was 7 seconds. During the first second, a fixation cross was visible. The following imagination interval was marked by a visual cue. It started at second one and lasted for 4 seconds and was followed by a short relaxation period. For classification, time series of 1.5 seconds were extracted from the ECoG recordings, starting 0.5 seconds after the cue onset.*

5.3.2 Data Preprocessing

Starting half a second after the cue onset, a window of 1.5 seconds duration was extracted from the time series of each electrode. For every trial and every electrode an EEG sequence was obtained that consisted of 1500 samples. Following the IFS concept, the linear trend from every sequence was removed. A forward-backward AR model (see section 2.3.3) of order three was fitted to each sequence. The concatenated model parameters of the channels together with the descriptor of the imagery task (i.e. +1, -1) formed one training point. For a given number n of EEG channels, a training point (x, y) is therefore a point in $\mathbb{R}^{3n} \times \{-1, 1\}$.

6 Results

The following sections 6.1 to 6.3 show the results of experiment with EEG, MEG and ECoG data. Based on these results, the performance and other quality metrics (see Section 4.3) of the proposed IFS concept of Section 4 are analyzed.

6.1 Offline Classification Performance

This section concentrates on the aspect of classification performance in the proposed IFS concept. The IFS concept is applied to signals of three slightly different movement imagery experiments which used either EEG, MEG or ECoG as a recording technique (section 5.1, 5.2, 5.3). Based on the EEG data, three competing feature selection methods are compared in an offline analysis: a filter approach (Fisher Criterion, see 2.5.1) and two embedded approaches (Recursive Channel Elimination (RCE) and l_0-Optimization) introduced in section 4.2.3. All three methods perform channel selection and deliver a channel ranking. The main focus of this section is on the classification performance of the three approaches in terms of cross-validation error. It will be investigated if the number of channels can be reduced without a loss of performance. The characteristics of the three methods are compared, e.g. which degree of channel reduction can be achieved before the classification error starts to increase.

The winning method on the EEG data is then transfered to the MEG and ECoG data sets to check the transferability of the method onto new signal types. There again, the performance is tested in terms of feature subset sizes and estimated classification error.

As the proposed signal processing concept IFS stands for an individual optimization of features and classifiers for each patient or healthy user, the following analysis of classification performance is performed individually for each subject.

6.1.1 EEG Signals

This offline analysis is based the EEG data sets of five subjects described in section 5.1. Prior to the channel selection and individually for each subject s, the regularization parameter C_s for later SVM trainings was estimated via 10-fold cross-validation from the training data sets. Estimating the regularization parameter repeatedly during the channel selection process for each number of channels might have improved the following reported results, but also might have increased the chance of overfitting. The estimation of the generalization error for all 39 stages of the channel selection process[1] was carried out using linear SVMs for classification.

Error Estimation

If the generalization error of a feature selection method had to be estimated, a somewhat more elaborated procedure was used compared to the somewhat simpler error estimation for a classifier (see section 2.4.5). An illustration of the procedure is given in figure 6.1.
The whole data set of a subject is split up into 10 folds (*F1* to *F10*) as for usual cross-validation. In each fold, the channel selection procedure (*CS* in figure 6.1) is performed based on the *train* set of the fold only. This leads to a ranking of the 39 EEG channels which is specific for that fold. The ranking expresses an estimate of the importance of every channel for the classification task. If a channel selection method estimates high importance for a channel, then it will be located at the top of the list, while unimportant channels are positioned at the end of the ranking. As figure 6.1 shows, 39 classifiers (named $C_F^h, h = 1, ..., 39$ in the figure) are then trained for each fold F. As the classifier C_F^h has to estimate the performance of the BCI system using the h best suited channels (according to a feature selection method), it is trained on only the h best ranked features of the train set of fold F. Analogously the classifier is tested on the same h channels of the test set of F. For each fold, this results in 39 test errors E_F^1 to E_F^{39}.

During a last step, the corresponding test errors are averaged over all folds. This leads to an estimate of the generalization error for the 39 possible different numbers of selected channels. Please note that the error estimation procedure does not deliver a single unified ranking of channels, but rather one ranking per CV-fold.

The estimation results are depicted in figure 6.2. The first five plots show the individual generalization error for the five subjects against the different numbers of channels chosen by the three channel selection methods. The sixth plot in the bottom right corner shows the generalization error of the three methods averaged over the five subjects.

[1] In fact, methods RCE and l_0-Opt perform rather a channel *removal* than a channel selection.

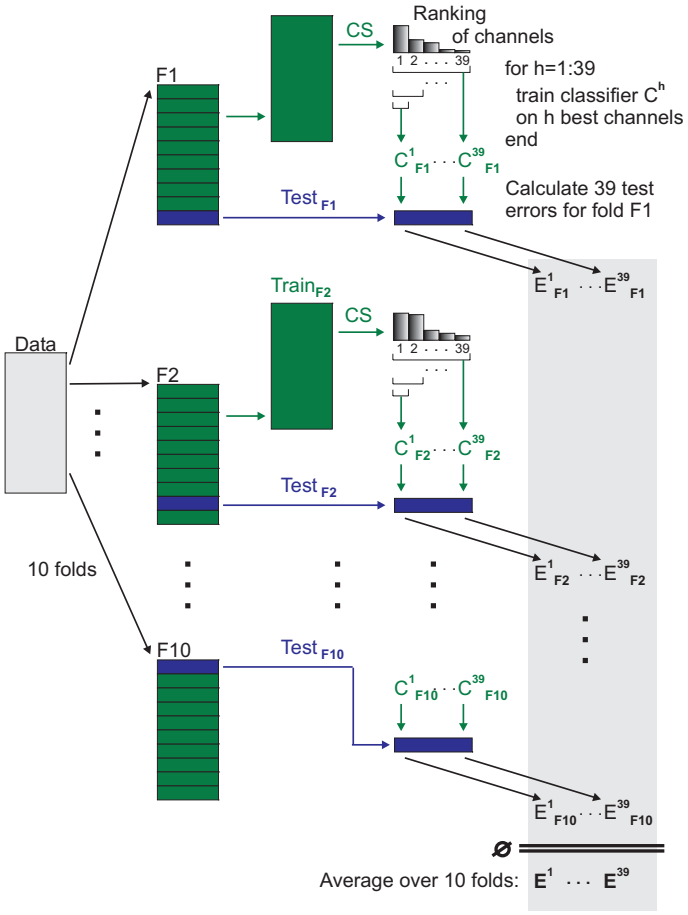


Figure 6.1: *Illustration of the procedure for feature selection and error estimation using 10-fold cross-validation.*

All five data sets can be classified well with error rates between 27 and 12% CV error. A closer look at Figure 6.2 and Table 6.5 shows that the two embedded methods RCE and l_0-Opt prove to be capable of selecting relevant channels, whereas the filter method based on the Fisher Criterion fails for some subjects and performs much worse on average. Especially for small numbers of channels RCE is slightly superior over the l_0-Opt as its error curve shows lower values for small channel sets and a sharper bend around 10 channels. For larger numbers of channels the performance of l_0-Opt is comparable to RCE.

	Subject					
Method	A	B	C	D	E	Average
Fisher	21	39	39	37	32	33.6
RCE	**18**	**12**	**19**	**12**	39	20.0
l_0-Opt	**18**	39	20	21	**27**	25.0

Table 6.1: *Comparison of three channel reduction methods. The table shows the number of EEG channels required for minimal cross-validation error.*

It is possible to reduce the number of EEG channels significantly using the RCE method for the investigated experimental paradigm. Furthermore this can on average be done without a loss of classification performance. Very nice results are gained for subject *D*: using only 8 EEG channels yields the same error estimate as the error obtained using all 39 channels. On the data set of subject *B* not only the number of channels could be reduced, but in addition a clear reduction of the cross-validation error was possible: it decreased from of 24.5% for 39 channels to 20.75% using 12 channels only.

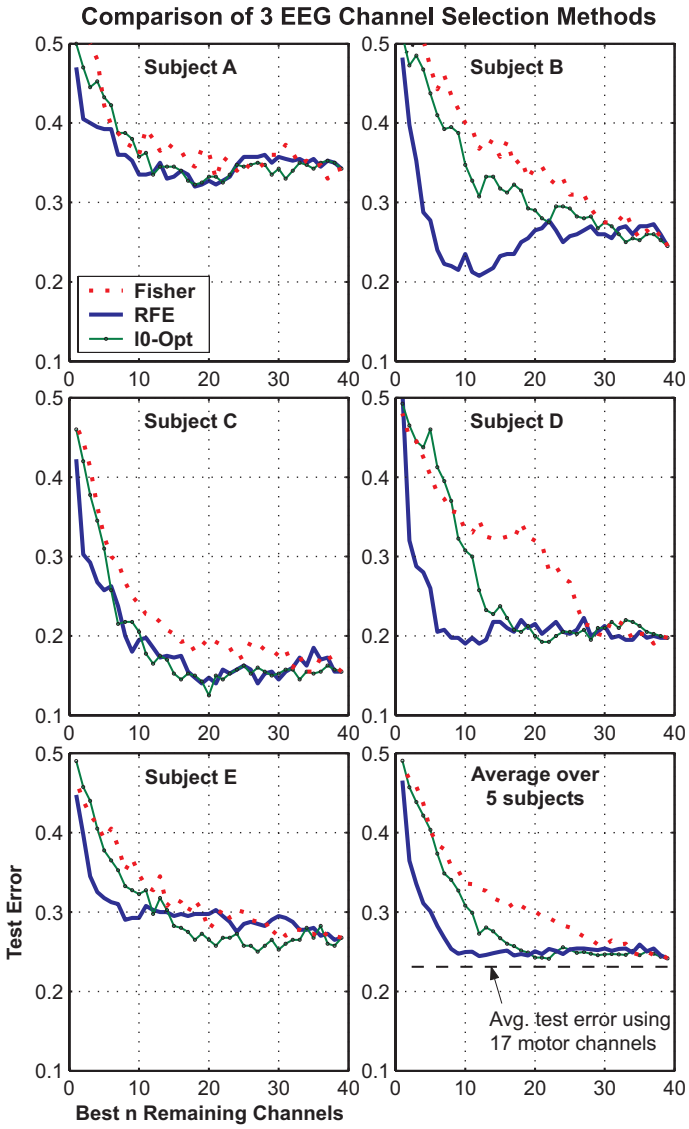


Figure 6.2: *Comparison of the three channel selection methods* Fisher Criterion, RCE *and* l_0-Opt *individually for five subjects and averaged over the subjects. For 1 to 39 best remaining channels, the corresponding test error is estimated and plotted. Method RCE allows the strongest reduction of number of channels for all subjects.*

As it is not tractable to test all ($\approx 10^{11}$) possible combinations of channels to find the overall best combination, it is difficult to provide a gold standard for channel selection. A surrogate for the gold standard might be the 17 channels located over or close to the motor cortex which can be considered a very good solution. This assumption is a valid one as the used paradigm was a left-/right imagined hand movement task. For rating the overall performance of the RCE method a classifier was trained using only these 17 channels. The result averaged over the five subjects is plotted as a dashed baseline in the bottom right figure. The average error rate for RCE (taken over all subjects) of 24% using 12 channels is very close to the error of the baseline which is 23% for 17 channels.

6.1.2 Discussion

The overall performance varies from subject to subject, but the five data sets are all classifiable between 12% and 27% of CV error. This shows that the proposed IFS is a suitable framework to exploit classification relevant information from the original high-dimensional EEG signal. Despite of the varying performance, the comparison between the filter method (Fisher Criterion), the Zero-Norm optimization and the RCE method showed clear superiority of RCE. With this embedded method, the number of channels could on average be reduced stronger than with the other competing

methods, before the estimated CV error increased. The IFS based on RCE bears the potential to save a huge amount of the time and work, which is momentarily necessary for long patient training studies. This reduction can mainly be caused by reduced electrode montage times, and for some patients even due to improved classification rates. Ideally, an initial calibration recording at the first experimental session would reveal the most useful channels for a new user so that in the following sessions the reduced montage would suffice.

The channel reduction with Zero-Norm optimization was still much better than for Fisher, but stopped too early compared to RCE. For the data of subject B the RCE method was even able to identify and remove noise channels during the first iterations. This leads to a clear improve in error for the reduced channel set compared to the set of all channels.

A comparison between RCE and a gold standard for the applied motor imagery task was surprising. On average for the five subjects only one percent of difference in error was observed between RCE and the 17 channels chosen according to prior knowledge. Particularly with regard to the decreased number of channels for RCE, this difference can be neglected.

Based on the results with EEG data, RCE was considered superior to Fisher and Zero-Norm optimization and thus applied in the following analyses of ECoG and MEG data.

6.1.3 ECoG Signals

Based upon the experiences with EEG data, the channel selection methods Fisher Criterion and on the l0-Norm are not tested any more with this new data type. Instead, the RCE method which has yielded best results is applied for ECoG signals. More specifically, this section analyzes ECoG data sets recorded from three subjects described in section 5.3. All results were gained in an offline analysis. Like for the EEG data, it is analyzed how well SVMs can *learn* classification from the data sets and how localized the classification relevant information is, i.e. how many of the recording positions are necessary to obtain high classification performance. The central question is, whether the strong reduction of recording channels observed for EEG can be reproduced for the ECoG data as well.

The following four types of subsets of ECoG channels will are used to assess the performance of RCE:

- *All:*
  the complete data containing all channels as specified in section 5.3

- *RCE_b:*
  This is the *b*est subset of channels suggested by the RCE method. As the RCE only outputs a ranked list of all channels, a validation method must be used to find the optimal number of channels. For every l in the range of one to the total number of channels, a 10-fold cross-validation error is calculated on the data of the l best-ranked channels. The validation procedure is identical to the one described for EEG in section 6.1.1. Finally the subset of channels which leads to the lowest error estimate is chosen. The number of channels contained in *RCE_b* is referred to as b.

- *RCE_2:*
  This subset consists of the two best-ranked channels. The underlying assumption for the choice of this channel set is that the relevant information might be much more localized for ECoG than for EEG due to a higher spatial resolution of the ECoG (see section 2.2). Thus a small number of correctly chosen channels might contain sufficient information for classification purposes.

- *Random_2:*
  This set is formed by two channels drawn at random.

To evaluate the four subsets, it is necessary to estimate the classification performance of an SVM that is trained on each one of the subsets. Therefore the prediction error of such an SVM has to be calculated on a separate test set. An additional level of cross-validation was introduced for this purpose. To reduce the variability of the results, its outer loop was repeated 50 times. The scheme is described as pseudo code in algorithm 1 for the example of *RCE_b*.

Due to the extra loop, the double CV-scheme is computationally expensive. This was implemented for the ECoG analysis, because of the considerably smaller number of examples of the ECoG data set and the slightly different problem setting in this section: Instead of investigating the channel reduction performance of RCE only, now four competing methods for choosing the subset are investigated - finding the optimal b has turned into a subproblem.

Solving this subproblem is the job of the inner 10-fold CV (line 6): After the error has been estimated for all number of channels, the number showing the smallest error is chosen. This inner step is comparable with the CV scheme that has been used for the evaluation of the EEG data in section 6.1.1. Figure 6.2 shows examples of these estimates. Using only 100 ECoG trials (for subject III) compared to 400 EEG trials can of course lead to bigger random variations of the error curve than those in figure 6.1.1. Choosing the optimal number of channels (b) based on a noisy curve can be source of

Algorithm 1 *Error estimation scheme for channel subset **RCE_b** using double CV.*

Require: preprocessed ECoG data of one subject
1: **for** (cntMainFolds = 1 to 50) **do**
2: split data randomly: 80% training set, 20% test set
3: with training set do: {contains all channels}
4: 10-fold CV: find *good* ridge r
5: rank channels with RCE using r
6: 10-fold CV: estimate number of channels b for subset **RCE_b**
7: reduce training set to b best ranked channels
8: with training set do: {reduced channels}
9: 10-fold CV: find *good* ridge r_red
10: train SVM S using ridge r_red
11: reduce test set to b best ranked channels
12: with test set do: {reduced channels}
13: test S
14: save test error and number of good channels b
15: **end for**
Output: mean error, standard deviation, mean b

overfitting. The assessment of generalization error rates independent of such overfitting effects was the main motivation for the implementation of the outer loop. As the 20% of extra test data in the outer loop have not participated in the decision for b the reliability of the error estimates is improved and the potential effect of overfitting is excluded. Finally the double CV-scheme estimates the average performance very precisely while its variability during the 50 repetitions is described by the variance. This allows for a fair comparison of the four subset types *All*, *RCE_b*, *RCE_2* and *Random_2*.

For estimating the performance of type *RCE_2* subsets, the algorithm needs to be varied only slightly: line 6 has to be replaced by simply setting $b := 2$ (see algorithm 4 in appendix B). The algorithms for type *All* and *Random_2* subsets are straightforward and even simpler. They can be found in appendix B as algorithms 3 and 5.

	All		*RCE_b*		*RCE_2*	*Random_2*
patient	#chan	error	#chan	error	error	error
I	74	0.382 ± 0.071	5.8	**0.243** ± 0.063	0.244 ± 0.078	chance level
II	64	**0.257** ± 0.076	21.5	0.268 ± 0.080	0.309 ± 0.086	0.419 ± 0.123
III	84	0.4 ± 0.1	5.0	0.233 ± 0.13	**0.175** ± 0.078	chance level
Average	74	0.346	10.8	0.245	0.243	chance level

Table 6.2: *Comparison of the classification accuracy of SVMs trained on the data of different ECoG channel subsets: All ECoG channels, the subset determined by Recursive Channel Elimination (RCE), the subset consisting of the two best ranked channels by RCE and two randomly drawn channels. The mean errors of 50 repetitions are given along with the standard deviations. In addition, the original number of channels and the average subset size proposed by RCE is given. The smallest error values are printed in bold face. Note that for subjects I and III the channel selection does increase performance compared to all channels (values printed in red).*

Table 6.2 summarizes the results of the error estimations for the four subset types. They show that the generalization ability can on average be significantly increased by both RCE subset types.

For **patient I** both RCE based subset types perform equally well. The error changes from 38% to 24% when using the channel subsets suggested by RCE while only 5.8 (*RCE_b*) or 2 (*RCE_2*) channels are needed. This is a very strong reduction to less than 8% (*RCE_b*) or less than 3% of the full number of channels.

For **patient II** the channel selection is not as successful. In this case, the channel selection process does not yield a decreased error, but rather an slight increase of approximately 1% for *RCE_b* and 5% for *RCE_2*. The necessary number of channels go down to 34% of the original 64 channels for *RCE_b*.

The high error for **patient III** of 40% can be reduced to 23% using 5 channels on average by *RCE_b* and even down to *17.5%* for *RCE_2*.

The direct comparison of the results using the two best ranked channels to two randomly chosen channels shows how well the RCE ranking method works: The average error can be divided in half, and for the case of patient three the error

drops from chance level to only 17.5%.

The reason for the big difference in performance for patient III for the subsets *All* and *RCE_2* might be that out of the 84 electrodes, only 20 are located over or close to the motor cortex. As the remaining 60 channels possibly contribute much noise they prevent learning of the classifier, while RCE successfully identifies the important electrodes. In contrast to patient III, the electrodes of patient II are all more or less located close to the motor cortex. This explains why data from two randomly drawn channels can yield a classification rate better than chance. Furthermore patient II had the fewest electrodes implanted and thus the chance of randomly choosing an electrode close to an important location is higher than for the other two patients.

6.1.4 Discussion

This section analyzed ECoG data recorded from three epilepsy patients during a motor imagery experiment. Although only few data points were collected, a number of conclusions can be drawn.

The data of all three patients is classifiable reasonably well, although the position of the recording grid was chosen according to the epilepsy monitoring and not primarily in order to cover the motor cortex. The error rates range from 17.5% to 25.7%. Although this is far from perfect classification, it has to be considered that the results were gained from only 1.5 seconds of data from each trial and that extremely few training points (100-200) were available. The proposed feature selection method shows a very effective feature reduction: it successfully identifies subsets of ECoG-channels. This reduction does not only lead to a good, but even to an improved classification performance. Especially if RCE achieves a strong reduction of channels like for subject I and III, the error drops impressively below the error on all channels. But even on average (including subject II), RCE leads to a strongly improved classification rate compared to a classifier that is based on the data of all available channels. This example illustrates the optimal case, when non-informative features (that prevent proper classification) are removed by feature selection.

Poor classification rates using two randomly drawn channels and good classification rates using the two best-ranked channels by RCE suggests that classification relevant information for ECoG is focused on relatively small areas of the cortex.

The comparison of *RCE_b* and *RCE_2* shows that the assumption of Graimann [GHLP04] for very small functional cortex areas does not generally hold. In their work, the authors had restricted the search of good features due to localization arguments to only one ECoG channel. The present analysis has instead shown that for two of the three data sets it was advantageous to use even more than 2 channels. Nevertheless, task relevant information seems to be more localized in ECoG compared to EEG.

6.1.5 MEG Signals

The analysis of the ECoG data has revealed that the important channels are clearly more localized in ECoG than in EEG data. This explained why for some users even very few good channels leads to satisfying classification results. As MEG has a higher spatial resolution than EEG and as it is recorded in a shielded environment, the situation might be similar for MEG data.

The data of 10 subjects was analyzed separately, according to the procedure outlined in Algorithm 2. In step 2 of the algorithm, the data is randomly split into a training set which contains 80% of the data and a test set which contains the remaining 20%. On the basis of the training data, the value of the ridge-hyperparameter for SVM is selected which leads to the smallest CV-error (this happens in step 3 of Algorithm 2). This ridge value is used by the RCE procedure which outputs a ranking of the 150 MEG channels (step 5).

In this study, a new method for choosing the number of channels b is introduced: the approach was that the number of channels should be minimized, while the error must not increase significantly. This was done in an offline analysis of the MEG experiments, based on the IFS concept with RCE.

The core idea is to taken the variance of the error curve into account. The smallest subset of channels, whose error is still in the range of one or two standard errors (*STE*) of the minimal point of the curve might still be a good solution but contain only a few channels.

In the steps following ridge selection, the training data is restricted to the $N \in \{1, ..., 150\}$ best ranked channels. For each N, the generalization error of an SVM trained with the reduced data using the cross-validation technique is performed. The number b of best channels is selected to be the minimum number of channels yielding an error estimate which deviates less than either one standard error or less than two standard errors from the minimal error estimate (these two criteria will be called *RCE_STE1* and *RCE_STE2* respectively). This procedure is visualized in Figure 6.3 for the range of one standard error (*RCE_STE1*). In this example, the 18 best ranked channels yield the lowest average error (marked in green) and the 11 best ranked channels would be selected. Note that the estimates of the CV error show random fluctuations.

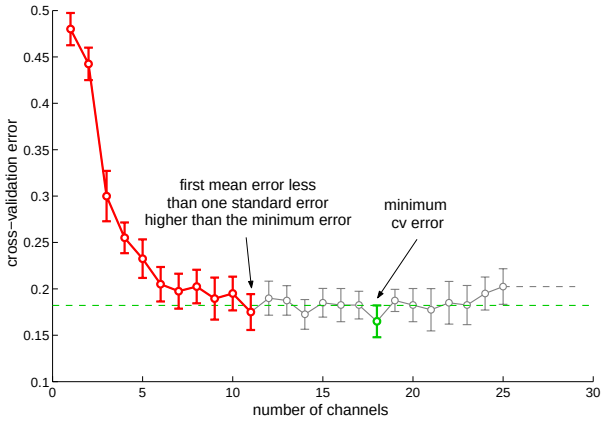


Figure 6.3: *Scheme for the estimation of channel subset size. Every point on this schematic graph corresponds to an estimate of the 20-fold cross-validation error (y-axis) using the N best ranked channels only. The bars denote standard errors.*

In step 7 the data of the b best ranked channels is extracted. The ridge is optimized again and an SVM using the best ridge is trained. Finally the SVM is tested on the reduced test set. Like in the previous section, this procedure is repeated fifty times to obtain stable results.

As before, the overall error of the channel subsets are estimated in a double CV scheme. Compared to earlier algorithms, it is important to notice, that the inner CV steps now loop 20 times instead of 10 times which leads to more reliable estimates of the standard error values for each step of RCE.

Algorithm 2 *Error estimation scheme for channel subset **STE1** using double CV.*

Require: preprocessed MEG data of one subject
1: **for** (cntMainFolds $= 1$ to 50) **do**
2: split data randomly: 80% training set, 20% test set
3: with training set do: {all channels}
4: 20-fold CV: find *good* ridge r
5: rank channels with RCE using r
6: 20-fold CV: estimate number of channels b_{STE1} using r
7: reduce training set to b_{STE1} best ranked channels
8: with training set do: {reduced channels}
9: 20-fold CV: find *good* ridge r_red
10: train SVM S using ridge r_red
11: reduce test set to b_{STE1} best ranked channels
12: with test set do: {reduced channels}
13: test S
14: save test error and number of good channels b_{STE1}
15: **end for**
Output: mean error, standard deviation, mean b_{STE1}

The error rates obtained with Algorithm 2 on data from ten subjects are summarized in Table 6.3. When using all channels the cross-validation error ranges from chance level (subjects $\mathcal{C}$ and $\mathcal{J}$) to 8% (subject $\mathcal{D}$). When using the STE1- or STE2-estimate to determine a feature subset, the average error (taken over the subjects) increases only slightly from 29.9% to 31.8% (STE1) or 31.2% (STE2).

The number of channels was reduced significantly as can be seen in Table 6.4. The STE2-estimate suggested on average channel sets of size 16.6 and the STE1-estimate suggested channel sets of size 7.1. Note that this is less than 5% of the original 150 channels.

6.1.6 Discussion

The application of the IFS concept with RCE on data of an MEG experiment showed, that the method is well-suited to reduce the number of necessary channels without significant loss of performance. For a MEG BCI, this reduction is only of minor use, as the application of extra channels does not require more time than the application of a few

	STE1-estimate	STE2-estimate	all channels
$\mathcal{A}$	0.297 ± 0.076	0.285 ± 0.076	0.258 ± 0.076
$\mathcal{B}$	0.327 ± 0.072	0.328 ± 0.072	0.287 ± 0.072
$\mathcal{C}$	0.484 ± 0.076	0.473 ± 0.076	0.441 ± 0.076
$\mathcal{D}$	0.098 ± 0.052	0.085 ± 0.052	0.080 ± 0.052
$\mathcal{E}$	0.378 ± 0.071	0.395 ± 0.071	0.403 ± 0.071
$\mathcal{F}$	0.230 ± 0.071	0.227 ± 0.071	0.239 ± 0.071
$\mathcal{G}$	0.339 ± 0.076	0.310 ± 0.076	0.313 ± 0.076
$\mathcal{H}$	0.237 ± 0.065	0.218 ± 0.065	0.193 ± 0.065
$\mathcal{I}$	0.335 ± 0.070	0.333 ± 0.070	0.323 ± 0.070
$\mathcal{J}$	0.460 ± 0.078	0.470 ± 0.078	0.456 ± 0.078
Avg.	0.318 ± 0.113	0.312 ± 0.119	0.299 ± 0.116

Table 6.3: *Classification errors and standard deviations for the MEG data of ten subjects. The rightmost column shows CV-errors obtained by using the data of all channels. The error rates for STE1- or STE2 subsets are contained in the first two columns.*

	STE1-ESTIMATE	STE2-ESTIMATE	ALL CHANNELS
$\mathcal{A}$	6.1 ± 3.7	13.5 ± 11.2	150
$\mathcal{B}$	15.0 ± 21.7	33.3 ± 30.6	150
$\mathcal{C}$	13.1 ± 17.2	30.9 ± 30.1	150
$\mathcal{D}$	6.8 ± 3.6	15.7 ± 17.7	150
$\mathcal{E}$	1.2 ± 0.8	5.4 ± 16.8	150
$\mathcal{F}$	1.7 ± 1.0	4.1 ± 10.9	150
$\mathcal{G}$	3.3 ± 2.5	8.5 ± 6.0	150
$\mathcal{H}$	8.2 ± 5.0	21.0 ± 22.9	150
$\mathcal{I}$	5.6 ± 9.6	12.5 ± 17.2	150
$\mathcal{J}$	10.0 ± 15.4	21.3 ± 23.9	150
AVG.	7.1 ± 4.6	16.6 ± 10.0	150

Table 6.4: *This table summarizes the average sizes of MEG channel subsets and their standard deviations. The subsets were suggested by the STE1- or STE2-estimates for ten subjects. On average the STE2 method results in subsets of size* 16.6*. Using the STE1-estimate achieves an average subset size of 7.1 .*

channels. But it is important to see, that the IFS concept is not only applicable for EEG data but also for sources of brain signals, with localization characteristics better than EEG. As we will see, feature selection also has the advantages that its results allow for the detection of muscle artifacts and for an interpretation of the brain activity patterns. Furthermore, the reduced channel set leads to a much faster signal processing system.

6.2 Comparison with Prior Knowledge

The previous section showed the impressive ability of the new signal processing concept to reduce the number of necessary channels enormously while the classification performance is not significantly decreased, and in some cases, such as Patients I and III in the ECoG experiment, significantly increased. This ability accelerates the experimental preparation and in general takes computational load from the signal processing flow.

An open and interesting question is, whether the selected channels can provide some extra information about the BCI task, and (of course) whether the selected channels seem to be adequate for an expert considering the mental task at hand.

The mental tasks used in the experiments of this thesis were chosen deliberately. Imagined right or left hand, tongue or finger movement are long used in the BCI community and the physiological background can be believed to be known to great extent. The brain data for example, that is recorded during imagined left-/ or right hand movement can be classified satisfactorily if channels close to the motor cortex, e.g. around positions C3 and C4, are chosen [WM94]. This does not mean that these tasks cannot be classified with data from any other channels, but plausible solutions are expected to contain channels closest to these cortical areas. Of course the RCE method for channel selection has no notion about this *a priori* knowledge but decides a suitable channel subset based on the data alone.

The following subsections will examine the channel subsets that were selected by the RCE during the proposed new signal processing concept. Their plausibility is discussed and some interesting visualization methods based upon these results are introduced.

6.2.1 EEG Signals

Table 6.5 contains the channel rankings, which are obtained by applying RCE to the data set of each subject. As the RCE method has outperformed CF and l_0-Opt, the rankings in table 6.5 were exclusively calculated by RCE. To interpret the table it is useful to have a closer look at figure 6.4.

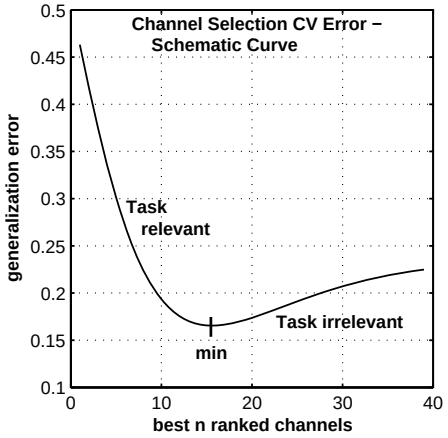


Figure 6.4: *Idealized generalization error curve for a balanced two-class problem using a channel selection method in the presence of irrelevant channels. When removing channels iteratively the classification error decreases slightly until all irrelevant channels are removed. Removing more channels results in an increase of error.*

It shows an idealized curve for an estimate of the generalization error when using a channel or feature selection method. Very often it is possible to reduce the number of channels without loss of performance. If these estimated generalization error curves are known (as for the example of the EEG data in figure 6.2), a heuristic estimate on the number of irrelevant channels can be obtained for each subject. One entry in each column of table 6.5 is underlined. The row number of that entry is an estimate for the rank position that divides task relevant channels from task irrelevant ones. E.g. for subject *D* figure 6.2 shows a local minimum of the RCE generalization error curve at 10 channels. Thus the best 10 selected channels can be used without increasing the error estimate.

The positions of the 17 channels over or close to the motor cortex were marked with a gray background. Except for very few of them, these channels have a high rank. This observation gives a positive answer to the question, whether RCE is able to identify physiologically reasonable channels.

Taken one step further, these five rankings can be analyzed in an inverted way by examining the highest ranks. In four of the subjects, only a few non-motor channels were ranked above the marked minimum-error positions (underlined

Table 6.5: RCE Ranking of 39 EEG Channels

	Subjects				
Rank	A	B	C	D	E
1	C4	CP4	CP4	FC4	CP4
2	CP4	C3	CP3	C4	CPz
3	CP2	C4	C4	CP2	C2
4	C2	FC4	C2	CP1	FC3
5	Cz	FT9	C1	C3	C4
6	FC4	FT10	CPz	FC3	C1
7	FC2	CP1	CP2	C2	FCz
8	C3	C1	C3	C1	FC4
9	CP3	F6	F1	FC2	C3
10	F1	Fp2	FC1	FC1	POz
11	F2	FC1	FC2	FT10	P6
12	C1	AFz	C5	FCz	O10
13	FC3	C2	FT7	F2	FC1
14	CPz	P6	F2	FT9	C6
15	CP1	CP2	FC3	F1	C5
16	FCz	P1	C6	C5	Cz
17	P2	EOG	P1	F5	CP2
18	P1	FC3	CP1	C6	O1
19	C6	Cz	O1	POz	O9
20	AFz	C6	POz	AFz	TP8
21	F5	TP8	TP7	FT8	CP1
22	C5	P2	Fp2	Fp2	P1
23	FT9	POz	P5	P2	F1
24	FC1	F2	P6	P1	F2
25	FT7	FC2	FC4	O10	FT7
26	POz	O10	EOG	O9	TP7
27	O2	O1	FCz	P6	P2
28	P6	CP3	AFz	O1	O2
29	EOG	FCz	Cz	P5	FT8
30	P5	P5	FT10	EOG	FT10
31	FT10	TP7	F5	Cz	F5
32	Fp2	O9	TP8	CPz	EOG
33	FT8	CPz	P2	F6	P5
34	O1	O2	O9	O2	CP3
35	TP8	F5	O2	TP7	FC2
36	O9	FT7	O10	CP3	FT9
37	O10	F1	F6	CP4	Fp2
38	F6	FT8	FT8	FT7	AFz
39	TP7	C5	FT9	TP8	F6

The ranking of the 39 EEG channels was calculated by the RCE method. The 17 channels over or close to motor areas of the cortex are marked with gray background for all five subjects. Underlined rank positions mark the estimated minimum of the RCE error curve for every subject from which on the error rate increases prominently (see figure 6.2 for the individual error curves).

ranks). Either these channels contain useful information about the classification task as well (for example if their cortical regions help to prepare the mental task), or these channels reveal slight uncertainties of the RCE method. This point cannot be answered definitively, but the analysis of the channel ranking of subject B might give a hint: for this subject the channels FT9, FT10, F6 and FP2 are rated relevant according to the ranking. While it is still plausible to find channel F6 early in the ranking (it is still close to motor areas, compare figure 5.1), the channels FT9, FT10, and FP2 are far away from motor areas and furthermore prone to pick up artifact signals from jaw muscles (FT9 and FT10) or from eye movement (FP2) due to their proximity to these artifact sources. To assess the significance of artifact usage, the following action was taken:

- The classification error was estimated using exactly the seventeen motor channels and compared to the error using the motor channels plus FT9, FT10, FP2, and EOG. Indeed by adding the four artifact channels the error could be reduced from 24% to 21%.
- An SVM was trained based on these artifact channels only. The performance was poor: only 0.55% accuracy could be reached in a 10-fold CV SVM training although the ridge was explicitly optimized for this test.

These findings indicate that, although feedback was not provided during the EEG experiment, this subject showed

task-related muscle activity. However, performance did not rely on muscle activity alone. The other four subjects did not accompany the left/right tasks with task-related muscle movements. This observation was supported by visual inspection and frequency analysis of the raw EEG signal - only very little muscle activity or other forms of artifacts could be detected.

It is concluded that the RCE method is capable of estimating physiologically meaningful EEG channels for the imagined left/right hand paradigm. Furthermore the method can be applied to detect the unwanted use of artifacts for BCI control.

Visualization Based on Channel Rankings

For visualization purposes a score calculated by RCE was assigned to each channel. The channels below the underlined entries of table 6.5 receive a score of 0. The ones above the underlined entries are mapped to the gray value scale according to their rank. Figures 6.5 and 6.6 show the task relevant channels for the five subjects. Black regions in both plots mark channels irrelevant for the classification task whereas white regions mark relevant ones.

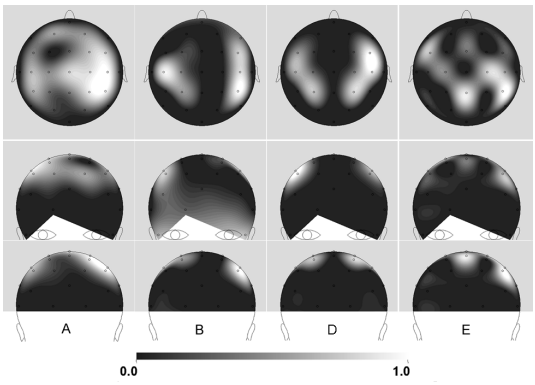


Figure 6.5: *Visualization of task relevant regions for subjects A,B,D and E (one subject per column) during imagined hand movements. The score for each channel was obtained by using the RCE method and is based on the full duration of 5 seconds. The top row depicts the view from above, the second and third row show the frontal view and view from the back. Please see the left column of figure 6.6 as well for the corresponding mapping of subject C.*

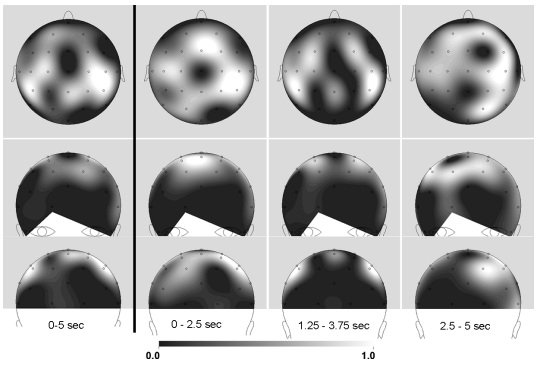


Figure 6.6: *Visualization of task relevant regions for subject C (top, front and back view). The leftmost column shows the scores obtained by RCE based on the complete duration of 5 seconds. The remaining three columns show the development of the scores over time. The rankings were obtained by applying the RCE method separately on the three shorter, overlapping time windows.*

For all subjects the informative regions are located close to the motor cortex. Subject *D* shows a clear and symmetrical concentration of important channels. The second column of figure 6.5 shows that subject *B* has additional important channels outside the motor area probably resulting from muscle and eye activity (as discussed above). This is especially detectable by the visualization in frontal view.
As the generalization error was minimal for the data of subject *C* a closer examination of this data was performed.

Columns 2 to 4 of figure 6.6 visualize the spatial distribution of task specific information *over time.* The training data was split into three overlapping windows each of 2.5 seconds length. For every time window, a separate channel selection via RCE was applied. It can be observed that the three resulting score patterns vary from window to window. This can be caused by variation in the channel selection results or by the fact that the task related activation pattern changes over time.

6.2.2 ECoG Signals

Electric stimulation of the implanted ECoG electrodes helps to identify cortical regions that are covered by the electrode grid prior to surgery. This information can be used to validate the results of the RCE method. Its result are reasonable, if RCE mainly or exclusively picks channels that have been identified as so-called motor channels by electric stimulation.

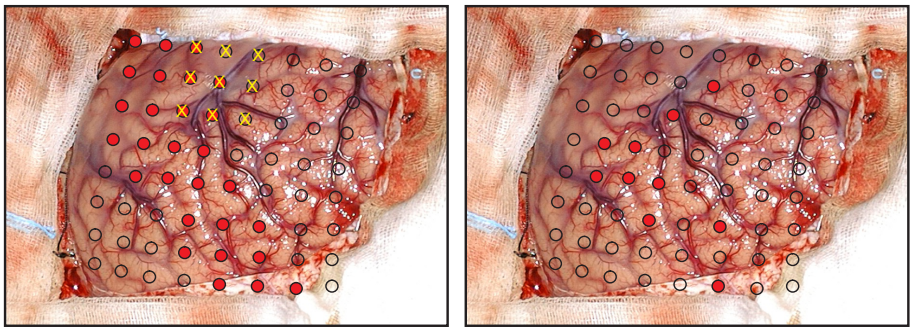


Figure 6.7: *Results of electric stimulation of the cortex. The red dots on the left picture mark the motor cortex of patient II as identified by the electric stimulation. The positions marked with yellow crosses correspond to the epileptic focus. The red points on the right image are the best ranked channels by Recursive Channel Elimination (RCE). The RCE-channels correspond well to the results from the electric stimulation diagnosis.*

The results of this kind of analysis is shown in Figure 6.7 for the case of patient II, for whom a quite detailed stimulation diagnosis was available. The ten best ranked RCE-channels are marked in the right plot of Figure 6.7. They correspond well with the results from the electric stimulation (left plot). Only one channel chosen by RCE (top right red marked channel) is slightly outside the area of motor channels.

6.2.3 MEG Signals

Using MEG data, the RCE method again suggests very reasonable channel subsets. For subjects $\mathcal{H}$ and $\mathcal{D}$ the results are visualized in Figure 6.8. Of course, a dedicated electric stimulation is not available for the MEG subjects in order to provide ground truth about the position of the motor and pre-motor areas. However, the dewar helmet of the MEG apparatus effectively restricts the head position during the signal recording in a way that the mapping of MEG channels to cortical areas is quite reliable. Figure 2.4 is useful to estimate the quality of the RCE results: For both subjects,

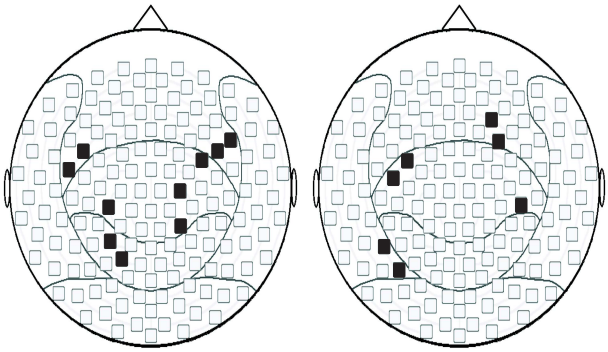


Figure 6.8: *Examples for best ranked MEG channels. The left figure displays the position of the ten best ranked MEG channels for subject $\mathcal{H}$, the right figure displays the position of the seven best ranked MEG channels for subject $\mathcal{D}$. The channels are located over or close to the motor cortex.*

the chosen sensor subsets lie over or close to the motor cortex which agrees well with the underlying cognitive process expected for the experimental task. In the case of the MEG channels, RCE did not give high rank to possible artifact channels located at the exterior of the scalp.

6.2.4 Discussion

For all experiments and signal types, the IFS method worked reliably, although no explicit prior knowledge about the mental task or about its underlying neural substrates was incorporated. Nevertheless, channels that are well-known to be important from a neuro-physiological point of view were consistently selected by RCE whereas task irrelevant channels were disregarded. In the one single case when this did not hold true, the method has proved useful to reveal the use of task related muscular activity.

The results suggest that the RCE method can be used for new experimental paradigms in future BCI research - especially if no a priori knowledge about the location of important channels is available.

Furthermore, in some cases medical staff has to be convinced about the validity of the outcome of IFS. The proposed visualization of channel scores via IFS scores can achieve this task. It supports the analysis of BCI experiments, reveal patterns of classifiable channels or sensors that are pick up misleading artifacts or visualize the spatial change of task relevant information over time- in short, the it can ease the choice of channel subgroups. Therefore visualization should be used if possible.

6.3 Transferability Across Subjects

If data from several subjects but for the same task is available, the question arises, whether a set of channels selected for one subject is useful for other subjects as well, that means, whether generalized conclusions can be drawn about channels relevant for the classification of a certain mental task across subjects. This will be analyzed on the basis of the BCI experiments based on EEG data introduced in Section 5.1).

The transfer of EEG channel rankings can be difficult for several reasons:

- The head shapes might vary between subjects. This limits the comparability of electrode positions and channel selection outcomes.
- Subjects might use different mental representations for a task, even if they are instructed carefully.
- Cortex areas important for the mental task are probably organized slightly differently between subjects. This limits the comparability of localized activity patterns.

Luckily, motor imagery tasks involve a comparably big part of the cortex and small dislocations of EEG electrodes (e.g. around typical motor positions C3 and C4) usually do not lead to profound error increase for the classification of brain activity.

Nevertheless it is very important to investigate the reliability of cross-subject channel selection: on the one hand, even a slightly increased classification error leads to a large drop in the information rate for a BCI system [SKSP03]; on the other hand, mental tasks that do not show the advantages of motor imagery will more and more be focused on by BCI research, as the existing systems will be expanded to multi-class BCIs or for patients whose motor areas are not intact.

The following pages show results for the RCE method on cross-subject data. First of all, RCE is applied to combined data of all five subjects (Section 6.3.3). Its results are compared with the individual channel rankings obtained from the five subjects. Then the transfer of rankings is investigated (Section 6.3.4): RCE calculates rankings of data combined from 4 subjects before these rankings are tested on the unseen data set of the last subject.

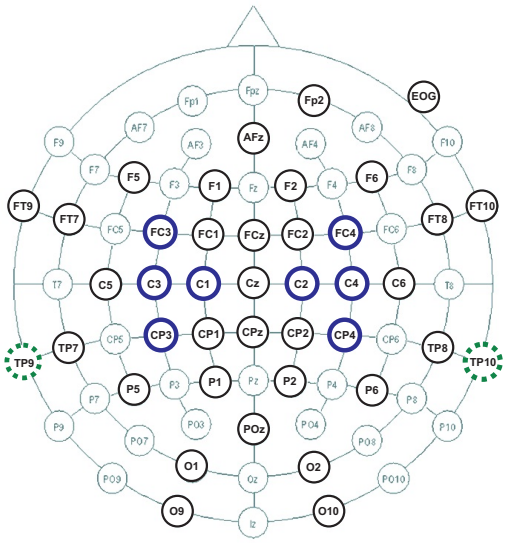


Figure 6.9: *The positions of 39 EEG electrodes within the standard 10/20 system used for data acquisition are marked by black circles. The two referencing electrodes are marked by green dotted circles. Eight electrodes over or close to the motor cortex are shown in bold blue circles (positions C1, C2, C3, C4, FC3, FC4, CP3 and CP4).*

6.3.1 Additional Data Preprocessing

In addition to the IFS data preprocessing of the EEG data introduced in section 5.1, one simple further preprocessing step was necessary. Before the data sets from several subjects were combined for cross-subject channel selection, an additional centering and linear scaling of the data was performed. This was done individually for each subject and trial in order to make corresponding AR coefficients comparable between trials of all subjects.

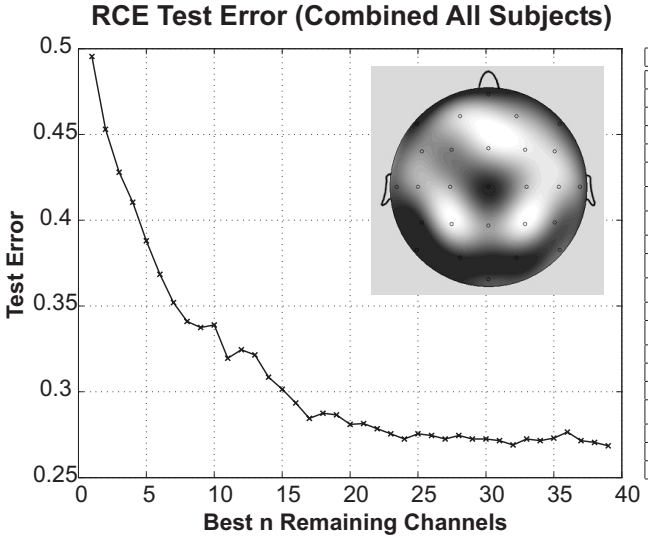
6.3.2 Generalization Error Estimation

For model selection purposes the generalization error of classifiers was estimated via 10-fold cross-validation. As the generalization error of IFS with RCE had to be estimated, a somewhat more elaborated procedure was used. An illustration of this procedure is given in figure 6.1. The whole data set is split up into 10 folds (*F1* to *F10*) as for usual cross-validation. In each fold F, the channel selection (*CS* in figure 6.1) is performed based on the train set of F only, leading to a specific ranking of the 39 EEG channels. For each fold F, 39 classifiers $C_F^h, h = 1, ..., 39$ are trained as follows: C_F^h is trained on the h best[2] channels, respectively, of the train set of F and tested on the corresponding channels of the test set of F. For each fold, this results in 39 test errors (E_F^1 to E_F^{39}). During the last step, the corresponding test errors are averaged over all folds. This leads to an estimate of the generalization error for every number of selected channels.

6.3.3 Channel Selection on Combined Data

The channel selection method RCE was applied on a training data set that had been combined from the five AR data sets.

The estimation of the average generalization error for all 39 stages of the channel selection process with RCE was carried out using linear SVMs as classifiers with parameter C previously determined by 10-fold cross-validation. Figure 6.10 shows the development of the estimated classification error for all 39 steps of the RCE on the combined data.



Ranking

Rank	Position	Rank	Position
1	CP2	21	POz
2	CP1	22	F6
3	FC2	23	AFz
4	FCz	24	TP8
5	F1	25	Cz
6	C4	26	P1
7	FC4	27	CP3
8	C2	28	P2
9	F2	29	FT9
10	C1	30	P5
11	C3	31	FT10
12	CPz	32	TP7
13	FC3	33	FT8
14	FT7	34	Fp2
15	FC1	35	F5
16	C6	36	O1
17	CP4	37	O9
18	P6	38	EOG
19	C5	39	O10
20	O2		

Figure 6.10: *RCE results for a combined data set of all 5 subjects. The graph shows a test error estimation for the n best channels. The error values were estimated by 10-fold cross-validation. The table on the right side shows the channel ranking performed on the combined data. Eight channels which are located closest to the motor cortex (see figure 6.9) are printed with gray background. The surface map visualizes this ranking: The 24 best ranked electrodes were mapped to gray scale values. Bright areas of the surface map correspond to relevant channels (according to RCE) whereas dark areas show less relevant electrodes.*

For this combined data set the test error was minimal (26.9%) when using data from 32 or more EEG channels but further reduction down to 24 channels increased the test error only marginally. Reducing the number of channels to fewer than the best 17 channels leads to a strong increase of the test error.

Throughout the ranking in the table of figure 6.10, artifact or task-irrelevant channels appear only in the last ranks (e.g. EOG, occipital channels, FT9, FT10 etc.). Direct comparison between figure 6.2 and figure 6.10 reveals that the curve in figure 6.2 shows smaller error rates: The performance of a classifier trained on the RCE channels of combined data is worse than the average performance of classifiers trained on the individual RCE channels of single subject data.

[2] In this context, *best* means according to the calculated ranking of that fold.

6.3.4 Transfer of Channel Selection Outcomes to New Subjects

In this section, it is analyzed whether RCE can find a subgroup of EEG channels that generalizes well to unseen subjects. In the following, different ways to obtain channel rankings are described, some of which are based on data of more than one subject. However, these rankings are always tested on the data of one subject only. Table 6.6 provides an overview over these ranking modes.

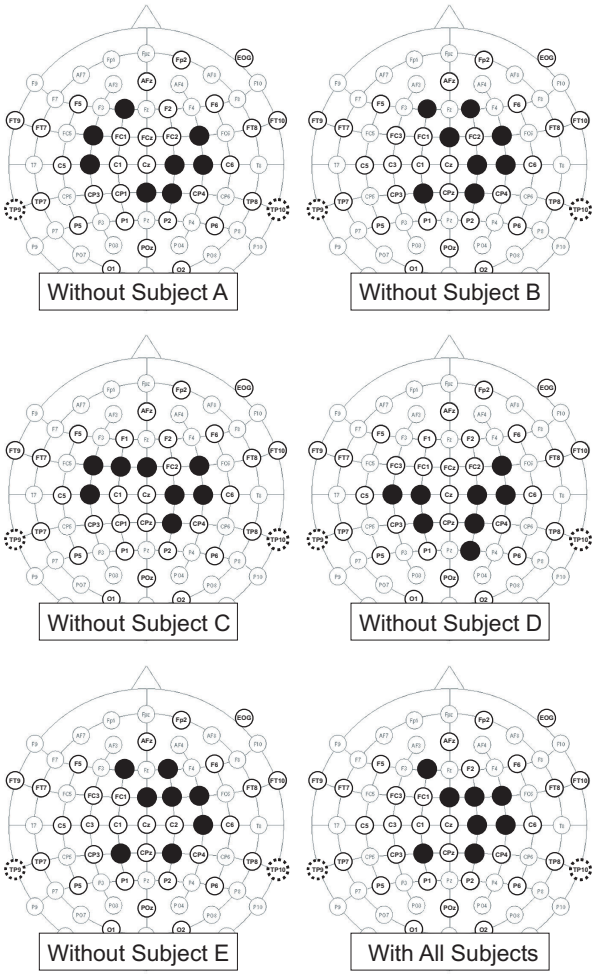


Figure 6.11: *The eight best channel positions across subjects. The database consists of data from 5 subjects. The channels were ranked 5 times using the channel selection method RCE, each time using the data of four subjects only. The electrode positions marked in bold are the 8 best ranked ones and are consistently located over or close to the motor cortex, although the method was not provided with prior knowledge about the motor imagery task. In the following this type of ranking will be referred to as* Best 8 (cross).

Cross Subject Modes

To implement the cross subject modes, the following process is iterated: Data of one subject is removed from the combined data base. On the remaining data the RCE method determines a channel ranking. To obtain test errors via 10-fold cross-validation on the data of the removed subject, this ranking is exploited in two different ways:

- *Best 8 (cross)*: The channel subset used for testing consists of the eight best-ranked channels. The resulting 8 best channels are plotted in figure 6.11.
- *Best n (cross)*: The channel subset used for testing consists of the n best ranked channels. The number n is chosen so that the expected cross-validation error on the four subjects is minimized. Note that this choice does not depend on the data of the fifth test subject.

Table 6.6: Ranking Modes Overview

Mode	**Ranking Method**	**Ranking based on**	**Description**
Motor 8	a priori knowledge	single subject	8 channels over or close to motor cortex
Random 8	(random)	single subject	random 8 channels
Best n (single)	RCE	single subject	n channels with highest rank that minimize CV error
Best 8 (single)	RCE	single subject	8 channels with highest rank
Best n (cross)	RCE	four subjects	n channels with highest rank that minimize CV error
Best 8 (cross)	RCE	four subjects	8 channels with highest rank

Explanation of the ranking modes used for the comparison shown in figure 6.12. The rankings were calculated on different types of data sets: on data from single subjects or on combined data sets (for cross-subject tests). Testing of the ranking modes was always performed on the data of one single subject.

As this process is repeated for every subject who was left out, the error values of the modes *Best 8 (cross)* and *Best n (cross)* can be averaged over five repetitions.

For Comparison: Single Subject Modes

For the fixed mental task of motor activity and imagery, the EEG literature suggests the channels CP3, CP4 and adjacent electrodes (e.g. [PNSL98]). This *a priori* knowledge can be used to define a generally good subgroup of EEG channels by the electrode set consisting of FC3, FC4, C1, C2, C3, C4, CP3, CP4. These positions are marked in boldface in figure 6.9. The corresponding test mode is referred to as *Motor 8*.

If no prior knowledge of a task and no channel selection methods were available, a random choice of channels would be the single solution. For comparison reasons the mode *Random 8* is included. Its test error is the average of ten repetitions of choosing eight random channels, optimizing the regularization parameter C and testing this random subset via 10-fold cross-validation on the data of one subject.

For the two modes *Best 8 (single)* and *Best n (single)* the RCE method was applied to the individual data of single subjects only. These modes used subgroups of the eight best channels and n best channels (see above) for calculating the test error via 10-fold cross-validation. It can be expected that the ranking for data from single subjects leads to more accurate classification results and can reveal task-related artifact channels [LSH$^+$04] that might not be present in data from other subjects.

Figure 6.12 shows the results for the 6 modes. The rightmost block contains an average taken over subjects for each of the modes. From the average results it can be observed that

- The mode *Motor 8* is not optimal: *Best n (single)* and *Best 8 (single)* perform much better[3].
- Mode *Best 8 (cross)* performs almost as well as the motor channel mode. Although it can be concluded that the RCE method fails to find an optimal channel subset, the results suggest that when transferring channel positions across subjects the expected performance is not much worse than the one using prior knowledge.
- The subset of 8 random channels performs surprisingly well. This finding suggests that the structure of the data can successfully be captured by the SVM even if only a few channels close to the motor cortex are (by chance) contained in the channel subset. However all other modes show better error estimates.
- The performance of *Best n (cross)* mode is comparable to the results of the *Best 8 (single)* mode (23%). Note however, that this is a comparison between a classifier using 8 channels only and one that used (on average) 27

[3] In figure 6.2 the choice of motor channels results in a lower classification error than the error from the RCE method. This is due to the fact that the regularization parameter C or ridge was not optimized for a specific ranking as was done in this analysis.

channels. The cross-validation error averaged over the five subjects is 26% for the choice of 27 random channels (not plotted in figure 6.12).

- The best performing mode is *Best n (single)*. On average it only uses n=14 channels and yields an error as low as 21.8%.

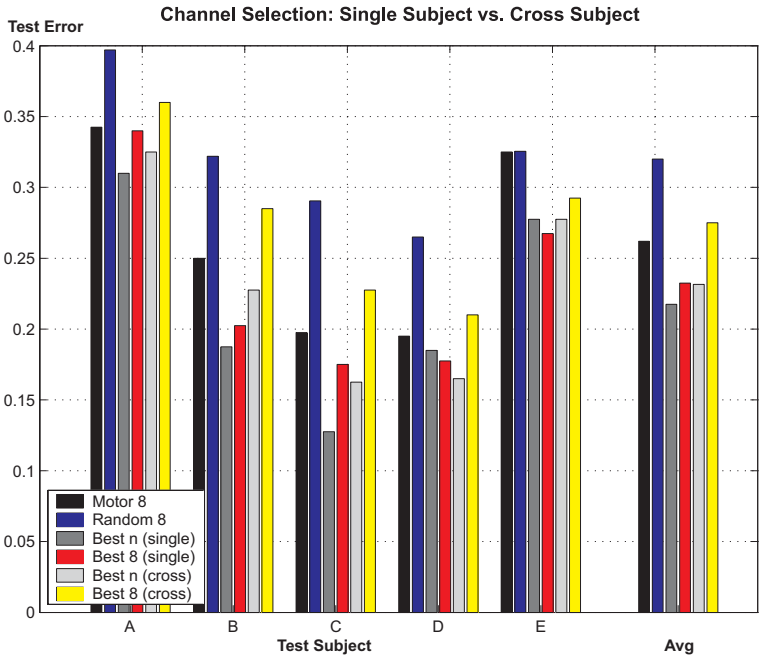


Figure 6.12: *Comparison of the test errors of six different ranking modes for single subjects (A to E) and the test errors of these modes averaged over the five subjects (Avg). For each mode and subject, the regularization parameter C was estimated separately. All test errors were obtained using 10-fold CV. The first mode* Motor 8 *tests the classification error for 8 channels over or close to the motor cortex, whereas* Random 8 *is based on 8 randomly chosen channels. The modes* Best n (single) *and* Best 8 (single) *both test channel sets, whose rankings were calculated based on the specific subject only. The modes* Best n (cross) *and* Best 8 (cross) *test channel sets, whose rankings were calculated based on all other subject's data, who but did not incorporate data from the test subject.*

Best results were obtained with RCE rankings from single subjects. A comparison reveals that they outperform motor rankings (including prior knowledge about the task) by about 5% absolute error. The transfer of RCE rankings from the data of multiple subjects to a new subject (cross mode) leads to a small decrease in average performance. Impressingly, this cross-mode is still about 2% better on average than the motor rankings, which were chosen by prior knowledge.

6.3.5 Discussion

The performance of three different types of rankings was analyzed for individual subjects: the ranking including channels over the motor cortex only, the ranking obtained by RCE from the data of that subject, and the ranking obtained by RCE from the data of the other four subjects.

It could be shown that RCE rankings on the combined data of *multiple* subjects are robust in the sense, that the rankings consistently are in agreement with the EEG-literature on motor imagery tasks.

The above results also show, that an individual channel ranking is slightly preferable over cross-subject rankings for the experimental paradigm investigated here, which is not surprising. However, the difference is surprisingly small, facing the numerous sources of possible inter-individual differences. The best cross-subject mode even outperforms a choice of channels selected by prior knowledge.

Even though the statistical basis is small with only 5 subjects, the results encourage the application of cross-subject channel selection in BCI experiments under certain conditions:

1. when standard motor paradigms are to be used with AR features for many subjects and EEG,

2. when for a number (e.g. 5 to 10, to err on the side of caution) of subjects IFS channel selections have already been performed, and

3. when a speed-up of experiment duration is needed for the remaining subjects.

So far it is of course unclear, whether these findings do generalize for different paradigms or for feature pre-processings other than AR.

6.4 Performance of the Implementation

For the feedback training of a user, it is advantageous to provide feedback as quick as possible in order to maximize the effect of reinforcement learning. In case of the proposed signal processing concept IFS the feedback could be given after the class label of a trial was determined by the Matlab software. The algorithms were chosen from the very beginning under the consideration of runtime. For example, both implementation variants for the estimation of the AR model were tested exhaustively. The C++ implementation was approximately three times faster than the Matlab implementation. This speedup can be a clear advantage if a BCI system has to be set up with very high sampling rates and a great numbers of recording channels. Using the C++ implementation, a recorded trial could be processed very quickly: even for the computationally most demanding experiments (150 MEG channels and 625 Hz sampling frequency), the class label codes were generated in the online experiment (see Section 6.5) within less than 100 ms of delay.

The computationally very demanding double cross-validation schemes that were used in offline analyses for parameter studies (see Algorithm 2 in Section 6.1.5) were performed in parallel on a Linux cluster of about 10 dual-core Pentium 4 machines and took about 1 to 20 hours for each type of analysis, depending on the number of inner and outer cross-validation folds.

6.5 Online Classification Performance

The previous sections reported results of offline analyses, and online feedback was not presented to the subjects. In this section, the transfer of channel subsets and trained classifiers from offline to online BCI experiments will be investigated.

6.5.1 MEG Signals

For the MEG experiment, it is described how four of the subjects used a trained BCI system. The system has been trained on the individual data of each subject according to the IFS setting. The training was performed with data from the calibration recording. It included the selection of a channel subset with RCE and the training of a linear SVM. In the feedback phase of the online experiment, the task was to write a short name by mere thought control.

The part of the experiment with feedback was described as stage 2 and stage 3 in section 5.2. It was carried out directly after the machine learning phase (stage 1). Since the subjects had to wait while their data was analyzed, it was not possible to estimate all parameters as described for the extensive offline analyses. Instead, parameter settings were used, that potentially were slightly suboptimal.

For every subject 200 training points had been collected in the first stage. The detrended data were preprocessed using an AR model of order 6. Similar to the offline analysis, a good ridge was selected and the channels were ranked by RCE using this ridge. As a single unified ranking and not an error estimate was needed, the expensive outer (50) folds of the double-CV scheme were not performed. The data set was restricted to the 20 best-ranked channels in order to avoid unnecessary loss of classification performance in case a too small number of channels would be chosen. Once more a good ridge value was estimated and an SVM was trained on the restricted data.

During stage 2, every trial was processed and classified online by the IFS architecture. The subjects were asked to perform the same task as before, but received some feedback: After every trial that was classified correctly by the SVM, a “Smiley” symbol was displayed to positively reinforce the subjects. Depending on their performance, the subjects completed two to four blocks of fifty trials. The subjects $\mathcal{B}$, $\mathcal{D}$, $\mathcal{F}$, $\mathcal{H}$ and $\mathcal{I}$ obtained an accuracy higher than 70% in their last block. On average, they showed a slightly decreased performance during the online feedback compared to the offline performance estimation. This difference can be attributed to the changed brain characteristics of the subjects due to the feedback. All subjects reported that the feedback trials felt more exciting for them than the trials during phase 1 without feedback.

For these subjects, the data sets collected during stage 1 and stage 2 were combined and again an SVM was trained on the combined data as described above. The SVM was then used in stage 3 of the experiment, the online spelling, which also included feedback.

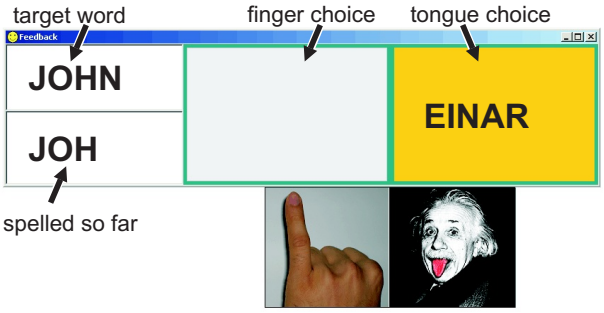


Figure 6.13: *Spelling interface during stage 3 of the MEG experiment. The screen shot demonstrates, how the subject spells the word* JOHN *(shown at the upper left box) by using motor imagery. The lower left box contains the letters* JOH *spelled so far. Every 5 seconds, the subject can produce a movement imagination to maneuver within the binary spelling tree by either choosing a branch and descending the tree towards its leaves (single letters) or not choosing it but instead changing to its sibling branch. In addition to letters, digits and punctuation marks, the speller tree regularly offered correcting steps and the possibility to delete spelled letters. In the present situation the next letter to write is among the letters of the right box (*N *within the branch* EINAR *at the present state). The subject can chose this box by imagining tongue movements (indicated by the thumbnail pictures added underneath the speller as reminders). In this case a subset of letters* EINAR *would appear next. Respectively, an imagined finger movements would communicate that the intended letter is not among the displayed ones and the empty left box would be chosen. In this case the speller changes to the sibling branch and a different set of letters will appear next. After a few iterations only one letter remains. If it is chosen by the subject, the letter will be appended to the word spelled so far. The computer system identifying the imagined movements is a combination of the feature selection technique RCE and regularized linear SVMs.*

The task for each of the remaining five subjects was to spell a short name by making a number of binary decisions via motor imagery. At the beginning of phase 3 the spelling interface displayed one half of the letters of the alphabet (including some special characters). If the first letter to be spelled was among the displayed ones, the subject had to imagine a tongue movement. To communicate that the designated letter currently was not displayed, the subject imagined a finger movement (see Figure 6.13). To help the subjects concentrate on the imagination task, the box of correct choice was highlighted. This spelling variant is sometimes referred to as “copy spelling” and useful for the training of subjects before proceeding to unrestricted free spelling. In the next iteration the selected subset of the alphabet was split in two parts again, and one of them was displayed to the subject. The subject made another choice by motor imagery and so on. On the last stage of this process a single letter had been determined. The letter had to be confirmed by a last single decision and was then displayed on the left part of the screen. Then the procedure started over again to allow the selection of further letters. The spelling algorithm allows also for the deletion of an already selected letter by choosing a delete symbol. Furthermore, the splitting algorithm was optimized so that it reflected the letter frequencies of the german language. For more details about the speller refer to [BGH+99].

As a result from the feedback experiments with MEG, four out of five subjects succeeded in spelling a short name. The fifth subject aborted the experiment after successfully spelling the first letter of the desired name. He reported decreasing classification performance and concentration problems. The names spelled by the other four subjects had 4.25 letters on average and their spelling took about 5 to 20 minutes.

6.5.2 ECoG Signals

Subject II of the ECoG experiments (see Section 5.3.1) could be recruited for a second session one week after the initial recordings. In that second session, some more training data of that day was collected (see Section 5.3). Based on the combined data set, a good ECoG channel subset was determined and a classifier was trained. Then the subject performed two feedback runs. During the first run of 50 trials, the subject was given feedback at the end of each trial. Here the subject only missed one out of 50 trials. Encouraged by this result, a second feedback run was started. In this run, the subject was asked to copy-spell the word *ANGELO*. For spelling, the same environment was used as the one described for MEG in Section 6.5.1. Again, the subject used binary decisions from motor imagery of the left little finger and the tongue.

Figure 6.14 shows screenshots of the experiment. The subject was able to accomplish the spelling task. If the feedback control had been free of errors, then the spelling of one letter would have taken 5 steps in the binary speller

Figure 6.14: *Screenshots from the ECoG online experiment. The left screenshot shows the patient sitting on his bed and looking at the feedback screen. The right screenshot shows the feedback screen, while the patient writes the name ANGELO in a copy-spelling task. The snapshot is taken during a binary decision step for the selection of the next letter L.*

tree on average. Compared to the first feedback run (without spelling), the second run showed more misclassified trails. To correct these trials, the subject had to perform more steps in the speller tree. In total, the spelling of the word took less than 10 minutes. The rate of correctly classified decisions was between 80% and 90%.

6.5.3 Discussion

The reported results from the spelling experiments show the performance in live BCI experiments. The necessary calculations for the channel selection and classifier training were performed within 10 minutes. During the feedback runs, the trained classifiers responed in realtime with a feedback value immediately after each trial. For the experimenters, who supervised the training of the BCI systems, the IFS concept simplified the process enormously. Instead of hand-choosing suitable channels and defining suitable frequency ranges in a spectrogram, the experimenters can use IFS to perform the necessary steps automatically, which even include an error estimate of the solutions found.

It can be concluded that the automated BCI processing including the IFS is very desirable for the use in future online BCI experiments.

7 Summary and Outlook

The thesis has introduced the difficult problem of feature selection in the context of Brain-Computer Interfaces (BCI) - a field, where a combination of low signal-to-noise ratio, high-dimensional data, small number of examples, new brain signal recording techniques, new experimental paradigms and the constraints of working with handicapped patients makes machine learning problems extremely challenging. Development must be guided by the recognition that a BCI system must be adapted individually to a user and that reduced EEG channel setups are in great demand.

For BCI classification tasks in general, it has been emphasized that individual learning is not only needed on the level of classifier training, but also on the level of feature selection. This has led to the new development of an individual feature selection (IFS) concept. Besides a robust workflow that is applicable during an experimental session, different channel selection approaches (one filter approach using Fisher score and two wrapper approaches based on SVMs) have been introduced as possible steps in IFS. In a direct comparison, the BCI-specific problem of channel selection and reduction was tackled on the basis of a series of EEG-BCI experiments conducted by the author.

Three competing methods were thoroughly evaluated in terms of offline classification accuracy and reduction of channel numbers. The winning method Recursive Channel Elimination (RCE), which was newly introduced by the author, went through a number of further evaluations:

- First, the method was applied to different types of brain signals (MEG and ECoG), collected by the author in two other series of BCI experiments with these emerging new recording techniques. On this data, the proposed method showed very good offline performance as well, e.g. a reduction of 95% of the channels and very low error rates.
- Second, a comparison of different stop-criterions for channel reduction was conducted in a double cross-validation evaluation, resulting in a simple but useful rule-of-thumb.
- Third, the channel subsets selected by the method were checked for plausibility during a well-known motor paradigm. Without using any prior knowledge about the task, IFS not only delivered reasonable, well-interpretable and easy-to-visualize results, but provided a tool for discovering objectionable non-neuronal artifacts in the data. Furthermore, its classification results were comparable to those channel sets which had been selected by prior knowledge.
- Fourth, the new concept was checked regarding different notions of transferability. One of them was the ability to transfer feature subsets to new user - in this setting, IFS with RCE showed very good generalization and could even outperform solutions based on expert knowledge.
- Last, but not least, the new concept was tested by the author in a series of online BCI experiments with feedback in a spelling paradigm and with MEG and ECoG data. This was the first time, that a BCI has been coupled with these brain signal sources in a feedback setting. For these new recording technologies, the implementation of the new concept was good for a reduction of 84% of the channels (in MEG) and proved the expected ease of operation and high performance. In the experiment series with MEG, four subjects started from scratch (no user training was involved) and ended spelling a short name within a single BCI session. For ECoG, this could be

It can be concluded that IFS based on RCE is of great value for BCI, particularly in exploring new experimental paradigms and new recording techniques where no prior knowledge is available, and also for every-day work with patients, where reduced channel sets are crucial for the acceptance of BCI.

Possible Extensions

The following possible future research topics are tapped by the results of this theses and could be studied in follow-up research:

If known BCI paradigms are not applicable for a patient due to his or her condition, then new experimental task settings have to be found. This is also the case if the known two-class paradigms are to be expanded to multi-class paradigms. A lack of prior knowledge about useful electrode sites or other features could be overcome by using the IFS concept.

The generalization ability of IFS has been tested in this thesis in many respects. However, the degree of reproducibility of ranking results under minor or major changes in the data can still be investigated. Here it will be interesting to see,

how well IFS still works on much worse signal-to-noise conditions, which could be investigated offline, for example, by introducing additional noise, leaving out samples, or by training on unbalanced classes.

Spatial filtering methods like the Common Spatial Patterns (CSP) method, Independent Component Analysis (ICA) or variants thereof have recently gained more interest in the field of BCI [LBCM05, KDB+04, DBK+06b, TDN+06, KSBM07, DBK+06a, FHLS06, FHS06]. They can concentrate the information spread over neighboring recording channels, but these methods still require a large number of sensors, and tend to deliver useful EEG components and undesired artifacts or background EEG that is not task related. The choice of a good set of spatial filters is often not trivial and is usually performed by hand. The proposed RCE method in IFS could be applied on the spatially filtered channels in order to automatize the selection of useful spatially filtered channels.

A Nomenclature

Mathematical Notation

General

$\mathbb{R}^d$	d-dimensional real-valued vector space
r^2	Correlation coefficient
l_0	l-Zero Norm
l_1	l-1 Norm
l_2	l-2 Norm (Euclidian Norm)
$O(\cdot)$	O-Notation for run time estimates of algorithms

Time Series

x	Time series signal
$x[n]$	Index of a sample point within signal x
r_{xx}	Autocorrelation of time series x

SVM

n	Number of training vectors (trials)
d	Data dimension (e.g. number of channels times the number of features per channel)
x	Data vector $(x \in \mathbb{R}^d)$
$X = (x_1, ..., x_n)$	(Training) data set $(X \in \mathbb{R}^{n \times d})$
y	Class label of one trial resp. of one data vector e.g. $(Y \in \{-1, 1\})$ for two-class problems
$Y = (y_1, ..., y_n)$	Class labels e.g. $Y \in \{-1, 1\}^n$ for two-class problems
ξ	Slack Variable
γ	Margin (distance of support vectors to the separating hyperplane)
w	Real valued weight vector defining the normal of a hyperplane
b	Real valued scalar defining the offset of a hyperplane
C	Real valued regularization parameter for *C formulation*
r	Real valued regularization parameter for *ridge formulation*

Fisher Score, RCE and $l0$-Optimization

R_j	Real valued score of feature j
$\mu(T)$	Mean value of set T
l	Number of channels showing minimal CV-error suggested by RCE
#	Abbrev. for "number of", e.g. #channels
r	Ridge value obtained for the full set of channels
r_{red}	Ridge value obtained for a reduced set of channels

Error Estimation

F_i	Data partition that belongs to the i-th fold of a cross validation procedure
C^h	Classifier trained with data from h best channels
C_F^h	Classifier trained with data from h best channels during fold F
E^h	Generalization error estimate for C^h

Abbreviations

ALS	Amyotrophic Lateral Sclerosis
ANN	Artificial Neural Network
AP	Action Potential
AR	Autoregressive
ARMA	Autoregressive Moving Average
BCI	Brain Computer Interface
BMI	Brain Machine Interface
BOLD	Blood Oxygenation Level
BSS	Blind Signal Separation
CV	Cross Validation
ECoG	Electrocorticogram
EEG	Electroencephalogram
EOG	Electrooculogram
EPSP	Excitatory Postsynaptic Potential
ERD	Event Related Desynchronization
ERP	Event Related Potential
ERS	Event Related Synchronization
FFT	Fast Fourier Transform
fMRI	Functional Magnetic Resonance Imaging
HMI	Human Machine Interface
Hz	Hertz
ICA	Independent Component Analysis
K^+	Ion of the chemical element potassium
K	Kelvin
MA	Moving Average
MEG	Magnetoencephalogram
Na^+	Ion of the chemical element sodium
P300	Significant brain signal (duration approx. 300ms) after excitation
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PSD	Power Spectral Density
RCE	Recursive Channel Elimination
RFE	Recursive Feature Elimination
SCP	Slow Cortical Potential
SNR	Signal to Noise Ratio
SQUID	Superconducting Quantum Interference Device
SSER	Steady State Evoked Response
STE	Standard Error
SVM	Support Vector Machine
T, fT	Tesla, femto Tesla
V, μV	Volt, Microvolt

B Algorithms

Algorithm 3 *Error estimation scheme for channel subset **All** using double CV.*

Require: preprocessed ECoG data of one subject
1: **for** (cntMainFolds $= 1$ to 50) **do**
2: split data randomly: 80% training set, 20% test set
3: with training set do: {all channels}
4: 10-fold CV: find *good* ridge r
5: train SVM S using r
6: with test set do: {all channels}
7: test S
8: save test error
9: **end for**
Output: mean error, standard deviation

Algorithm 4 *Error estimation scheme for channel subset **RCE_2** using double CV.*

Require: preprocessed ECoG data of one subject
1: **for** (cntMainFolds $= 1$ to 50) **do**
2: split data randomly: 80% training set, 20% test set
3: with training set do: {all channels}
4: 10-fold CV: find *good* ridge r
5: rank channels with RCE using r
6: $b := 2$
7: reduce training set to b best ranked channels
8: with training set do: {reduced channels}
9: 10-fold CV: find *good* ridge r_red
10: train SVM S using ridge r_red
11: reduce test set to b best ranked channels
12: with test set do: {reduced channels}
13: test S
14: save test error and number of good channels b
15: **end for**
Output: mean error, standard deviation

Algorithm 5 *Error estimation scheme for channel subset* ***Random_2*** *using double CV.*

Require: preprocessed ECoG data of one subject
1: **for** (cntMainFolds $= 1$ to 50) **do**
2: split data randomly: 80% training set, 20% test set
3: with training set do: {all channels}
4: 10-fold CV: find *good* ridge r
5: determine 2 random channels c_1,c_2
6: reduce training set to channels c_1,c_2
7: with training set do: {reduced channels}
8: 10-fold CV: find *good* ridge r_red
9: train SVM S using ridge r_red
10: reduce test set to channels c_1,c_2
11: with test set do: {reduced channels}
12: test S
13: save test error and number of good channels b
14: **end for**
Output: mean error, variance

Bibliography

[Ame91] American Electroencephalographic Society, *Guidelines for standard electrode position nomenclature*, Journal of Clinical Neurophysiology **8** (1991), 200–202.

[BCM02] Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller, *Classifying single trial EEG: Towards brain computer interfacing*, Advances in Neural Information Processing Systems (NIPS) (T.G. Diettrich, S. Becker, and Z. Ghahramani, eds.), vol. 14, MIT Press, 2002, pp. 157–164.

[BDK+06] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Michael Schröder, John Williamson, Roderick Murray-Smith, and Klaus-Robert Müller, *The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell*, Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006, Verlag der Technischen Universität Graz, 2006, pp. 108–109.

[BDK+07] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Volker Kunzmann, Florian Losch, Gabriel Curio, and Klaus-Robert Müller, *The Berlin Brain-Computer Interface: Machine-learning based detection of user specific brain states*, Toward Brain-Computer Interfacing (Guido Dornhege, Jose del R. Millán, Thilo Hinterberger, Dennis McFarland, and Klaus-Robert Müller, eds.), MIT press, Cambridge, MA, 2007, in press.

[Ber29] Hans Berger, *Über das Elektrenkephalogramm des Menschen*, Archiv für Psychiatrie und Nervenkrankheiten **87** (1929), 527–570.

[BFOS84] L. Breiman, J. Friedman, J. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, 1984.

[BGH+99] Niels Birbaumer, N. Ghanayim, Thilo Hinterberger, I. Iversen, B. Kotchoubey, Andrea Kübler, J. Perelmouter, E. Taub, and H. Flor, *A spelling device for the paralysed*, Nature **398** (1999), 297–298.

[BHKN03] Niels Birbaumer, Thilo Hinterberger, Andrea Kübler, and Nicola Neumann, *The thought-translation device (TTD): Neurobehavioural mechanisms and clinical outcome*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **11** (2003), no. 2, 120–123.

[Bis95] C.M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[BKG+00] Niels Birbaumer, Andrea Kübler, N. Ghanayim, Thilo Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor, *The Thought Translation Device (TTD) for completely paralyzed patients*, IEEE Transactions on Rehabilitation Engineering **8** (2000), no. 2, 190–193.

[BML01] S. Baillet, J.C. Mosher, and R.M. Leahy, *Electromagnetic brain mapping*, IEEE Signal Processing Magazine **18** (2001), no. 6, 14–30.

[BR92] A. L. Blum and R. L. Rivest, *Training a 3-node neual network is np-complete*, Neural Networks **5** (1992), no. 1, 117–127.

[CCY+90] D. Cohen, B.N. Cuffin, K. Yunokuchi, R. Maniewski, C. Purcell, G.R. Cosgrove, J. Ives, J.G. Kennedy, and D. L. Schomer, *MEG versus EEG localization test using implanted sources in the human brain*, Annals of Neurology **28** (1990), 811–817.

[CH03] D. Cohen and E. Halgren, *Magnetoencephalography (neuromagnetism)*, Encyclopedia of Neuroscience (G. Adelman, ed.), Elsevier Science, Amsterdam, 2003.

[CLC+03] J. M. Carmena, M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. Nicolelis, *Learning to control a brain-machine interface for reaching and grasping by primates*, PLoS Biol. **E42** (2003).

[Coh68] D. Cohen, *Magnetoencephalography, evidence of magnetic fields produced by alpha-rhythm currents*, Science **161** (1968), 784–786.

[CV95] C. Cortes and V.N. Vapnik, *Support vector networks*, Machine Learning **20** (1995), 273–297.

[DBC03] Guido Dornhege, Benjamin Blankertz, and Gabriel Curio, *Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials*, Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering. Capri 2003, 2003, pp. 591–594.

[DBCM04] Guido Dornhege, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller, *Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms*, IEEE Transactions on Biomedical Engineering **51** (2004), no. 6, 993–1002.

[DBK+06a] Guido Dornhege, Benjamin Blankertz, Matthias Krauledat, Florian Losch, Gabriel Curio, and Klaus-Robert Müller, *Combined optimization of spatial and temporal filters for improving brain-computer interfacing*, IEEE Transactions on Biomedical Engineering **53** (2006), no. 11, 2274–2281.

[DBK+06b] ———, *Optimizing spatio-temporal filters for improving brain-computer interfacing*, Advances in Neural Inf. Proc. Systems (NIPS 05) (Cambridge, MA), vol. 18, MIT Press, 2006, pp. 315–322.

[DS02] Dennis Decoste and Bernhard Schölkopf, *Training invariant support vector machines*, Machine Learning **46** (2002), no. 1-3, 161–190.

[DSW00] E. Donchin, K. M. Spencer, and R. Wijesinghe, *The mental prosthesis: Assessing the speed of a P300-based brain computer interface*, IEEE Transactions on Rehabilitation Engineering **8** (2000), no. 2, 174–179.

[EAW93] Roger Eckert, Raimund Apfelbach, and Elke Weiler, *Tierphysiologie*, Thieme, 1993.

[Eng93] J. Engel, *Surgical treatment of the epilepsies*, 2 ed., ch. Presurgical evaluation protocols, pp. 740–742, Raven Press Ltd., New York, 1993.

[FHLS06] J. Farquhar, N. J. Hill, T. N. Lal, and B. Schölkopf, *Regularised CSP for sensor selection in BCI*, Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006 (Graz), Verlag der Technischen Universität Graz, 2006, pp. 14–15.

[FHS06] J. Farquhar, N. J. Hill, and B. Schölkopf, *Optimizing spatial filters for BCI: Margin- and evidence-maximization approaches*, 2006, p. 1.

[FLA+06] R. Ferenets, T. Lipping, A. Anier, V. Jantti, S. Melto, and S. Hovilehto, *Comparison of entropy and complexity measures for the assessment of depth of sedation*, IEEE Transactions on Biomedical Engineering **53** (2006), no. 6, 1067–1077.

[GHEV04] G. Garcia, U. Hoffmann, T. Ebrahimi, and J. Vesin, *Direct brain-computer communication through EEG signals*, Tech. report, EPFL, 2004.

[GHLP04] Bernhard Graimann, Jane E. Huggins, Simon P. Levine, and Gert Pfurtscheller, *Toward a direct brain interface based on human subdural recordings and wavelet-packet analysis*, IEEE Transactions on Biomedical Engineering **51** (2004), no. 6, 954–962.

[Goh97] K. Gohlenhofen, *Physiologie: Lehrbuch, Kompendium, Fragen und Antworten*, Verlag Urban & Schwarzenberg, 1997, ISBN 3-541-16351-8.

[Gro87] Stephen Grossberg, *Competitive learning: From interactive activation to adaptive resonance*, Cognitive Science **11** (1987), 23–63.

[GWBV00] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, *Gene selection for cancer classification using support vector machines*, Journal of Machine Learning Research (2000), no. 3, 1439–1461.

[Hay95] S.S. Haykin, *Adaptive filter theory*, Prentice Hall, 1995.

[HKK+00] Thilo Hinterberger, Boris Kotchoubey, Jochen Kaiser, Andrea Kübler, Nicola Neumann, Juri Perelmouter, Ute Strehl, and Niels Birbaumer, *Anwendungen der Selbstkontrolle langsamer kortikaler Potentiale*, Verhaltenstherapie **10** (2000), 219–227.

[HLS+06] N. J. Hill, T. N. Lal, M. Schröder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. E. Elger, B. Schölkopf, A. Kübler, and N. Birbaumer, *Classifying EEG and ECOG signals without subject training for fast BCI implementation: Comparison of non-paralysed and completely paralysed subjects*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **14** (2006), no. 2, 183–186.

[HLT+07] N. J. Hill, T. N. Lal, M. Tangermann, T. Hinterberger, G. Widman, C. E. Elger, B. Schölkopf, and N. Birbaumer, *Classifying event-related desynchronization in EEG, ECoG and MEG signals*, Toward Brain-Computer Interfacing (Guido Dornhege, Jose del R. Millán, Thilo Hinterberger, Dennis McFarland, and Klaus-Robert Müller, eds.), MIT Press, Cambridge, Mass., 2007, pp. 235–260.

[HNK+07] Thilo Hinterberger, Femke Nijboer, Andrea Kübler, Tamara Matuz, Adrian Furdea, Ursula Mochty, Miguel Jordan, Thomas Navin Lal, N. Jeremy Hill, Jürgen Mellinger, Michael Bensch, Michael Tangermann, Guido Widman, Christian E. Elger, Wolfgang Rosenstiel, Bernhard Schölkopf, and Niels Birbaumer, *Brain-computer interfaces for communication in paralysis: A clinical experimental approach*, Toward Brain-Computer Interfacing (G. Dornhege, J. del R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, eds.), MIT Press, Cambridge, Mass., 2007, pp. 43–64.

[HSF+06] L.R. Hochberg, M.D. Serruya, G.M. Friehs, J.A. Mukand, M. Saleh, A.H. Caplan, A. Branner, D. Chen, R.D. Penn, and J.P. Donoghue, *Neuronal ensemble control of prosthetic devices by a human with tetraplegia*, Nature **442** (2006), no. 7099, 164–171.

[JC00] D. D. Jensen and P. R. Cohen, *Multiple comparisions in induction algorithms*, Machine Learning **38** (2000), no. 3, 309–338.

[JP49] H. Jasper and W. Penfield, *Electrocorticograms in man: Effects of voluntary movement upon the electrical activity of the precentral gyrus*, Arch. Psychiat. Nervenkr. **183** (1949), 163–174.

[Kay88] Steven M. Kay, *Modern spectral estimation*, Prenctice Hall, 1988.

[KBCM07] Roman Krepki, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller, *The Berlin Brain-Computer Interface (BBCI): towards a new communication channel for online control in gaming applications*, Journal of Multimedia Tools and Applications (2007).

[KDB+04] Matthias Krauledat, Guido Dornhege, Benjamin Blankertz, Florian Losch, Gabriel Curio, and Klaus-Robert Müller, *Improving speed and accuracy of brain-computer interfaces using readiness potential features*, Proceedings of the 26th Annual International Conference IEEE EMBS on Biomedicine, San Francisco, 2004.

[KDB+07] Jens Kohlmorgen, Guido Dornhege, Mikio Braun, Benjamin Blankertz, Klaus-Robert Müller, Gabriel Curio, Konrad Hagemann, Andreas Bruns, Michael Schrauf, and Wilhelm Kincses, *Improving human performance in a real operating environment through real-time mental workload detection*, Toward Brain-Computer Interfacing (Guido Dornhege, Jose del R. Millán, Thilo Hinterberger, Dennis McFarland, and Klaus-Robert Müller, eds.), MIT press, Cambridge, MA, 2007, in press.

[KNK+01] Andrea Kübler, Nicola Neumann, Jochen Kaiser, Boris Kotchoubey, Thilo Hinterberger, and Niels Birbaumer, *Brain-computer communication: Self-regulation of slow cortical potentials for verbal communication*, Arch Phys Med Rehabil **82** (2001), 1533–1539.

[KNM+05] Andrea Kübler, Femke Nijboer, Jürgen Mellinger, Theresa M. Vaughan, Hannelore Pawelzik, Gerwin Schalk, Dennis C. McFarland, Niels Birbaumer, and Jonathan R. Wolpaw, *Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface*, Neurology **64** (2005), 1775–1777.

[KNW+04] Andrea Kübler, Nicola Neumann, Barbara Wilhelm, Thilo Hinterberger, and Niels Birbaumer, *Predictability of brain-computer communication*, Journal of Psychophysiology **18** (2004), 121–129.

[KSBM07] Matthias Krauledat, Michael Schröder, Benjamin Blankertz, and Klaus-Robert Müller, *Reducing calibration time for brain-computer interfaces: A clustering approach*, Advances in Neural Inf. Proc. Systems (NIPS 06), vol. 19, MIT press, 2007, accepted.

[KSU+01] Boris Kotchoubey, Ute Strehl, C. Uhlmann, Susanne Holzapfel, M. König, W. Fröscher, V. Blankenhorn, and Niels Birbaumer, *Modification of slow cortical potentials in patients with refractory epilepsy: A controlled outcome study*, Epilepsia **42** (2001), no. 3, 406–416.

[KVP05] Julian Kronegg, Svyatoslav Voloshynovskiy, and Thierry Pun, *Analysis of bit-rate definitions for brain-computer interfaces*, Proceedings of the Int. Conf. on Human-computer Interaction (HCI), 2005.

[LBCM05] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller, *Spatio-spectral filters for improved classification of single trial EEG*, IEEE Trans. Biomed. Eng. **52** (2005), no. 9, 1541–1548.

[LHW+05] Thomas N. Lal, Thilo Hinterberger, Guido Widman, Michael Schröder, Jeremy Hill, Wolfgang Rosenstiel, Christian E. Elger, Bernhard Schölkopf, and Niels Birbaumer, *Methods towards invasive human brain computer interfaces*, Advances in Neural Information Processing Systems (NIPS) 17 (Cambridge, MA) (Lawrence K. Saul, Yair Weiss, and Léon Bottou, eds.), MIT Press, 2005, pp. 737–744.

[LL95] Joel F. Lubar and Judith O. Lubar, *Neurological basis and neurofeedback treatment of ADHD*, Journal of Neurotherapy (1995).

[LLD76] D. Lehmann, W. Lang, and P. Debruyne, *Controlled EEG alpha feedback training in normals and headache patients*, Archiv für Psychiatrische Nervenkrankheiten **221** (1976), 331–343.

[LSH+04] Thomas N. Lal, Michael Schröder, Thilo Hinterberger, Jason Weston, Martin Bogdan, Niels Birbaumer, and Bernhard Schölkopf, *Support vector channel selection in BCI*, IEEE Transactions Biomedical Engineering **51** (2004), no. 6, 1003–1010.

[LSH+05] Thomas Navin Lal, Michael Schröder, N. Jeremy Hill, Hubert Preissl, Thilo Hinterberger, Juergen Mellinger, Martin Bogdan, Wolfgang Rosenstiel, Niels Birbaumer, and Bernhard Schölkopf, *A brain computer interface with online feedback based on magnetoencephalography*, Proceedings of the 22nd International Conference on Machine Learning (ICML) (Bonn), 2005.

[LSK+05] Robert Leeb, Reinhold Scherer, Claudia Keinrath, Christoph Guger, and Gert Pfurtscheller, *Exploring virtual environments with an EEG-based BCI through motor imagery*, Biomedizinische Technik **52** (2005).

[LSW+04] Eric C. Leuthardt, Gerwin Schalk, Jonathan R. Wolpaw, Jeffrey G. Ojemann, and Daniel W. Moran, *A brain-computer interface using electrocorticographic signals in humans*, Journal of Neural Engineering **1** (2004), 63–71.

[MKD+04] K.R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, *Machine learning techniques for brain-computer interfaces*, Biomed. Tech. **49** (2004), 11–22.

[MPSPR05] G.R. Müller-Putz, R. Scherer, G. Pfurtscheller, and R. Rupp, *EEG-based neuroprosthesis control: A step towards clinical practice*, Neuroscience Letters **382** (2005), 169–174.

[MSK+06] Klaus-Robert Müller, Michael Schröder, Matthias Krauledat, Guido Dornhege, Gabriel Curio, and Benjamin Blankertz, *Das Berliner Brain-Computer Interface*, Künstliche Intelligenz KI **4** (2006), 61–61.

[MTD+07] Klaus-Robert Müller, Michael Tangermann, Guido Dornhege, Matthias Krauledat, Gabriel Curio, and Benjamin Blankertz, *Machine learning for real-time single-trial EEG analysis: From brain-computer interfacing to mental state monitoring*, Journal of Neuroscience Methods (2007), in review.

[NWB04] Christian Neudert, Maria Wasner, and Gian Domenico Borasio, *Individual quality of life is not correlated with health-related quality of life or physical funcions in patients with amyotrophic lateral sclerosis*, Journal of Palliative Medicine (2004), 551–557.

[PB00] J. Perelmouter and N. Birbaumer, *A binary spelling interface with random errors*, Transactions on Rehabilitation Engineering **8** (2000), no. 2, 227–232.

[PdS99] G. Pfurtscheller and F.H. Lopes da Silva, *Event-related EEG/MEG synchronization and desynchronization: basic principles*, Clinical Neurophysiology **110** (1999), no. 11, 1842–1857.

[PNK94] P. Pudil, J. Novovicova, and J. Kittler, *Floating search methods in feature-selection*, PRL **15** (1994), no. 11, 1119–1125.

[PNM+03] Gert Pfurtscheller, Christa Neuper, Gernot R. Müller, Bernhard Obermaier, Gunter Krausz, Alois Schlögl, Reinhold Scherer, Bernhard Graimann, Claudia Keinrath, Dimitris Skliris, M. Wörtz, G. Supp, and C. Schrank, *Graz-BCI: state of the art and clinical applications*, IEEE Trans. Neural Systems and Rehabilitation Engineering **11** (2003), no. 2, 177–180.

[PNSL98] G. Pfurtscheller, C. Neuper, A. Schlögl, and K. Lugger, *Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters*, IEEE Transactions on Rehabilitation Engineering **6** (1998), 316–325.

[PPF94] M. Pregenzer, G. Pfurtscheller, and D. Flotzinger, *Selection of electrode positions for an EEG-based brain computer interface (BCI)*, Biomed. Technik **39** (1994), no. 10, 264–269.

[PR50] W. Penfield and T. Rasmussen, *The cerebral cortex of man. a clinical study of localization of function*, The Macmillan Comp., New York, 1950.

[RBEL84] Brigitte Rockstroh, Niels Birbaumer, Thomas Elbert, and Werner Lutzenberger, *Operant control of EEG and event-related and slow brain potentials*, Biofeedback and Self-Regulation **9** (1984), no. 2, 139–160.

[RLD+02] Jan Raethjen, Michael Lindemann, Matthias Dümpelmann, Roland Wenzelburger, Henning Stolze, Gerd Pfister, Christian E. Elger, Jens Timmer, and Günther Deuschl, *Corticomuscular coherence in the 6-15 Hzefs band: is the cortex involved in the generation of physiologic tremor?*, Experimental Brain Research **142** (2002), no. 1, 32–40.

[RWP97] H. Ramoser, J. R. Wolpaw, and G. Pfurtscheller, *EEG-based communication: Evaluation of alternative signal prediction methods*, Biomed. Technik **42** (1997), no. 9, 226–233.

[SBR+03] Michael Schröder, Martin Bogdan, Wolfgang Rosenstiel, Thilo Hinterberger, and Niels Birbaumer, *Automated EEG feature selection for brain computer interfaces*, Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering, 2003, pp. 626–629.

[SKL04] A.-D. Sperfeld, J. Kassubek, and A. C. Ludolph, *Aktuelle Aspekte in der Diagnostik und Therapie der Amyotrophen Lateralsklerose (current aspects in diagnostics and therapy of amyotrophic lateral sclerosis)*, Aktuelle Neurologie **31** (2004), no. 5, 209–215.

[SKSP03] A. Schlögl, C. Keinrath, R. Scherer, and G. Pfurtscheller, *Information transfer of an EEG-based brain-computer interface.*, Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering, 2003, pp. 641–644.

[SS02] B. Schölkopf and A.J. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[TDN+06] Ryota Tomioka, Guido Dornhege, Guido Nolte, Benjamin Blankertz, Kazuyuki Aihara, and Klaus-Robert Müller, *Spectrally weighted common spatial pattern algorithm for single trial eeg classification*, Tech. Report 40, Dept. of Mathematical Engineering, The University of Tokyo, July 2006.

[TDT+94] Camilo Toro, Günther Deuschl, Robert Thather, Susumu Sato, Conrad Kufta, and Mark Hallett, *Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG*, Electroencephalogr. Clin. Neurophysiol. **93** (1994), 380–389.

[Vap95] V.N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, 1995.

[WBH+01] J. R. Wolpaw, Niels Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, *Brain computer interface technology: A review of the first international meeting*, IEEE Transactions on Rehabilitation Engineering **8** (2001), no. 2, 164–173.

[WEBS05] J. Weston, A. Elisseeff, G. BakIr, and F. Sinz, *The spider machine learning toolbox*, `http://www.kyb.tuebingen.mpg.de/bs/people/spider/`, 2005.

[Wel67] Peter D. Welch, *The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms*, IEEE Trans. Audio Electroacoustics **AU-15** (1967), 70–73.

[WEST03] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Michael Tipping, *Use of the zero-norm with linear models and kernel methods*, Journal of Machine Learning Research **3** (2003), no. 7-8, 1439–1461.

[WM94] J.R. Wolpaw and D.J McFarland, *Multichannel EEG-based brain-computer communication.*, Electroencephalography and Clinical Neurophysiology **90** (1994), 444–449.

[WM04] Jonathan R. Wolpaw and Dennis J. McFarland, *Control of a two-dimensional movement signal by a non-invasive brain computer interface in humans*, Proceedings of the National Academy of Sciences of the United States of America (PNAS) **101** (2004), no. 51, 17849–17854.

[WMNF91] Jonathan R. Wolpaw, Dennis J. McFarland, Gregory W. Neat, and Catherine A. Forneris, *An EEG-based brain-computer interface for cursor control*, Electroencephalogr. Clin. Neurophysiol. **78** (1991), 252–259.

[WMVS03] Jonathan R. Wolpaw, Dennis J. McFarland, Theresa M. Vaughan, and Gerwin Schalk, *The Wadsworth Center Brain-Computer Interface (BCI) research and development program*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **11** (2003), no. 2, 204–207.

[YL03] Lei Yu and Huan Liu, *Feature selection for high-dimensional data: A fast, correlation-based filter solution*, Proceedings of the twentieth international conference on machine learning ICML, 2003, pp. 856–853.

[YL04] L. Yu and H. Liu, *Efficient feature selection via analysis of relevance and redundancy*, Journal of Machine Learning Research **5** (2004), 1205–1224.