Using Temporal Evidence and Fusion of Time-Frequency Features for Brain-Computer Interfacing

by

Gireesh S. Dharwarkar

A thesis presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2005

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Brain-computer interfacing (BCI) is a new method of human-machine interaction. It involves the extraction of information from the electroencephalogram (EEG) through signal processing and pattern recognition. The technology has far reaching implications for those with severe physical disabilities and has the potential to enhance machine interaction for the rest of the population. In this work we investigate time-frequency analysis in motor-imagery BCI. We consider two methods for signal analysis: adaptive autoregressive models (AAR) and wavelet transform (WAV). There are three major contributions of this research to single-trial analysis in motor-imagery BCI. First, we improve classification of AAR features over a conventional method by applying a temporal evidence accumulation (TEA) framework. Second, we compare the performance of AAR and WAV under the TEA framework for three subjects and find that WAV outperforms AAR for two subjects. The subject for whom AAR outperforms WAV has the lowest overall signal-to-noise ratio in their BCI output, an indication that the AAR model is more robust than WAV for noisier signals. Lastly, we find empirical evidence of complimentary information between AAR and WAV and propose a fusion scheme that increases the mutual information between the BCI output and classes.

Acknowledgements

First and foremost I would like to thank my advisor, Dr. Otman Basir, for his wisdom and for believing in this area of research. His drive and enthusiasm in everything he does has been inspirational and an example for me. I am indebted to my beloved fiancé, Andrea, who, during challenging times in this research, convinced me I could accomplish anything; she has been my rock. I am grateful for my parents, Sadashiv and Shobha, and my sister, Nita, whose encouragement and influence throughout my life has provided me with the foundation to succeed in this research. I would also like to thank my friends at the PAMI lab; I could always count on them for sharing ideas and having many laughs along the way. Lastly, I would like to thank the researchers in the BCI group at the Institute of Bomedical Engineering, University of Technology, Graz, Austria for providing data for this research.

This is dedicated to the ones who were the first to teach me a love for knowledge, my parents, Sadashiv and Shobha Dharwarkar.

Contents

List of Figures

Chapter 1

Introduction

Brain-computer Interfacing (BCI) is an exciting new technology for communication and interaction with machines. It has far reaching benefits for persons with severe physical disabilities and has the potential to enhance machine interaction for the rest of the population. This technology bridges several disciplines of study: computer and systems engineering in the form of signal processing, pattern recognition, and machine intelligence; electrophysiology; neuroscience; cognitive science; and psychology. In this chapter we provide an introduction to the technology. This includes an overview of the BCI system itself as well as major developments in the field with reference to literature. Finally, based on the major developments that are identified, the scope of this thesis is defined.

1.1 Introduction to Brain Computing

Interaction with computers, whether in the form of a PC or embedded in the myriad of devices people use on a daily basis, is ubiquitous in modern society. The primary challenge is for the user to convey their intentions to the machine in an efficient manner. Advances to this end not only have significant impact on the productivity of society, but also quality of life, as the barriers to communication with the devices upon which we have come to rely cause frustration and stress. Conventional interfaces, such as a keyboard and mouse, make use of a fraction of the information that humans can convey. Other forms of communication that have been integrated into human-machine interfaces (HMI) more recently include speech, hand gestures and even facial expression has received attention from researchers [1]-[3]. Another pressing challenge in HMI is improving accessibility to all persons. Much like buildings have been revamped to accommodate persons with disabilities in latter decades, HMIs must undergo a similar revolution to improve accessibility in modern society.

Brain computing involves the extraction of information directly from the brain through real-time analysis of its electrical activity. In this endeavor, the HMI ascertains the intention of the user by converting electrical activity of the brain into a control signal for devices. In their full potential, brain-computer interfaces (BCI)

are the epitome of user-centric design. They ascertain the will of the user directly from the source, the mind.

BCI's have much to contribute in addressing the challenges of developing HMI's that employ multi-modal information to increase efficiency and information throughput. All of the other modalities originate from the brain; perhaps extracting information from the brain directly can provide some redundancies to improve the overall reliability of the system as well as provide some complimentary information to improve performance [4]. BCI's have a more pressing role in HMI's designed for those with severe physical disabilities resulting in locked-in syndrome. Numerous diseases disrupt the neural pathways that control muscles: Amyotrophic lateral sclerosis (ALS), muscular dystrophies, cerebral palsy, multiple sclerosis and brain and spinal cord injury. Two million people suffer from these disorders in the United States alone [5]. BCI's can provide a revolutionary means for such persons to access and interact with the world around them that most or none of the other modalities can offer.

This research area has been active for ten to fifteen years and successful rudimentary control of devices has been achieved [6]-[8]. The full potential of the technology is yet to be realized. When it is realized the benefits to society will be far reaching.

1.2 The Electroencephalogram

The encephalogram (EEG) is a recording of electrical activity originating from the brain. It is recorded on the surface of the scalp using electrodes made of a highly conductive metal. In this research, the EEG is the signal from which information is extracted to ascertain the intention of the user. All of the data used in this work was retrieved non-invasively (on the surface of the scalp). In this section some background information is provided about the EEG signal. Although the content in the rest of this thesis is focused on signal processing and pattern recognition, in this section some biological, historical and other information about the signal is presented to provide a context for the signal processing challenges.

1.2.1 The Biology of EEG

The brain consists of billions of neurons making up a large complex neural network. Below is a diagram of a neuron. It has several components: the **soma** is the cell body of the neuron and contains the nucleus, which houses genetic information; the **dendrites** extend from the soma and receive chemical messages from other neurons; the **axon** transmits electro-chemical signals to other neurons; the **myelin sheath** consists of fatty tissue cells that insulate the electrical current flowing through the axon; finally the **bouton** is responsible for converting an electrical signal to a chemical signal to be received by other neurons [9].

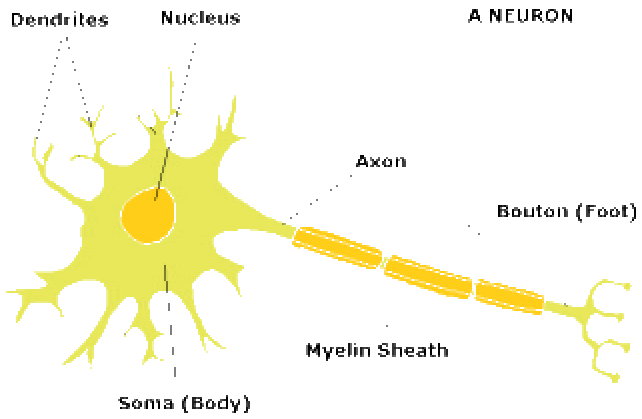

Figure 1-1- The Neuron: the most basic unit in the neural-network of the brain

The processing of information takes place by the firing or pulsing of many individual neurons. The pulse is in the form of membrane depolarization traveling along the axons of neurons. A series of pulses in the neurons, also known as a spike train, encodes the information processes of the neural network [10]. The EEG is the electrical field potential that results from the firing of many neurons. Thus, there is a relationship between the spike train and the EEG and the latter also provides information about neural-network activity [11].

1.2.2 A Brief History of EEG

The first recordings of human EEG on chart paper can be traced back to the Austrian psychiatrist Dr. Hans Berger who published his work in 1929 [12]. In his work he made 73 recordings from a single subject and found regular oscillations at

10 Hz. He named this rhythm the a wave and found that the best recordings came from the occipital region (over the visual cortex at the lower rear of the skull) with another reference electrode on the forehead. Since then neurologists and clinicians have found that scalp-recorded EEG has a frequency range of 0.5 to 40 Hz. The most common categorization of EEG sub bands is ? (0 - 4 Hz), ? (4 – 8 Hz), a (8 – 13 Hz) and ß (13-38 Hz) [13].

EEG has had wide medical applications, from studying sleep stages to diagnosing neurological irregularities and disorders. It was not until the 1970's that researchers considered real-time analysis of EEG, which implied the signal could be used for communication and control. With the computer advances that ensued, active research in EEG utilization for communication has occurred in the last ten to fifteen years.

1.2.3 Signal Conditioning Challenges

There are many complications in acquiring good quality recordings of EEG. The signal itself is very week, in the order of 5-100 µV. At the scalp, EEG is no longer a direct expression of brain activity. Between the brain and scalp are layers of cerebrospinal fluid, bone, and skin, all of which attenuate the signal. This causes poor signal-to-noise ratio. In addition to attenuation, these layers alter the signal more fundamentally through volumetric conduction. This scenario lends itself well

to rather complex *inverse problem* modeling; however, the computational demands of such an approach are not well suited for real-time processing.

For consistent recordings it is important to maintain integrity in the contacts between the electrodes and the scalp, otherwise contact impedance can hamper the quality of the signal. A conductive paste is used to decrease contact impedance and electrode migration. High-gain amplifiers are used to bring the signal levels up to the required level for analog-to-digital converters. Unfortunately, this also significantly amplifies background electrical noise at 60 or 50 Hz depending on the part of the world the system is used. Fortunately, for most brain-computing applications the frequencies of interest are between 0.5 to 40 Hz. However, due to the low signal-to-noise ratio careful analog filter design is required to properly attenuate the background electrical noise.

There are a myriad of other noise sources, commonly referred to as artifacts in the literature, introduced by the body itself. Movement of eyes during recordings produces the electrooculogram (EOG) signal, which can be detected by EEG sensors, particularly those that are at the front of the scalp. Muscle activity in the head emits an electromyogram (EMG) that can also interfere with the EEG signal. Other artifacts are caused by the electrocardiogram (ECG) (electrical activity from heart tissue), sweat, and head and body movement. These artifacts are often within

the frequency ranges that coincide with EEG. EOG is particularly a concern as its amplitude is quite large compared to EEG.

It is apparent that signal conditioning and pre-processing of the EEG in and of itself is a difficult problem. To address this challenge noise cancellation is often used. In this approach, an electrode is placed just above the eyes to acquire the EOG and subtract it from the electrode. More sophisticated approaches, based on information theory, model the problem as *blind source separation* and use *Independent Component Analysis* (ICA) [15], [16]. This method has been used successfully and has the advantage of not requiring additional electrodes, yet identifying artifacts that are statistically independent of EEG processes. EEG preprocessing is an area of research on its own. In this work, pre-processing is not the focus and the datasets used have undergone some minimal pre-processing that will be further discussed in section 2.3.

1.2.4 The International 10-20 System for Electrode Placement

The most common selection for the location of electrode placement is based on an international standard termed the Ten-Twenty system. This standard was established by an international committee and published in 1958 [17]. The positions of the electrodes are relative to landmarks on the skull, mainly the inion, naison and mastoid processes. The term Ten-Twenty stems from the fact that the distance

between bony points, such as the inion and nasion, is divided into segments that are either 10% or 20% of the total length. Thus, the placement specification is scaleable to accommodate all head sizes [18]. Figure 1-2 depicts the arrangement for 19 electrodes. Each electrode is assigned a letter and number describing its position on the skull. The letter describes the area of the brain to which the electrode corresponds such as frontal (F), central (C), and temporal (T). The number refers to the side of the head the electrode is located, where odd numbers indicate the left side and even numbers denote the right side.

The standard specifies the relationship between the electrodes and the parts of the brain. The correspondences between electrodes and the brain were determined in two ways. Metal clips were placed along the fissures of the brain during open brain surgery and X-rays were taken while the electrodes were on the scalp. Secondly, using cadavers, holes were drilled through the skull to the brain at the designated electrode positions. Ink was applied to the holes and the brain was removed from the skull to analyze the markings [18]. For more details on the Ten-Twenty system refer to [17] and [18].

Electrode	Brain Region
Fp1	Left fronto-polar
Fp2	Right fronto-polar
F3	Left superior frontal
F4	Right superior frontal
F7	Left inferior
F8	Right anterior
Fz	Mid frontal
C3	Left central or rolandic
C4	Right central or rolandic
Cz	Vertex
T3	Left mid temporal
T4	Right mid temporal
T5	Left posterior temporal
T6	Right posterior temporal
P3	Left parietal
P4	Right parietal
Pz	Mid parietal
O1	Left occipital
O2	Right occipital

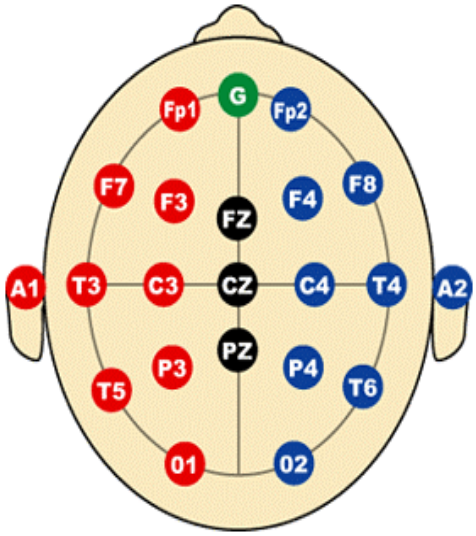

Figure 1-2 – The Ten-Twenty system for electrode placement

1.3 Overview of the Brain-Computer Interface

A brain-computer interface (BCI) is a direct communication channel between a person's brain and a computer or machine. The goal of EEG-based BCI research is

to identify the user's intentions from their EEG and translate these commands into a control signal for a device such as a wheelchair, mouse cursor, or prosthesis. The motivation for this research is to enable severely physically disabled people to access the world around them. In the longer-term BCI technology may be a useful modality for the public at large. Figure 1-3 illustrates the major components of a BCI as well as the flow of information.

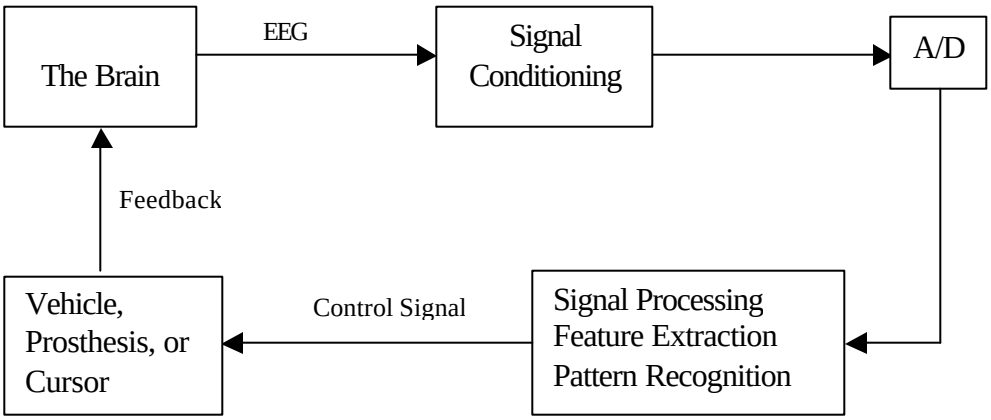

Figure 1-3 - The brain-computer interface

The EEG is acquired via electrode(s) placed on the scalp of the subject. See section 1.2.3 for information about the required signal conditioning upon EEG acquisition. Next the signals must be amplified considerably. The majority of relevant information resides in a frequency range of 0.5-40Hz and band pass filtering is desirable to remove irrelevant frequencies. The signal is ready for processing by a computer, but first it must be discretized by an analog-to-digital

(A/D) converter. Once acquired by the computer, the signal is processed to extract features that provide some insight into the user's intention. These features are classified to determine which of a finite number of user intentions, for a given BCI system, the user is communicating. Once the user intention is determined a control signal can be applied to a device, such as a wheel chair. In Figure 1-1 a feedback loop connects the output of a BCI back to the user. This feedback loop may or may not exist depending on the application of the BCI, but proves to be very useful. In control theory closed-loop systems generally outperform their open loop counter parts. Similarly, the feedback in Figure 1-3 enables the user to adjust their thought process to improve the performance of the system, thus, the learning process is a harmonious one: the system learns to accommodate the user while the user learns to use the system. One of the objectives in BCI research is to minimize the burden on the user to adapt to the system. In most cases the feedback is inherent to the activity. For example, if the user is employing the BCI to move a mouse cursor on a computer screen, the resulting movements are the feedback.

1.4 Cognitive Tasks and Methods in EEG Communication

Currently, BCI's fall into four major categories based on what aspect of the EEG they analyze and which part of the brain the signal originates. Consequently, each type of BCI requires a different set of mental tasks. In this section, we provide an overview of each of the cognitive tasks.

1.4.1 Visual Evoked Potentials

In visual evoked potentials (VEP) the communication occurs through the use of visual information by analyzing the EEG originating from the visual cortex (O1 and O2 in Figure 1-2). The evoked potential is a change in the EEG resulting from an external stimulus [19]. The first use of VEP can be traced back to the work of Jacques Vidal in the 1970's [20]. In this work, VEP is analyzed to detect changes in eye gaze. This tracking of eye gaze can be used, for example, to move a cursor.

Other research efforts, more recently, employ systems that present the user with several symbols or buttons [21], [22]. In [21], they present 64 symbols with different colour alterations in the visual interface. The user focuses on the symbol they would like to select. The system determines the user's selection by comparing the resulting VEP to those acquired during training. In [22], they present several buttons to the user on a screen, each flashing at a different frequency. The photic

driving response over the visual cortex is assessed by the system. This response will closely match the frequency of the button under the user's gaze.

1.4.2 P300 Evoked Potential

P300 or 'oddball' BCI's take advantage of a response in the EEG that occurs when the user observes sporadic stimulus among routine stimuli. The stimuli can be auditory, visual, or somatosensory. The response occurs 300 ms after the anomaly and can be detected in the EEG over the parietal cortex (P3 and P4 in Figure 1-2) [23], [24]. One example of a BCI that uses the P300 property is described in [25]. In this system the user manipulates a virtual keyboard through a matrix of letters displayed on a screen. A single row or column flashes every 125 ms and the user is instructed to pay attention to the key of their choice by counting the number of times it flashes. A P300 response in the EEG occurs after the row and column of the desired letter flashes. The results in [25] imply an achievable communication rate of 1 word per minute (4.8 symbols/min with 90% accuracy). Other work increases this communication rate to 5.45 symbols/min with 92% accuracy [26]. Research efforts in P300 based BCIs utilize tactile and auditory input [27]. This work is particularly beneficial for those with visual impairment.

1.4.3 Slow Cortical Potentials

Slow cortical potentials (SCPs) are low frequency amplitude changes in the

EEG that occur over 0.5 to 10 s. SCPs are generated in the cortex and result from thoughts relating to movement, such as imagining maneuvering in a familiar room. When the cortex is active a reduction in the SCP is observed and when it is inactive an increase occurs [28]. Initial work in SCPs has resulted in the BCI commonly referred to as the "thought translation device" (TTD) [29]. Typically, SCP based BCIs allow the user to make binary decisions by either activating or deactivating the cortex. In [29] each trial is 4 seconds long and the system provides the user with feedback in the form of a cursor that moves to the top or bottom of the screen. The user can move the cursor up or down by controlling their SCP. In this study, ALS patients use the technique with considerable success; it is used to spell words on a screen using a binary decision tree to select letters of the alphabet. This system produces a low information transfer rate of 0.15-3 letters per minute.

1.4.4 µ-Rhythm and Motor Imagery

The EEG rhythm known as µ is related to processes in the motor cortex, which comprises brain activity related to movement of body parts. The µ-rhythm has components in a and ß and is observed over the motor cortex in most people above the age of two years [30]. Activity in the µ-rhythm most relevant to BCIs is event-related (de)synchronizations (ERD/ERS). An ERD is attenuation in the µ-rhythm that accompanies movement or preparation of movement. The ERS is an

amplitude increase, which occurs after the completion of movement-related thought [31]. ERD occur even when one thinks about or imagines movement of body parts, commonly referred to as motor-imagery. This feature makes the cognitive task suitable to those with physical disabilities who cannot actually move their limbs. Typical cognitive tasks in such BCI's are imagination of left vs. right hand movements [32]-[35]. Other work has involved motor imagery involving other somatosensory output [36], [37]. In these communication protocols, the user conveys their intention in predefined time intervals ranging in length from 4 seconds to 9 seconds. Each cycle has an initial idle period and then a cue is presented to the user to indicate that the system is ready classify their intention. After the cue the system knows that it is analyzing task-relevant EEG and deciphers the motor imagery (e.g. left vs. right imagined hand movement). This interface provides a binary control signal that can be used for a simple switch or navigating through a menu. Some of the systems not only facilitate binary classification but also assign magnitude to decisions [37], [38]. This feature enables some subjects to control the magnitude, which is useful for applications such as maneuvering a cursor [36].

1.4.4.1 Asynchronous Communication Protocols

All of the cognitive-task protocols that have been presented above in sections 1.4.1 to 1.4.4 are synchronous. That is they depend on predefined widows of time

where the system knows the user is providing meaningful input. Such protocols have two glaring drawbacks: (1) they limit the flexibility of the user and (2) they limit the information transfer rates. Recently, however, there is research that demonstrates promise for asynchronous protocols within motor-imagery BCIs [39]-[41]. In this work, methods are developed to detect motor events relevant to the system's context. Thus, the system detects the onset of relevant input and can automatically distinguish idle and meaningful periods. During periods of relevant input, the system can apply pattern recognition algorithms for distinguishing the type of motor event. These methods make use of an EEG feature known as movement-related potential (MRP), which are DC EEG spikes that occur at the onset of planned movement.

1.5 Scope

Although all four of the major types of BCIs have been successful in achieving rudimentary control by subjects, the work in this thesis focuses on motor-imagery. There are several reasons for investigating motor-imagery over the other areas. This section outlines those reasons in a preamble to defining the scope of this work.

A glaring weakness of the P300 and VEP approaches is that they require a

stimulus to interact with the user. Fundamentally, these systems have primary control over user interaction. This drawback limits the use of these approaches in asynchronous communication protocols, because the user cannot provide input spontaneously. This in turn limits the application domain of the approach.

SCP communication not related to motor-imagery also has potential to function in an asynchronous environment since they do not require stimuli from the system. However, to the best knowledge of this research, a method has not yet been proposed for asynchronous communication in these methods. Hence, a compelling reason for furthering motor-imagery BCI is its proven feasibility in asynchronous protocols [39]-[41]. Another advantage of motor-imagery, over SCP, is that its cognitive task is more closely related to BCI tasks: moving a cursor left or right; or turning a wheelchair left or right.

Although there are good reasons to favour motor-imagery over other approaches, there has been little work to compare the performance of the four BCI types in their current state. In [42] they compare left vs. right motor-imagery to imagination tasks related to SCP for the same set of subjects in a synchronous protocol. The results for ten subjects age 29-54 indicate that motor-imagery has lower classification performance. Given the promise motor imagery has in other aspects, this is a compelling reason to further improve classification performance.

This thesis specifically investigates classification of left. vs. right hand movements. The analysis is performed off-line on pre-recorded EEG. Current time-frequency analysis and pattern recognition techniques in motor imagery are assessed and compared. Several novel approaches are proposed and investigated. For a detailed problem statement refer to section 3.1.

Chapter 2

Background

In the last chapter we provide compelling reasons for further investigating motor-imagery based BCI's. In this chapter we present a review of the literature specific to this area. In particular, we review current time-frequency and spatial feature extraction techniques and several voids in the research are identified. At the very least, the reader should be familiar with electrode nomenclature of the International Ten-Twenty System (section 1.2.4) and EEG characteristics related to motor imagery (section 1.4.4) prior to reading the material in this chapter.

2.1 Developments in Time-Frequency Analysis of Motor Imagery EEG

In EEG motor-imagery analysis researchers employ time-frequency analysis to identify the occurrence of ERD/ERS and MRP. There are two major time-frequency paradigms that are in common use. One is the autoregressive (AR)

model, its variant the adaptive autoregressive (AAR) model, and, more recently, wavelet analysis. This section reviews the research trends in each as applied to motor-imagery EEG. For a detailed description and formulation of AR, AAR, and wavelet refer to sections 4.1, 4.2, and 5.1 respectively.

Autoregressive modeling has a long-standing tradition in EEG analysis and its origins can be traced back to the late 1960's and early 1970's [43]-[46]. There are several reasons AR modeling prevails in EEG signal processing: it is a maximal entropy spectral estimator and only a few parameters are required to describe spectral information [69]; it does not require *a priori* knowledge of relevant frequencies, as they can vary from subject-to-subject or even within the same subject; it describes the stochastic nature of EEG quite well; and the methods for their computation are well studied [47]. Conventional approaches to AR coefficient estimation assume the signal to be wide-sense stationary, such as the Burg and Yule-Walker methods [48]. These methods can be applied to non-stationary signals, such as EEG, by windowing the signal and approximating it to be stationary within the window. Depending on the size and shift of the window, traditional AR estimation algorithms can become computationally intensive. Furthermore the stationary assumption within the window is often poor and choosing the size of window is problematic with respect to the uncertainty principle in time-frequency analysis.

Instead, dynamic estimation techniques using Kalman filtering and other recursive least-square estimators are more computationally efficient [49] and suitable in real-time. These techniques allow the parameters to adapt with each sample of the signal, which is a better model of the EEG. In BCI literature, this approach to estimating autoregressive parameters is commonly referred to as adaptive autoregressive (AAR) modeling.

AAR modeling has been used with some success in BCI [32], [33]. In these studies the tasks to be distinguished are motor imagery of the dominant hand versus mental arithmetic using EEG from C3 and C4. Autoregressive parameters are adapted over time using a lattice-filter approach to minimizing the mean square error. The autoregressive coefficients are used as features and a Bayesian learning framework classifies them. They use a latent-space smoothing approach to ascertain certainty in decisions and reject low confidence decisions. In the strictest rejection they achieve an average performance of 86.5% over 7 subjects. An obvious drawback in this approach is the information loss in ignoring trials. Another disadvantage is that the two distinguishing tasks are not very intuitively related to each other or a particular application.

A significant number of studies have investigated Kalman filtering estimation of AAR parameters to distinguish left versus right imagined hand movements by the

Graz BCI group [50]-[56]. These studies are predominantly off-line and assess performance through cross-validation of many single trials. In each trial, however, an on-line classifier provides the user with feedback during data collection. Thus, the work investigates performance with feedback and in [53] classification accuracy of 85% to 95% is achieved for four subjects. There are three major contributions throughout these publications over previous AAR EEG analysis. Firstly, unlike [33] where they use completely different categories of cognitive tasks to improve separability, the Graz BCI achieves significant separability within the same cognitive-task (imagined hand movements). Secondly, in this work there is attention to assessing how long it takes to reach a decision. They accomplish this by performing cross-validation of the features collected at each instant in time over the trial. This assessment is very relevant to real-world applications of BCI where response time is of concern. Thirdly, they propose using the mutual information (MI) between the BCI output signal and the motor-imagery classes to assess the quality of communication. This analysis gives more insight into the reliability of the algorithm in a real on-line BCI.

AAR estimation techniques, such as Kalman filtering, do not yet have a direct and well-verified means for selecting optimal time-frequency resolution. Furthermore, no methods exist that offer a tractable multi-resolution approach in

AAR modeling. As discussed in section 1.4.4 relevant ERD features occur both in the µ and ß bands, therefore, multi-resolutional analysis should be beneficial. It is a natural progression for research efforts to consider wavelet analysis in motor-imagery BCI [39], [58], [63], [65], [68]. In [39] they use the wavelet transform coefficients to characterize MRP in asynchronous motor-event detection. In [63] they use wavelet analysis on subdural implants and achieve nearly 100% accuracy. This success can be largely attributed to a much higher signal-to-noise ratio in these invasive recordings. Also, EEG recorded directly from the surface of the brain has more information across the spectrum making multi-scale analysis very beneficial. There is little work that has applied wavelet analysis to non-invasive EEG in single-trial motor imagery using performance measures proposed in [56]. A notable contribution in this regard is in [58]. They apply wavelets scaled appropriately for the µ-rhythm a and ß bands and use the coefficients as features. Furthermore, they propose a method for combining decisions made over the trial to accumulate evidence. For a single subject they achieve 89.3 % accuracy in cross-validation.

2.2 Prospects and Challenges in Motor Imagery BCI

There has been little work to compare the performance of AAR and wavelet features in motor-imagery EEG. In [65] they compare AR and wavelet for three subjects for three mental states: relaxed, right, and left motor imagery. They suggest that AR is better based on the overall confusion matrix. However, distinction between left and right motor imagery was near random for both feature sets, which makes the comparison less credible. Also the wavelet approach was better at distinguishing the relaxed and motor-imagery states. Thus, the comparison is not conclusive. The proposed wavelet approach in [58] was the winning entry in the BCI Competition 2003 and competed with AR-based methods. However, it is unclear if the approach in [58] performed better because of one or more of the following reasons: (1) it used wavelet features; (2) it used temporal evidence accumulation which none of the other approaches used; or (3) the AR implementations were not as good as they could be (even the competition organizers comment on the variability in the quality of submissions and none of the other AR methods were published [64]) Therefore, the best time-frequency paradigm (wavelet vs. AAR) continues to be debatable and inconclusive.

Another aspect that requires attention is the utilization of all information the

EEG provides about motor imagery. MRP and ERD/ERS contribute complimentary information about motor imagery tasks [66]. In addition, the locations of ERD activation in regions of the motor cortex give insight into the imagined tasks. The method of common spatial patterns (CSP) can capture this information effectively [67]. In [68], they demonstrate that combining these complimentary features, particularly MRP and ERD (via AAR features), at the classifier level improves the information transfer rate. This is encouraging, since the information transfer rate is the single greatest drawback in non-invasive EEG based BCI's. Perhaps another question worth answering is whether or not there are complimentary features available within one or more of the information sources discussed above. The hierarchical structure of this idea is illustrated in Figure 2-1.

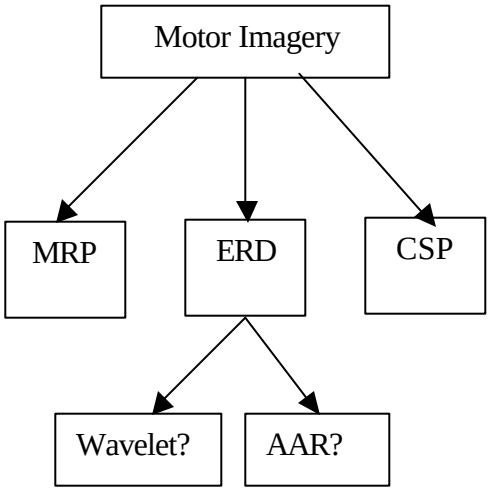

Figure 2-1 – This hierarchical structure demonstrates the fusion of various sources of information from the EEG related to motor-imagery. The novel idea is in the bottom layer, i.e. the suggestion that wavelet and AAR features could have complimentary information with respect to ERD.

2.3 Description of Data Set

Although data was analyzed off-line in this study, all signal processing is causal and applicable on-line. We consider three subjects imagining left and right hand movements using similar protocols as in [58]. The subjects are referred to as C1, B2, and A3. The dataset was provided by the Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology, Graz, Austria. For subject C1, the EEG was sampled at 128 Hz from three biopolar channels (C3, Cz, and C4) that were band-pass filtered allowing frequencies

between 0.5 to 30 Hz. The dataset consists of 280 trials that are 9 s long each. Figure 2-2 depicts the protocol for each trial. The first 3 s is an idle preparation period, at which point a visual cue in the form of an arrow appears pointing either to the left or the right. The user was instructed to perform the imagination task according to the direction of the arrow for the next six seconds. During this time the system provided the user with visual feedback using an on-line classifier. Subjects B2 and A3 had exactly the same protocol except their trial length was 8 s and feedback started after 4s.

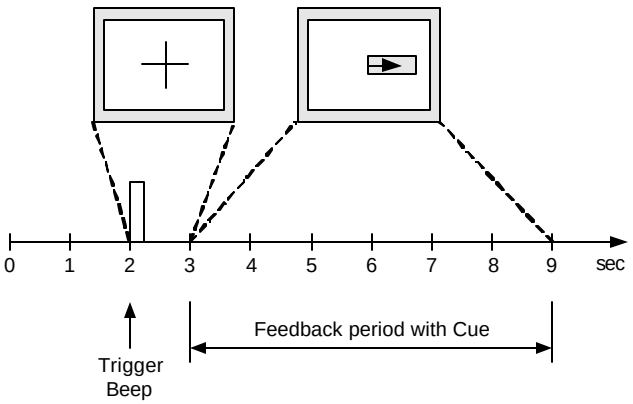

Figure 2-2 – The 9 second protocol for each imagined left/right hand movement.

Chapter 3

Overview of Research

In this section we discuss the research direction of this thesis with regard to advancing time-frequency pattern analysis in motor-imagery. We discuss how the objectives of this research fill various voids in current research.

3.1 Problem Statement

This thesis seeks to advance current methods of ERD analysis in motor-imagery classification. As discussed in Chapter 2, there are two major approaches to time-frequency analysis: adaptive autoregressive (AAR) models and wavelet analysis. In section 2.2 we briefly describe an approach that uses wavelet analysis and enhances classification by combining decisions over time, thus accumulating knowledge about the trial [58]. We propose that the AAR approach could benefit from a similar scheme that we refer to as the temporal evidence accumulation (TEA)

framework (section 4.3.2). We implement the TEA approach and compare its performance to that of a conventional method for classifying AAR features. The TEA approach to classification of both AAR and wavelet features is an interesting context in which to compare the two features sets. This has never been done in the literature and can provide valuable insight regarding individual strengths of each approach.

As discussed in the previous chapter, the literature suggests that AAR features model non-invasive EEG quite well. The strength of this feature set is perhaps its robustness to the large amount of noise present in non-invasive EEG. However, the obvious strength of the wavelet transform is its multi-resolution time-frequency analysis approach. This suggests that each approach offers different attributes in signal analysis. Therefore, after comparing the two feature sets under the TEA framework we investigate complimentary information between the two approaches. If adequate complimentary information exists, then there is good reason to investigate fusion to improve performance. Figure 3-1 summarizes the flow of investigation in this thesis.

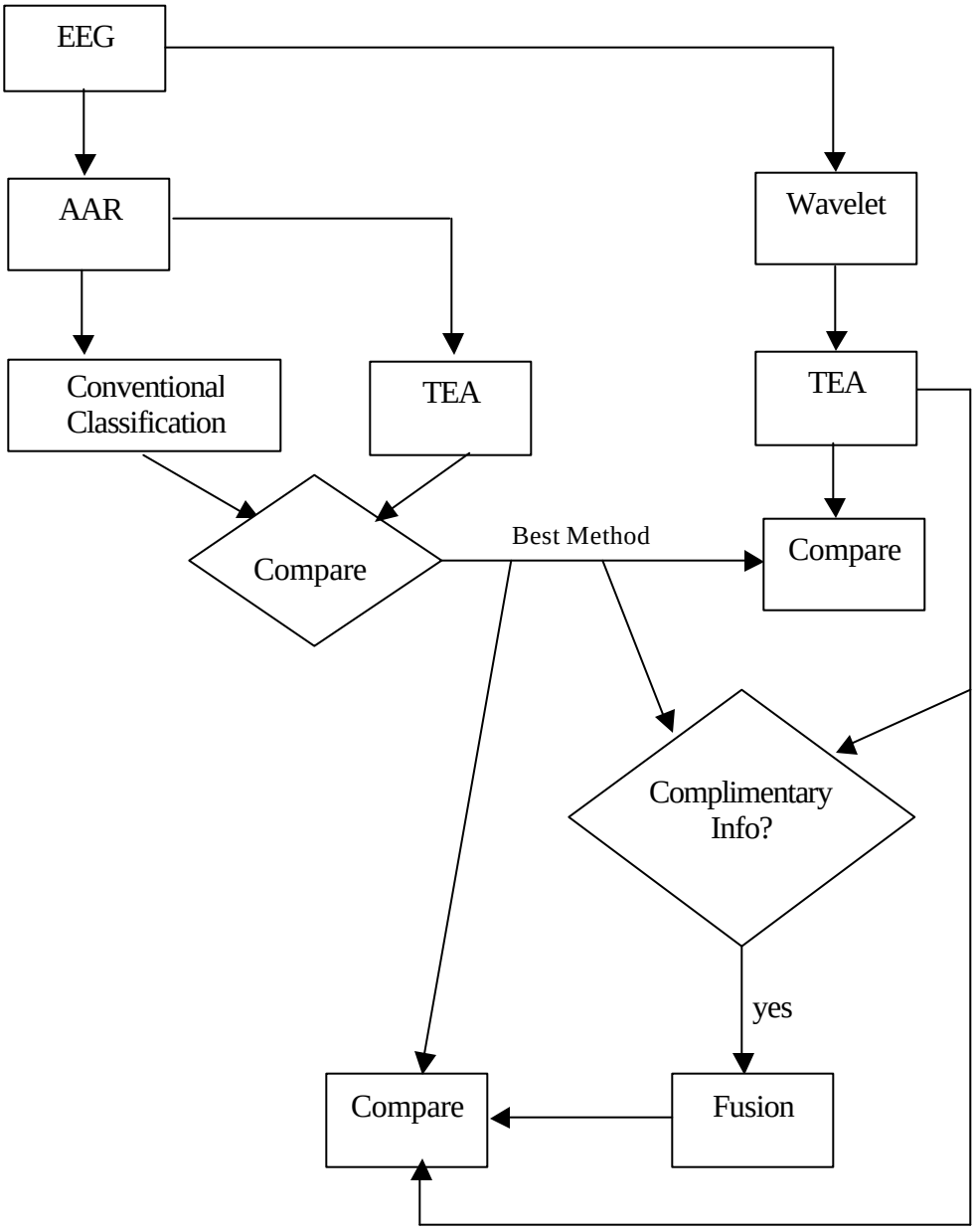

Figure 3-1 – Overview of Research Path in this Thesis

3.2 Thesis Organization

The remaining chapters carry out the investigation depicted in Figure 3-1. Chapter 4 describes the adaptive autoregressive model, the TEA framework, and how the framework can be applied to AAR parameters. It then compares the performance of TEA to a conventional classification approach. Chapter 5 presents the wavelet approach, under the TEA framework, for motor-imagery analysis and presents results for the three subjects. Chapter 6 compares the performance of wavelet and AAR under the TEA framework and investigates complimentary information between the two. Furthermore, an approach for fusing the methods is proposed to take advantage of complimentary information. In Chapter 7, conclusions from the various experiments are drawn and future work is recommended.

Chapter 4

Feature Set 1 – Adaptive Autoregressive Model

In the previous chapter we establish that two methods of spectral analysis are considered in this work: adaptive autoregression (AAR) and wavelet; in this chapter we focus on the former. We provide background theory in autoregression and describe a commonly used technique for extracting AAR parameters from EEG and classifying them. Finally, we propose a classification scheme that improves performance over the aforementioned conventional approach.

4.1 The Autoregressive Model

A stochastic process is a random variable that evolves over time. Therefore,

the process evolves according to probabilistic laws and cannot be predicted exactly [48]. According to this description, EEG may be modeled as a stochastic process. In order to detect and characterize an ERD and ERS one must estimate the power spectral density of the EEG. There are a myriad of techniques to estimate the power spectrum of a wide-sense stationary random processes. Although EEG is not stationary (see section 2.1), the stationary assumption is the starting point for this discussion on power spectral estimation. The existing estimation procedures can be broadly categorized as parametric and non-parametric. In pattern recognition applications parametric methods are particularly attractive because they summarize information concisely and translate well to feature vectors. The autoregressive (AR) model is one such approach that has been well studied in EEG analysis.

In autoregression, the model infers the *way* in which a random process $y(n)$ is synthesized [48]; $y(n)$ may be generated by applying an all-pole filter to a white-noise process $w(n)$ with zero mean and variance s_y^2. The idea is illustrated in the block diagram of Figure 4-1. Notice that the signal under analysis is actually the output of the system. This is in contrast to other spectral analysis techniques in which the signal of interest is decomposed as the input to filter(s). The filter coefficients a_m, therefore, provide information about the spectral content of $y(n)$ and are referred to as AR parameters. Equation (4.1) describes the relationship between

the spectral density of *y* denoted as S_y and the AR parameters for a p[th] order AR model [69].

$$S_{AR} = \frac{s_v^2}{|1 - \sum_{k=1}^{p} a_k e^{-jvk}|^2} \quad \textbf{(4.1)}$$

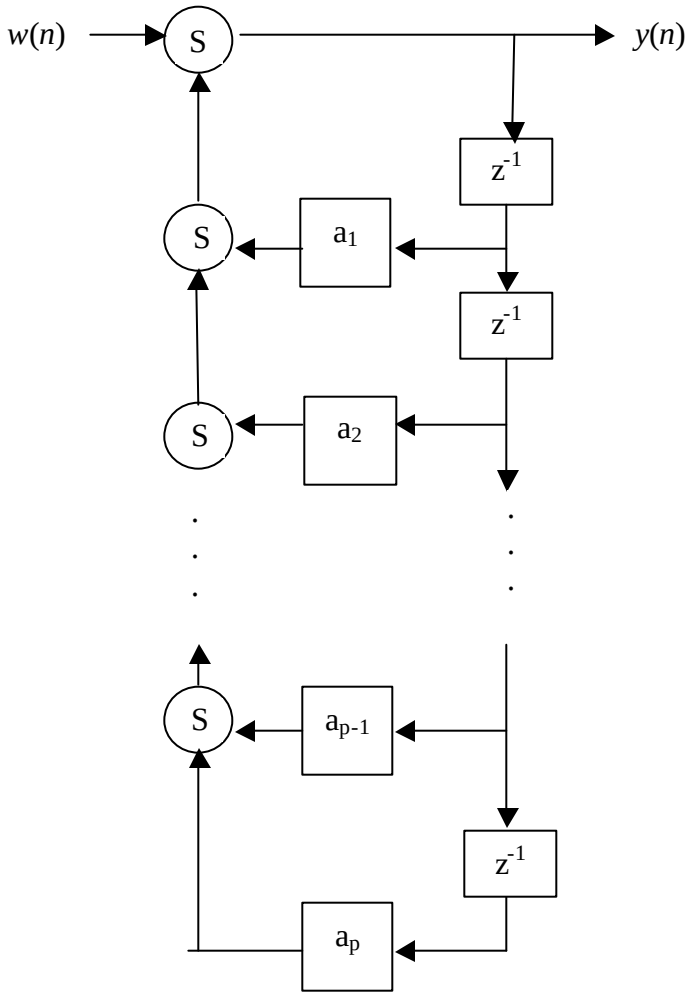

Figure 4-1 – The AR Process, where *w*(*n*) is the input (zero-mean white noise) and y(*n*) is the output (the signal to be modeled)

In the EEG analysis context the AR parameters are, of course, unknown. There are several methods to estimate AR parameters for a wide-sense stationary process such as the Yule-Walker and Burg's methods [48]. We do not further discuss these approaches because they apply to stationary processes. The estimation method used in this research is discussed in section 4.2.

The AR model itself is not an intuitive description of the process that actually occurs to produce EEG. The signal is a result of the firing of many neurons, not the all-pole filtering of zero-mean white noise. The AR model seems to characterize EEG well (see chapter 2), nonetheless, and there are good reasons for using it as discussed below. Firstly, the AR spectrum is closely related to the spectrum that can be estimated by the *maximum entropy method* (MEM) [69]. In fact, the two spectral estimation techniques are mathematically equivalent (see [69] for a proof). MEM is known to give the power spectrum that is associated with the most random time series possible for a given autocorrelation matrix. Thus, the AR parameters provide a lot of spectral information with few parameters. Secondly, AR modeling is rather robust with regard to noise, with which non-invasive EEG is significantly corrupted. This notion is quite intuitive since the model itself is driven by noise and because it is a maximal entropy estimate. Lastly, *a priori* information about relevant

frequencies is not required to estimate the AR coefficients. This is powerful because if the relevant frequencies change the same implementation is applicable. This is a concern in BCI since the EEG characteristics may change for a given subject over time.

4.2 The Adaptive Autoregressive Model

In the previous section the AR model is described for a wide-sense stationary process. EEG, however, is non-stationary and the AR parameters change with time. Conventional approaches to address non-stationary behavior of EEG have been to apply a moving window of AR estimates. Depending on the size and shift of the window, traditional AR estimation algorithms can become computationally intensive. Instead, we may estimate non-stationary AR parameters by inferring a dynamic model and using Kalman filtering [70]. Such an approach is referred to as adaptive autoregressive modeling (AAR) and has been successfully used in left-right single-trial motor imagery classification [50]-[55]. This section introduces the technique as applied to EEG signal processing.

4.2.1 Introduction to Kalman Filtering

The Kalman Filter is a dynamic, recursive estimator that uses a state-space model to represent an estimation problem. The state, at any given time n, changes;

we are interested in estimating the *state vector*, which consists of the fewest number of parameters to specofy the state. We may infer the state at time n from measurements in the system (in this case the EEG) at time n as well as the history of states in the system. The advantage of the Kalman Filtering formulation is its recursive nature, which requires knowledge of the previous state only. Therefore it is an attractive tool because it is computationally efficient and is a least-squares estimate of the unknown state.

There are two major components to the Kalman Filtering formulation: (1) a process model and (2) a measurement model. The process model describes the dynamics of the state and is given by

$$\mathbf{z}(n) = \mathbf{F}(n \mid n-1)\mathbf{z}(n-1) + v_{\mathrm{p}}(n), \qquad \textbf{(4.2)}$$

where $\mathbf{z}$ is a column vector of the state parameters, $\mathbf{F}$ is a matrix representing the dynamics of the state, and v_p is zero-mean white noise with a Gaussian distribution. The role of v_p in the model depends several contexts : (1) the dynamics that govern $\mathbf{z}$ may be purely stochastic ($\mathbf{F}=\mathbf{I}$), in which case v_p is the driving force for the dynamics; (2) if the system is believed to be deterministic and there is no noise term measurements would be completely ignored, in this case a small noise term v_p preserves the adaptability by restricting the confidence of the estimator; and (3) the mathematical model for $\mathbf{F}$ may only be an approximation of the actual dynamics and

v_p provides flexibility to accommodate unaccounted variations in the model. In order to initiate the recursion a boundary condition is necessary, namely

$$E[\mathbf{z}(0)] = \mathbf{z}_0 \text{ and } \text{cov}(\mathbf{z}_0) = \mathbf{P}_0 . \quad \textbf{(4.3)}$$

The dynamics matrix **F,** which models the state transitions, has two noteworthy properties:

1. Product Rule - $\mathbf{F}(n,m)\mathbf{F}(m,l) = \mathbf{F}(n,l)$
2. Inverse Rule - $\mathbf{F}^{-1}(n,m) = \mathbf{F}(m,n)$

where *l*,*m*, and *n* are integers.

The measurement model is given by

$$\mathbf{y}(n) = \mathbf{C}(n)\mathbf{z}(n) + v_m(n), \quad \textbf{(4.4)}$$

where ***y*** is a column vector of length *M* of the *M*-dimensional measurements, **z** is a column vector of the state parameters, **C** is the measurement matrix and v_m is zero-mean white Gaussian noise.

For a complete derivation of the solution for the unknown state using the process and measurement equations refer to [69]. Here we highlight the key parameters that provide conceptual insight into Kalman filtering and summarize the algorithm.

1. The Innovations Process

The innovations process is defined as

$$u(n) = \mathbf{y}(n) - \overline{\mathbf{y}}(n \mid g_{n-1}), \tag{4.5}$$

where u is the innovations process and $\overline{\mathbf{y}}(n \mid g_{n-1})$ is the minimum mean-square estimate of the observed measurement $\mathbf{y}(n)$ given all past observations denoted by g_{n-1}. Notice that $u(n)$ quantifies new information in the observed data $\mathbf{y}(n)$.

There are three important properties of the innovations process [69]:

1. According to the principle of orthogonality, The innovations process is orthogonal to all past observations $\boldsymbol{y}(n)$ described as

$$E[u(n)y^{*}(k)] = \mathbf{0},\ 1 \le k \le n-1. \tag{4.6}$$

2. The innovations $u(1),\ u(2),\ u(3) \ldots.\ u(n)$ are orthogonal to each other (white) described as

$$E[u(n)u^{*}(k)] = \mathbf{0},\ 1 \le k \le n-1. \tag{4.7}$$

3. There is a one-to-one correspondence between the observations $\boldsymbol{y}(1)$, $\boldsymbol{y}(2)$, $\boldsymbol{y}(3)$,….. $\boldsymbol{y}(n)$ and $u(1),\ u(2),\ u(3) \ldots.\ u(n)$.

2. Correlation of the Innovations Process and Predicted State Error

The correlation matrix for the innovations process is a necessary parameter and can be calculated by

$$\mathbf{Q}(n) = \mathbf{C}(n)\mathbf{P}(n, n-1)\mathbf{C}^{H}(n) + \mathbf{R}(n), \tag{4.8}$$

where $\mathbf{R}$ = cov($\boldsymbol{v}_{\mathrm{m}}$) and $\mathbf{P}(n, n\text{-}1)$ = cov[$\mathbf{z}(n) - \overline{\mathbf{z}}(n \mid g_{n-1})$], i.e. it is the correlation of the predicted state error. $\mathbf{P}(n, n\text{-}1)$ may be found recursively using the Riccati equation given by

$$\mathbf{P}(n, n-1) = \mathbf{F}(n)\mathbf{P}(n-1)\mathbf{F}^{H}(n) + \mathbf{R}(n). \tag{4.9}$$

$\mathbf{P}(n\text{-}1)$ is the updated estimate of $\mathbf{P}(n\text{-}1, n\text{-}2)$, i.e. $\mathbf{P}(n\text{-}1) = \mathrm{cov}(\mathbf{z}(n) - \overline{\mathbf{z}}(n \mid g_{n}))$ and may be calculated as

$$\mathbf{P}(n-1) = \mathbf{P}(n-1, n-2) - \mathbf{F}(n)\mathbf{G}(n)\mathbf{C}(n)\mathbf{P}(n-1, n-2). \tag{4.10}$$

$\mathbf{G}(n)$ is the Kalman gain and is further discussed below.

3. The Kalman Gain

The Kalman gain represents the algorithms confidence in new information provided by measurements and is given by

$$\mathbf{G}(n) = \mathbf{F}(n)\mathbf{P}(n-1)\mathbf{C}^{H}(n)\mathbf{Q}^{-1}(n). \tag{4.11}$$

4. **Estimation of State**

Finally, the estimate of the state parameters at time n can be calculated using the following

$$\overline{\mathbf{z}}(n) = \mathbf{F}(n)\mathbf{z}(n-1) + \mathbf{G}(n)u(n). \tag{4.12}$$

The parameters and algorithm are concisely summarized in Table 4.1 and Table 4.2.

Table 4.1 – Summary of Kalman Filter Parameters

Symbol	Definition	[*] Dimensions
$\mathbf{z}(n)$	State at time *n*	M by 1
$\mathbf{y}(n)$	Measurement at time *n*	N by 1
$\mathbf{F}(n)$	State transition matrix from time *n*-1 to *n*	M by M
$\mathbf{C}(n)$	Measurement matrix at time *n*	N by M
$\mathbf{W}(n)$	Correlation matrix of process noise	M by M
$\mathbf{R}(n)$	Correlation matrix of measurement noise	N by N
$\bar{\mathbf{z}}(n)$	Estimate of the state at time *n*	M by 1
$\mathbf{G}(n)$	Kalman gain at time *n*	M by N
$u(n)$	Innovation process at time *n*	M by N
$\mathbf{Q}(n)$	Correlation matrix of the innovations process at time *n*	N by N
$\mathbf{P}(n,n-1)$	Correlation matrix of the error in $\bar{\mathbf{z}}(n \mid g_{n-1})$	M by M
$\mathbf{P}(n)$	Updated Correlation matrix of the error in $\bar{\mathbf{z}}(n \mid g_{n})$	M by M

[*] *M* and *N* refer to the number of dimensions in the measurement and state, respectively

Table 4.2 – Summary of the Kalman Filter Algorithm

Input:

Measurements = {**y(**1**),** y**(2), ….. ,** y**(**n**)}**

Known Parameters:

Transition matrix = **F(**n)

Measurement matrix = **C**(n)

Correlation matrix of process noise = **W**(*n*)

Correlation matrix of measurement noise = **R**(*n*)

Computations:

$$\mathbf{G}(n) = \mathbf{F}(n)\mathbf{P}(n-1)\mathbf{C}^H(n)[\mathbf{C}(n)\mathbf{P}(n,n-1)\mathbf{C}^H(n) + \mathbf{W}(n)]^{-1}$$

$$u(n) = \mathbf{y}(n) - \overline{\mathbf{y}}(n \mid g_{n-1})$$

$$\overline{\mathbf{z}}(n) = \mathbf{F}(n)\mathbf{z}(n-1) + \mathbf{G}(n)u(n)$$

$$\mathbf{P}(n-1) = \mathbf{P}(n-1,n-2) - \mathbf{F}(n)\mathbf{G}(n)\mathbf{C}(n)\mathbf{P}(n-1,n-2)$$

$$\mathbf{P}(n,n-1) = \mathbf{F}(n)\mathbf{P}(n-1)\mathbf{F}^H(n) + \mathbf{R}(n)$$

Initial conditions:

$$\overline{\mathbf{z}}(0) = E[\mathbf{z}(0)]$$

$$\mathbf{P}(1,0) = E[(\overline{\mathbf{z}}(1) - E[\overline{\mathbf{z}}(1)])(\overline{\mathbf{z}}(1) - `E[\mathbf{x}(1)])^H]$$

4.2.2 Estimating AAR parameters using Kalman Filtering

A p^{th} order AAR model can describe a 1-D EEG signal using

$$y(n) = \mathbf{z}(n)^T * \mathbf{Y}(n-1) + v_m(n), \qquad (4.13)$$

where n is the discrete time index; y is the EEG sample; **z** is a p-element column vector of the autoregressive coefficients; **Y** is a column vector of the last p EEG samples; and v_m is a zero-mean, white-noise process . Notice that the autoregressive coefficients change with time to capture the dynamics of the EEG. In the context of the Kalman Filter, the EEG signal is the measurement and the autoregressive features comprise the state parameters. Thus, equation (4.13) is the measurement model where the measurement matrix **C** from equation (4.4) is a column vector of the last p samples of the EEG signal *y*. The process model is given by

$$\mathbf{z}(n) = \mathbf{z}(n-1) + v_p(n). \qquad (4.14)$$

where v_p is zero-mean white noise. In the context of the Kalman filter the dynamics matrix is identity, thus, the system is modeled as a random walk with small changes in the state.

[1]**Table 4.3 –** Summary of Kalman Filter Parameters for AAR model

Symbol	Definition	Dimensions
W(n)	Process noise variance	scalar
R(n)	Variance of measurement noise	scalar
$\bar{\mathbf{z}}(n)$	Estimate of AAR parameters (state)	p x 1
Y(n)	Measurement matrix (last p samples of y)	p x 1
G(n)	Kalman gain	p x 1
?(*n*)	Innovation process	scalar
Q(n)	Innovation process variance	scalar
P(n)	Predicted state-error correlation matrix	p x p
uc	Update coefficient	scalar

The Kalman filtering algorithm for AAR parameter estimation is given below (see Table 4.3 for description of parameters):

$$Q(n) = (1 - uc) * Q(n-1) \tag{4.15}$$

$$\mathbf{G}(n) = \mathbf{P}(n-1) \cdot \mathbf{Y}(n) / Q(n) \tag{4.16}$$

$$\hat{\mathbf{z}}(n+1) = \hat{\mathbf{z}}(n) + \mathbf{G}(n)u(n) \tag{4.17}$$

$$W(n) = uc * trace(\mathbf{P}(n-1)) / p \tag{4.18}$$

$$\mathbf{P}(n) = \mathbf{P}(n-1) - \mathbf{G}(n)^{\mathrm{T}} \cdot \mathbf{Y}(n-1) \cdot \mathbf{P}(n-1) + W(n) \tag{4.19}$$

The reader should notice that there are several simplifications from the general Kalman filtering equations given in the previous section. Since the state

[1] Note that the dimensions of the parameters are given specifically for the AAR estimation problem and have been simplified from Table 4.1

transition matrix is identity in this application, and the innovations process is scalar (the innovations correlation is simply a variance) equation (4.11) simplifies to equation (4.16). The identity transition matrix also simplifies equation (4.9) making two steps (update and prediction) for calculation of the state error correlation matrix unnecessary. Instead, we combine equations (4.9) and (4.10) (the updated and predicted steps) into one calculation to form equation (4.19).

There are other deviations in the above algorithm from the original Kalman filtering formulation that stem from the fact that the measurement and process noise variance is unknown; the algorithm given in Table 4.2 assumes these parameters are known, therefore, a slightly different approach is necessary. Calculation of the innovations process variance, $Q(n)$, requires the process noise variance, $R(n)$ (see equation (4.8)). There are several alternatives to calculate $Q(n)$ recursively without prior knowledge of $R(n)$ [47]. Based on findings in [47], where they compared these alternatives in EEG analysis for several subjects, we use equation (4.15). Equation (4.18) estimates the unknown measurement noise variance and is among the better performing methods among those studied in [47]. Note that in equations (4.15) and (4.18) the parameter *uc* (update coefficient) is introduced; in the above algorithm the unknown parameters can be traced back to the model order p and update coefficient *uc*. These parameters are significant and their selection governs the time-frequency

resolution of the analysis. Section 4.2.3 describes a method for selecting these parameters.

4.2.3 The REV Criterion and Time-Frequency Resolution

The update coefficient *uc* and model order *p* are to be selected so that the AAR estimates best describe the EEG signal *y*. In [71], they propose the relative error variance (REV) as a minimization criterion for selecting these parameters and it is given by

$$REV = \frac{\sum u(n)^2}{N} / \mathrm{var}(y), \qquad (4.20)$$

which is the mean squared error (innovation process) normalized by the signal variance. *N* is the total number of samples in the trial. The innovation process is a measure of the dynamics that the system could not predict. In essence it is a measure of the goodness-of-fit of the model and is, therefore, a good criterion to consider when selecting *p* and *uc*. In this work the REV criterion is used as a guideline for selecting *p* and *uc* but we do not necessarily use the parameter values that minimize the REV. Marginally lower REV values do not justify larger model orders if the increase in complexity and computation do not significantly improve classification performance.

An important realization is the implication of the model order and update coefficient on the time-frequency resolution and the principle of uncertainty. The

model order corresponds to the frequency resolution and the update coefficient is related to the time resolution [47]. Therefore, for a given update coefficient there is an optimal model order and vice versa.

4.3 Method

This section describes two methods for classifying AAR features as either left or right-hand imagined movements. Section 4.3.1 presents the conventional method (CONV) that has been used in several studies [50]-[56]. Section 4.3.2 describes a novel framework for temporal classification and demonstrates how the framework can be applied to AAR features. The objective is to classify the trial with a significant degree of certitude as quickly as possible starting from the time of feedback (see section 2.3 for a description of the protocol used in these trials).

4.3.1 Conventional Method (Linear-Discriminant Analysis)

Linear Discriminant Analysis (LDA) identifies the best discriminating (n-1)-dimensional hyper plane between two classes *C1* and *C2* in an n-dimensional feature space - for a more detailed discussion of (LDA), see [61]. Despite the simplicity of the method it continues to be a strong pattern analysis technique in BCI and performs better than or the same as more complicated and dynamic techniques such as Hidden Markov Models (HMM) [72]; the main problem in HMM's in BCI is

adequate *a priori* information for the topology. LDA classification of left/right motor-imagery may be performed by training (i.e. finding the linear discriminant for the training set) at each discrete instant in time. We may use cross-validation to assess performance over time by examining the error rate and other measures that will be discussed in section 4.4.1. In a BCI implementation of this algorithm a time during the trial would be selected, based on the cross-validation analysis, when decision making is at its best and the system would wait until that time to classify the trial.

Let $\mathbf{z}_{t,3}(n)$ and $\mathbf{z}_{t,4}(n)$ denote row vectors of the estimated p^{th} order AAR parameters for trial t at time n for electrodes C3 and C4, respectively. We concatenate the two vectors to form a single $2p$-dimensional feature vector giving

$$\mathbf{z}_t(n) = [\mathbf{z}(n)_{t,3}, \mathbf{z}(n)_{t,4}]. \qquad (4.21)$$

Using the LDA weight vector $\boldsymbol{w_t}$ and offset w_0, the classification decision $D_t(n)$ may be obtained by

$$D_t(n) = \mathbf{w}_t(n) \cdot \mathbf{z}_t(n) - w_0, \qquad (4.22)$$

where $D_t(n) > 0$ classifies trial t as right and $D_t(n) < 0$ as left at time n. The magnitude of D_t is an indication of the certitude of the decision.

4.3.2 Temporal Evidence Accumulation Framework

In the previous section we introduce a conventional approach using LDA that classifies the trial at each discrete instant in time independently from one another. We propose that this approach does not take advantage of the accumulated evidence the AAR features are providing over time. We apply a temporal evidence accumulation framework based on one proposed in [58], and investigate its success with AAR features.

Labeled training data from 10-fold cross validation of the dataset is used to infer a class conditional p-D Gaussian distribution for both classes Left (L) and Right (R). Thus we have the probability density for the feature set $\mathbf{z}_t(n)$:

$$p(\mathbf{z}(n) \mid c) = \frac{|\Sigma_{n,c}|^{-1/2}}{4\pi^2} \exp\{-\frac{1}{2}(\mathbf{z}(n) - m_{n,c})^T \Sigma_{n,c}^{-1}(\mathbf{z}(n) - m_{n,c})\}, \qquad (4.23)$$

where $c \in \{L, R\}$. The mean $m_{n,c}$ and covariance $\Sigma_{n,c}$ are estimated from the training data for c. We may classify $\mathbf{z}_t(n)$ by calculating the probability of belonging to a class using (4.23) and Bayes theorem as

$$p(c \mid \mathbf{z}(n)) = \frac{p(\mathbf{z}(n) \mid c)}{p(\mathbf{z}(n) \mid L) + p(\mathbf{z}(n) \mid R)}. \qquad (4.24)$$

In order to combine information throughout time the decision made at the current time $n = n_t$ is a weighted average of all the decisions for $n = n_t$; thus we have

$$p(c \mid \mathbf{z}(1)...\mathbf{z}(n_t)) = \frac{\sum_{n \le n_t} w_t \, p(c \mid \mathbf{z}(n))}{\sum_{n \le n_t} w_t}, \tag{4.25}$$

where w_t reflects our certitude in the decision making ability of the classifier at time n_t, and is derived from the probability of misclassification [58]. Although we cannot find the probability of misclassification directly we may calculate its upper limit using the Chernoff bound; an advantage of using an explicit probabilistic approach to classification. The weight w_t is defined as

$$w_t = 1 - \min_{0 \le g_t \le 1} \int p(z(t) \mid L)^{g_t} \, p(z(t) \mid R)^{1-g_t} \, dz \, . \tag{4.26}$$

Since we infer a Gaussian distribution, this integral can be expressed in closed form and the minimum solution can be found analytically or numerically. The Chernoff bound was chosen over the Bhattacharyya bound (see [61]) because, in general, it is a tighter bound. Although the Bhattacharyya method can be slightly more computationally efficient, we prioritize the tighter bound since Chernoff calculations are done in the training stage, which does not have real-time demands. For the derivation of the Chernoff bound and other details regarding its relationship with the Bhattacharyya bound see Appendix A.

4.4 Results

In this section we compare the performance of the conventional approach (CONV) described in section 4.3.1 to the temporal evidence accumulation (TEA) approach described in section 4.3.2 for the three subjects in this study. The REV criterion is used to select the model order and update coefficient.

4.4.1 Performance Measures

We employ three measures of performance using 10-fold cross validation: the time course of the percent misclassified (%Err); the signed decision magnitude, D(n); and the mutual information, I(n) [38].

D(n) reflects both the classifier's decision by its sign (D(n) > 0 → R, D(n) < 0 → L) and certitude by its magnitude . For CONV, the signed decision magnitude $D_{CONV}(n)$ is the shortest distance between $\mathbf{z}(n)$ and the LDA boundary (see equation (4.22)) and, for TEA, $D_{TEA}(n)$ is defined as

$$D_{TEA}(n) = 1 - 2 \cdot p(L \mid \mathbf{z}(1)...\mathbf{z}(n_c)) . \qquad (4.27)$$

The reasons for examining D(n) are two-fold: it gives a good indication of the separability of classes and the classifier's confidence in its decision; second, it collapses the problem into a single dimension lending itself well to computational simplicity and efficiency for information theoretic measures. D(n) can be modeled as

having two processes: one being the motor imagery task and the other being noise [38]. Information theory may be used to evaluate how much information D(*n*) provides about the motor imagery task. The entropy of D(*n*) is a measure of its uncertainty at time *n* and, inferring a Gaussian distribution, is defined as

$$H(D(n)) = \frac{1}{2}\log(2\pi e\sigma^2), \qquad (4.28)$$

where σ^2 is the variance of D(*n*) (see Appendix B for derivation in the Gaussian case and other details). The reduction in entropy from H(D(*n*)) to the within-class entropy H(D(*n*)|*c*) is the mutual information [38]. It is a measure of how closely related D(n) is to the motor imagery classes, and thus, quantifies the amount of information in D(*n*) that is relevant to the problem. The mutual information can be calculated by

$$I(D(n)) = \frac{1}{2}\log\left[\frac{\sigma^2}{0.5(\sigma_L^2 + \sigma_R^2)}\right], \qquad (4.29)$$

where σ_L^2 and σ_R^2 are the class conditional variances of D(n), assuming the noise process for D(*n*) is also Gaussian. If this assumption is incorrect, $I(D(n))$ in equation (4.29) is in fact the upper limit on the mutual information [38]. See appendix B for more details on mutual information.

4.4.2 REV plots and selection of Model Order and Update Coefficient

We calculate the REV criterion over a range of model orders and update coefficients. Figure 4-2 shows plots for model orders ranging from 3 to 6. In general the REV criterion did not improve for model orders greater than 6 so we do not plot them in the figure. The update coefficient is set to a range of 2^{-14} to 2^{-7} as the Kalman Filter exhibited instability at approximately $uc > 2^{-7}$ for all three subjects. For all subjects and model orders minimum REV values occur close to 2^{-7} and this value was chosen for the update coefficient. Since only minor improvements in the criterion occur for larger model orders we select a 3rd order AAR model. To check this selection, after the 3rd order model classification results were produced (section 4.4.3), we ran the experiments for higher model orders and the results were worse or as good as the 3rd order model.

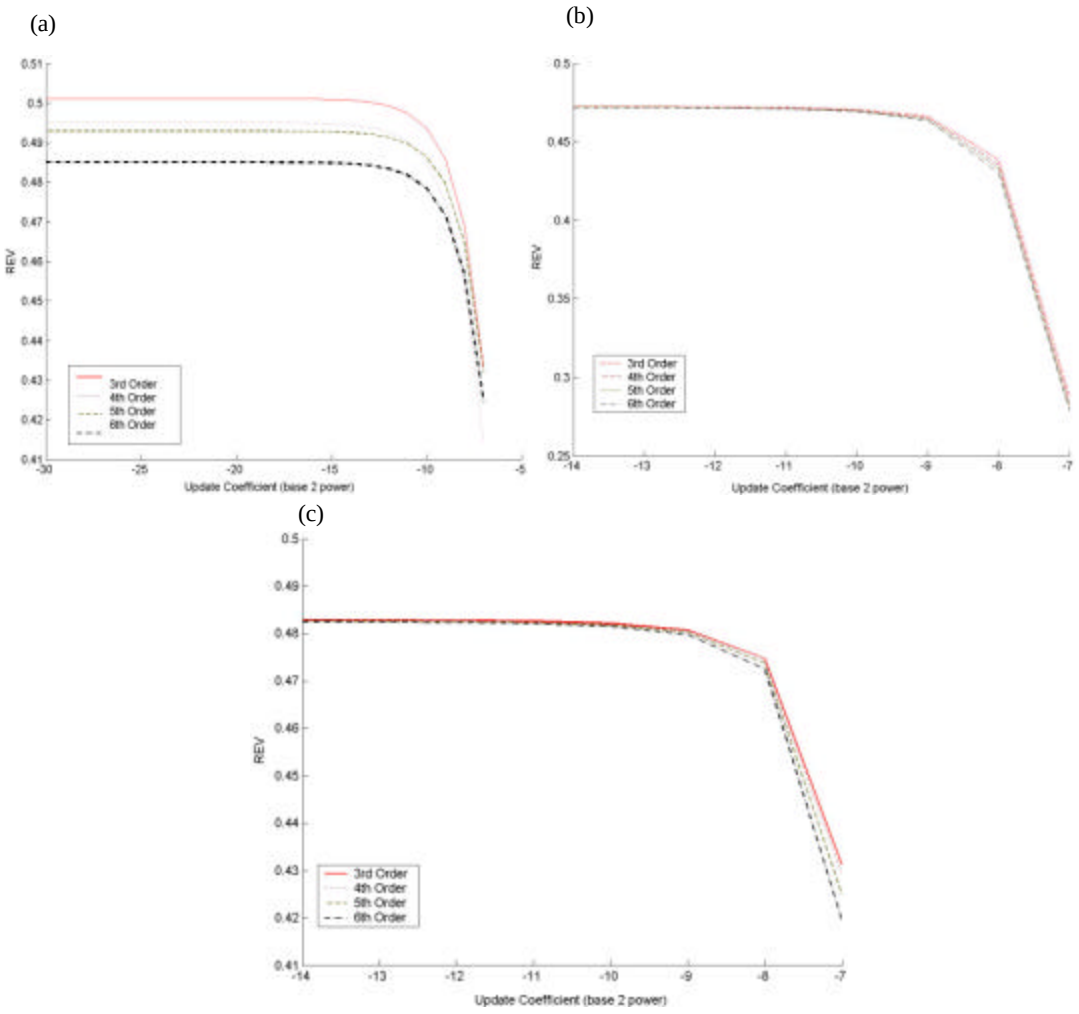

Figure 4-2 – REV criterion plot for several model orders over a range of update coefficients for (a) subject C1, (b) subject B2, and (c) subject C3

4.4.3 Performance Comparison

The time course of the classification results and mutual information for both experiments are depicted in Figure 4-3 to Figure 4-5, for subjects C1, B2, and A3, respectively.

For subject A1 The TEA method reaches a minimum error of 14.3% while CONV achieved a very inconsistent minimum of 18.2%. If one were to consider a moving average of the CONV error curve the minimum would be 21.6%. More significant than the reduction in minimum error is the substantial difference in variability throughout the time course. This is an important improvement in the TEA approach for enhanced predictability of the BCI system.

The advantages of the TEA approach are more compelling in the plot of mutual information. Notice that $I_{TEA}(n)$ reaches a maximum of 0.53 compared to 0.41 for $I_{CONV}(n)$. The latter maximum is inconsistent as there is more variability in $I_{CONV}(n)$. If we again consider a moving average of the curve, $I_{CONV}(n)$ has a maximum of only 0.33. In addition to attaining higher maximal mutual information, I_{TEA} demonstrates a more rapid rate of increase. This is a useful improvement in BCI applications where timely response to user input is desirable.

The time course of $D_{TEA}(n)$ and $D_{CONV}(n)$ are depicted in figure 2. Notice the greater consistency in $D_{TEA}(n)$ and faster rate at which the classification problem

is more separable.

The results for both TEA and CONV reflect an improvement in the separability of the classes over time followed by degradation. For CONV this is evident both in %Err and $I_{CONV}(n)$ in Figure 1 (a) and (b), respectively. For TEA this is evident in the Chernoff bound. This may be a result of a decrease in concentration and focus on the part of the user. There is an important difference in the way the two algorithms handle this problem. The CONV approach does not make use of the knowledge the training data provides about this degradation. The TEA method, however, measures uncertainty in the training data enabling it to predict the degradation in features in the latter part of the trial.

For subject B2 The reduction in error rate over time in TEA is much more consistent then CONV (see Figure 4-4 (a)). TEA reaches its minimum error rate of 27% 1.2 seconds before CONV reaches its minimum of 25%. Although CONV's minimum is lower it is only for an instant in time. In an actual BCI it would be difficult to predict the optimal response time for CONV. Furthermore, since response time is a concern in a BCI, the TEA's faster descent is attractive in many applications. The mutual information plot in Figure 4-4 (b) also demonstrates the more rapid increase in class relevance of the BCI task for TEA. In the latter part of the trial the mutual information is sustained in TEA and declines in CONV, also

demonstrating the robustness of the TEA algorithm. Although the average magnitude of the decision output for CONV is greater in the latter part of the trial (see Figure 4-3 (c) and (d)), it is inconsistent; the mutual information plot, which is directly related to the variance of the decision output, reflects this. Similar trends are observed for subject A3, but the improvements in the TEA approach are even more compelling (see Figure 4-5).

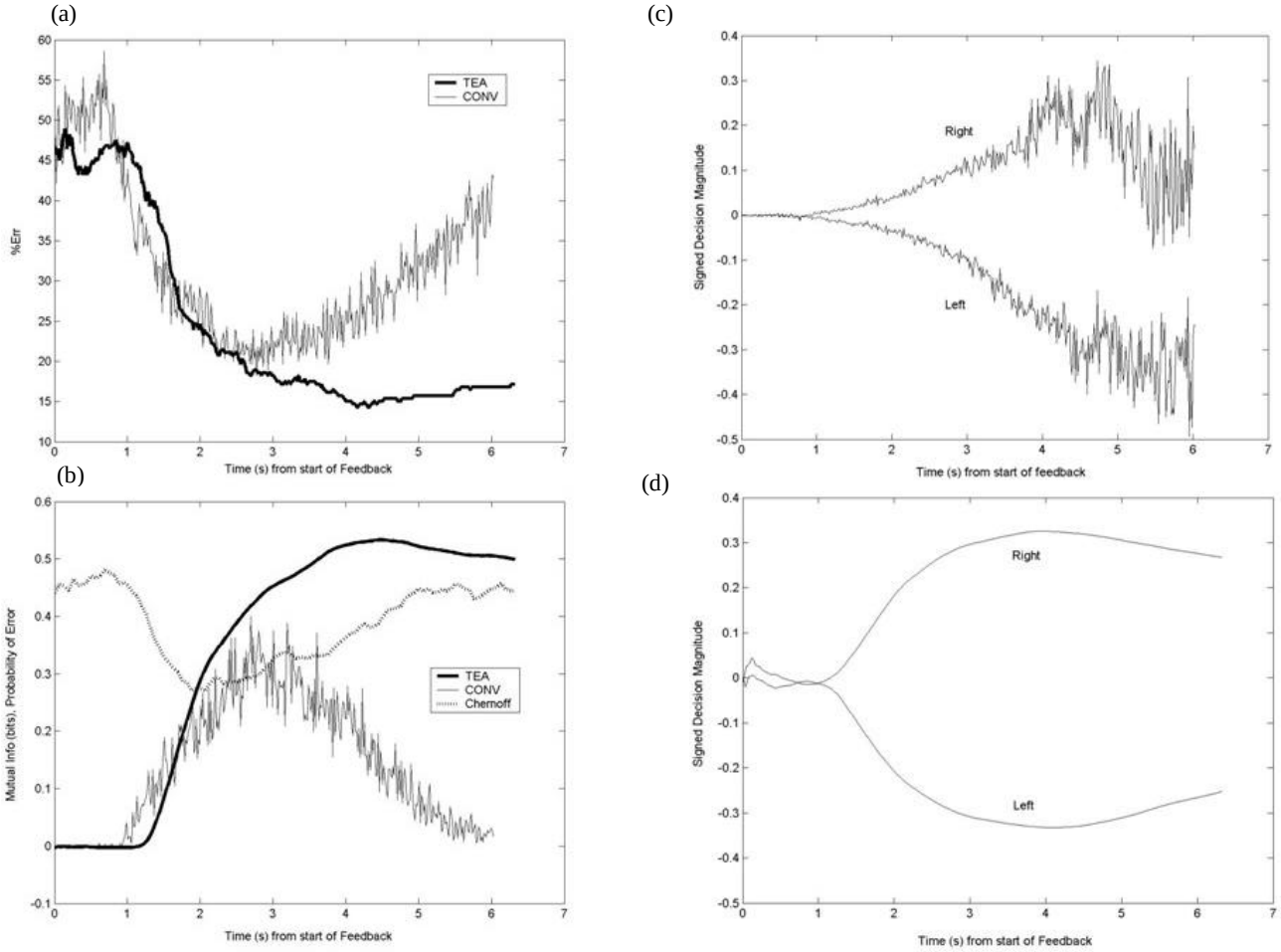

Figure 4-3 – SUBJECT C1: (a) Time course of percent error from 10-fold cross validation for the TEA and CONV approaches to classification, (b) the time course of the Chernoff bound (dotted line) and the mutual information between the decision output and classes for TEA (thick line) and CONV (thin line) methods, (c) time course of left and right decision outputs for the CONV (d) and TEA (b) methods. For CONV the distance to classification boundary has been scaled to be numerically comparable to TEA.

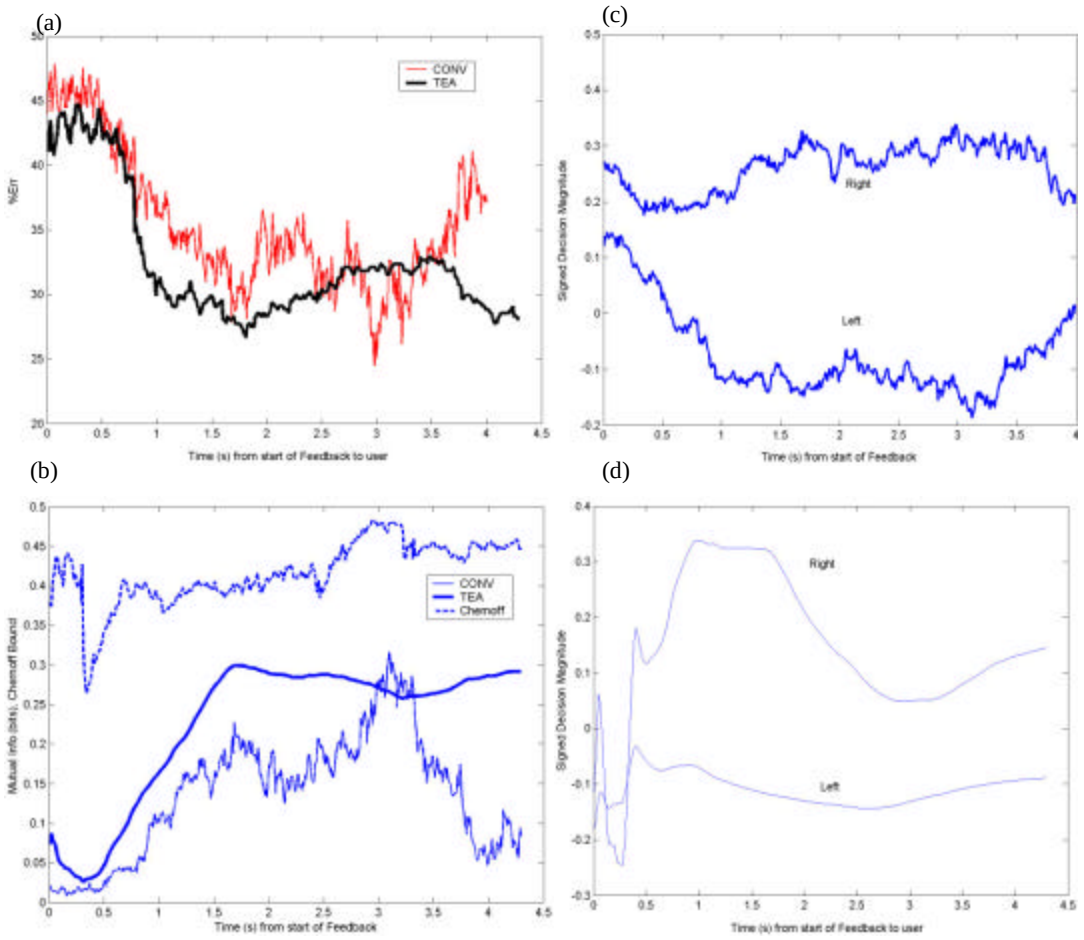

Figure 4-4 – SUBJECT B2: (a) Time course of percent error from 10-fold cross validation for the TEA and CONV approaches to classification, (b) the time course of the Chernoff bound (dotted line) and the mutual information between the decision output and classes for the (thick line) and CONV (thin line) methods, (c) time course of left and right decision outputs for the CONV (d) and TEA (b) methods. For CONV the distance to classification boundary has been scaled to be numerically comparable to TEA.

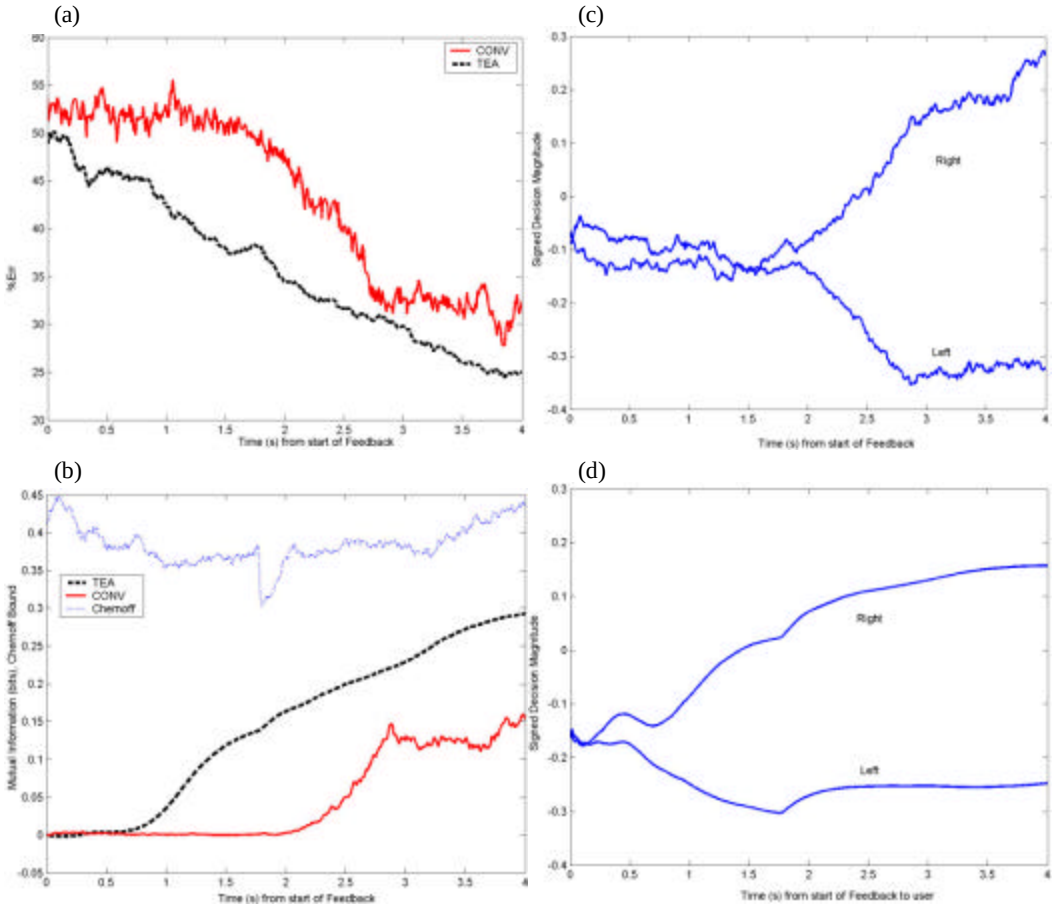

Figure 4-5 – SUBJECT A3: (a) Time course of percent error from 10-fold cross validation for the TEA and CONV approaches to classification, (b) the time course of the Chernoff bound (dotted line) and the mutual information between the decision output and classes for the TEA (thick line) and CONV (thin line) methods, (c) time course of left and right decision outputs for the CONV (d) and TEA (b) methods. For CONV the distance to classification boundary has been scaled to be numerically comparable to TEA.

Chapter 5

Feature Set 2 – Wavelet Transform

In this chapter we consider the second method of spectral analysis: the wavelet transform. We present some fundamental wavelet theory and describe a method for extracting wavelet coefficients from motor-imagery EEG. We then apply the TEA framework discussed in the previous chapter to classify the wavelet features and analyze the results.

5.1 Introduction to The Wavelet Transform

According to Fourier theory a continuous signal can be expressed as a weighted integral of complex sinusoids of varying frequency [59]:

$$f(t) = \frac{1}{2p} \int_{-\infty}^{\infty} F(jw) e^{jwt} dw \ , \qquad (5.1)$$

where $f(t)$ is the time domain signal and $F(j?)$ is the weight function for the complex sinusoids, also known as the frequency domain of the signal. Therefore, the frequency domain provides a measure of activity in the signal throughout the frequency spectrum and is defined as

$$F(jw) = \int_{-\infty}^{\infty} f(t) e^{-jwt} dt \quad . \qquad (5.2)$$

The major shortcoming in Fourier analysis is that, although it provides information about the frequency content of a signal, it does not indicate *when* frequencies occur in a signal. In EEG signal analysis we are interested in how the frequency content of the signal changes with time, known as time-frequency analysis. To overcome this disadvantage of the Fourier Transform (FT) the Short-Term Fourier Transform (STFT) was proposed. In this approach, the Fourier transform is applied to segments of the signal in time. An important consideration is how large to make the window of analysis. The smaller the window the more precise the information is with respect to time. Suppose we apply the most precise window of time possible, the Dirac impulse. This is equivalent to convolution of the Dirac impulse with the signal in the time domain, and corresponds to multiplication in the frequency domain. In the frequency domain the Dirac pulse contains all possible frequencies so the frequency information of the signal is corrupted. Therefore, there is a tradeoff

between time and frequency resolution. This is known as the Heisenberg Principle of Uncertainty as applied to time-frequency analysis.

The wavelet transform addresses the shortcomings of the FT and STFT with respect to the principle of uncertainty. Specifically, the modulated window that is shifted along the signal is done so at various scales [60]. Therefore, high frequencies are measured at sharper time resolutions than low frequencies to establish a compromise between the time-frequency resolution trade-off. The continuous wavelet transform may be applied as

$$w(s,t) = \int_{-\infty}^{\infty} f(t) y_{s,t}^{*}(t) dt \ , \qquad (5.3)$$

where w is the wavelet coefficient that corresponds to the frequency associated with the scale s and time t of the wavelet function $y_{s,t}(t)$. The wavelet function is scaled and shifted versions of a *mother wavelet* $y(t)$:

$$y_{s,t}(t) = \frac{1}{\sqrt{s}} y\left(\frac{t-t}{s}\right) \ . \qquad (5.4)$$

The mother wavelet function is not specified in equation (5.4) as it can take many forms that suit the specific application. For details regarding the types of common mother wavelets and the properties they all must satisfy refer to [60].

5.2 Method of Wavelet Analysis in Motor Imagery EEG

As discussed in section 1.4.4, motor imagery response can be found in the a band over the post central motor cortex and the ß band over the pre central motor cortex. Furthermore, there seems to be complimentary information in these two frequency bands and, thus, merit in exploiting both features [57]. With this in mind, we employ a method of wavelet analysis based on the work presented in [58]. In order to take full advantage of wavelet analysis it is useful to obtain *a priori*

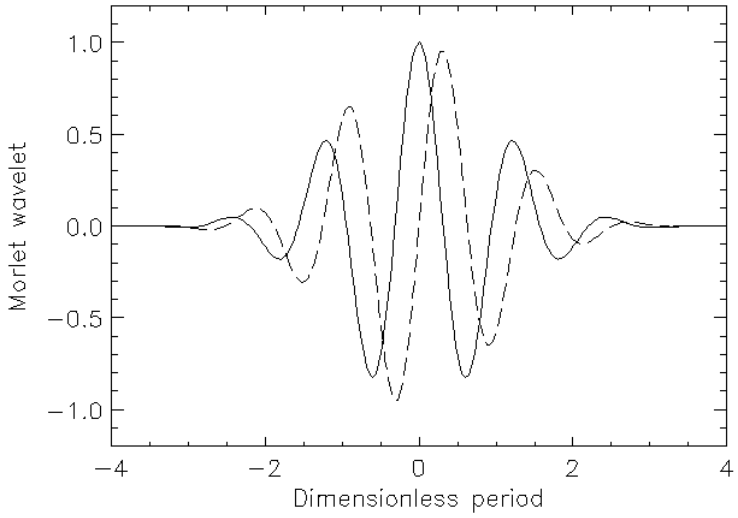

Figure 5-1 – The Morlet Wavelet

information about the dominant frequency within the aforementioned bands. This allows us to scale the mother wavelet precisely for these two relevant frequency bands. Therefore, the *a priori* knowledge allows us to limit the feature space to just two wavelet coefficients per electrode. In this study we use the Morlet mother wavelet, which is a modulated Gauss impulse (see Figure 4-1) and is given by

$$y(n) = \frac{1}{\sqrt{s}} p^{-\frac{1}{4}} e^{(iw_0 \frac{n-t}{s})} e^{-\frac{1}{2}(\frac{n-t}{s})^2}, \qquad (5.5)$$

where n is the discreet time index; s is the scaling factor corresponding to the target frequency; t is the temporal shift of the wavelet; and $?_0$ is the eigenfrequency of the wavelet.

The scaling factor governs the time-frequency resolution and is given by

$$s(f) = \frac{w_0 + \sqrt{2 + w_0^2}}{4pf}, \qquad (5.6)$$

where f is the target frequency. Note that increasing the eigenfrequency $?_0$ sharpens the frequency resolution while decreasing the time resolution.

The effective widths in the time and frequency domain are given by

$$t_{eff} = \sqrt{2}s, \qquad (5.7)$$

$$f_{eff} = \sqrt{2}(2ps)^{-1} , \qquad (5.8)$$

where t_{eff} is also known as the e-folding time and refers to the length of the wavelet (in time) after a $\frac{1}{e^2}$ reduction in the power of the wavelet, and f_{eff} is the corresponding width in the frequency domain.

As discussed in section 5.1, the wavelet coefficients $w(t,f)$ are the convolution of the signal with the mother wavelet:

$$w(t\,,\,f) = \frac{1}{\sqrt{s}} \parallel y(t) * y\,(t) \parallel , \qquad (5.9)$$

where y(t) is the 1-D EEG signal from a given electrode. In the application of this algorithm to the subjects in this study, the feature vector w_v consists of four coefficients at each discreet time n:

$$w_v(n) = (w_{C3}(n,\, f_a), w_{C3}(n,\, f_b), w_{C4}(n,\, f_a), w_{C4}(n,\, f_b))^T \ . \qquad (5.10)$$

where *C3* and *C4* refer to the corresponding electrodes in the international 10-20 system.

In order to make the off-line algorithm implemented in this work applicable on-line, all analysis is causal. To meet this requirement the extension of the wavelets in the time domain is limited to four times the e-folding time. Therefore, a

delay is required and is governed by the most course resolution in the time domain, that which corresponds to α. Hence, The feature vector $w_v(n)$ is not available until time $n+2t_{eff}$. The idea is depicted in Figure 5-2. Due to the necessary delay, classification of the trial does not occur until $2t_{eff}$ seconds after the start of feedback.

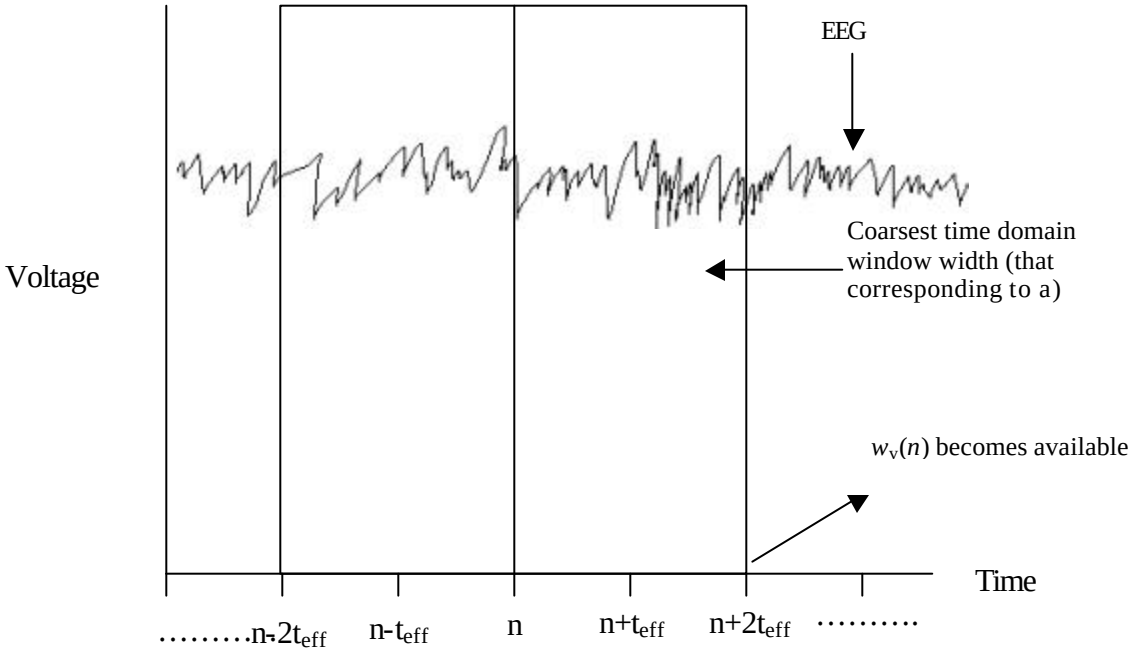

Figure 5-2 – Causal delay of coefficients due to windowing by wavelet

We perform cross-validation of the feature vector at each discrete instant in time over the 280 trials for each subject. The same performance measures are used to evaluate the approach as described in section 4.4.1. For classification, we apply the TEA framework as described in section 4.3.2.

5.3 Results

In this section we apply the wavelet transform method, described in the previous section, to the three subjects C3, B2, and A1. As in Chapter 4, three performance measures are used: the time course of the percent misclassified (%Err); the signed decision magnitude, D(*n*); and the mutual information, I(*n*) [38] (see section 4.4.1).

For most people relevant frequency bands in motor-imagery tend to be at $f_a = 10Hz$ and $f_b = 20Hz$; we scale the mother wavelet appropriately according to equation (5.6). We perform 10-fold cross-validation on several model parameter settings based on the prominent frequencies. The performance for the WAV method is presented in Figure 5-3, Figure 5-4, and Figure 5-5. The TEA framework classifies the wavelet features successfully with minimum error rates of 12%, 19%, and 29% for subjects C1, A2, and B3 respectively. Further insights into the results below will be discussed in the next chapter where the performance is compared to the AAR-based TEA method of Chapter 4.

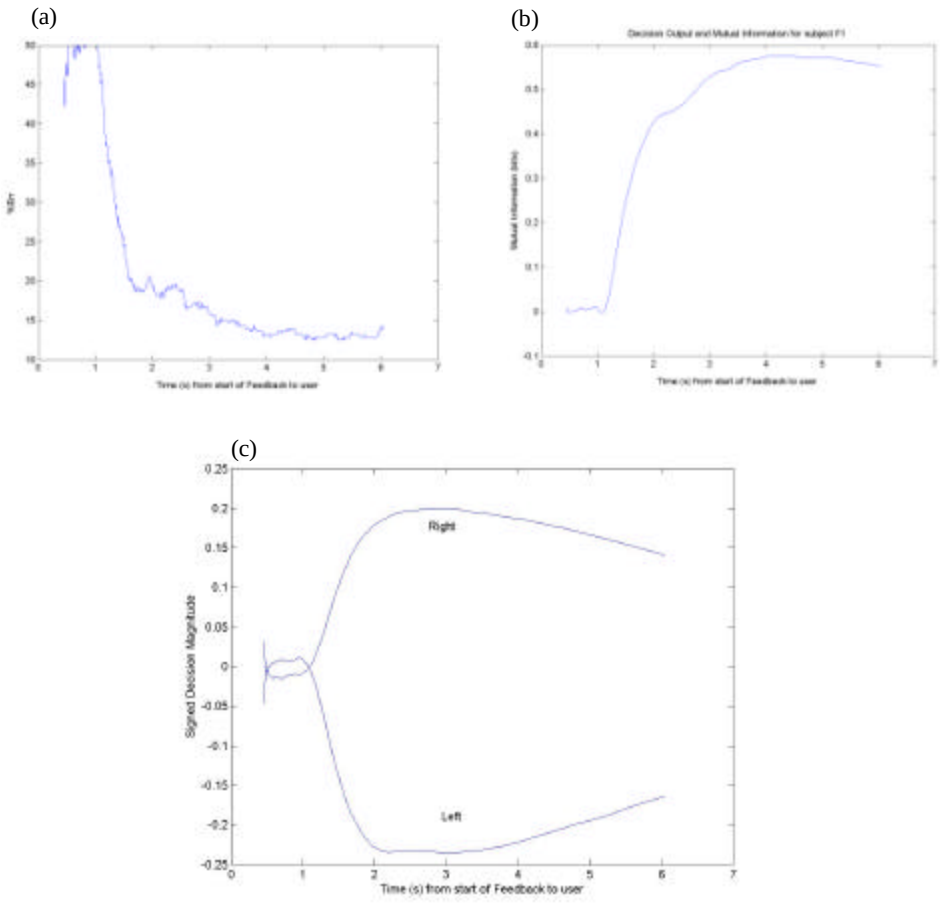

Figure 5-3 – SUBJECT C1: (a) Time course of percent error from 10-fold cross validation for the TEA approach using wavelet features, (b) the time course of the mutual information between the decision output and classes (c) time course of left and right decision outputs

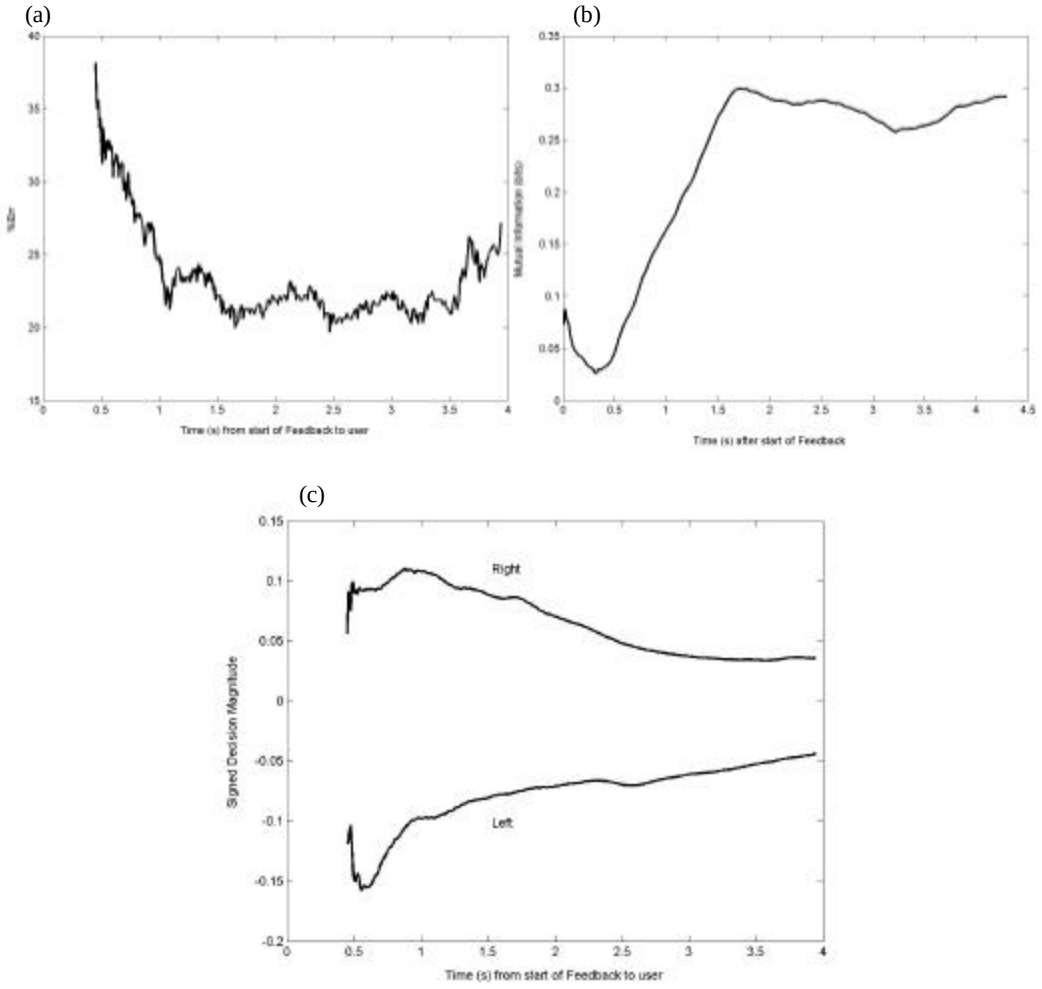

Figure 5-4 – SUBJECT B2: (a) Time course of percent error from 10-fold cross validation for the TEA approach using wavelet features, (b) the time course of the mutual information between the decision output and classes (c) time course of left and right decision outputs

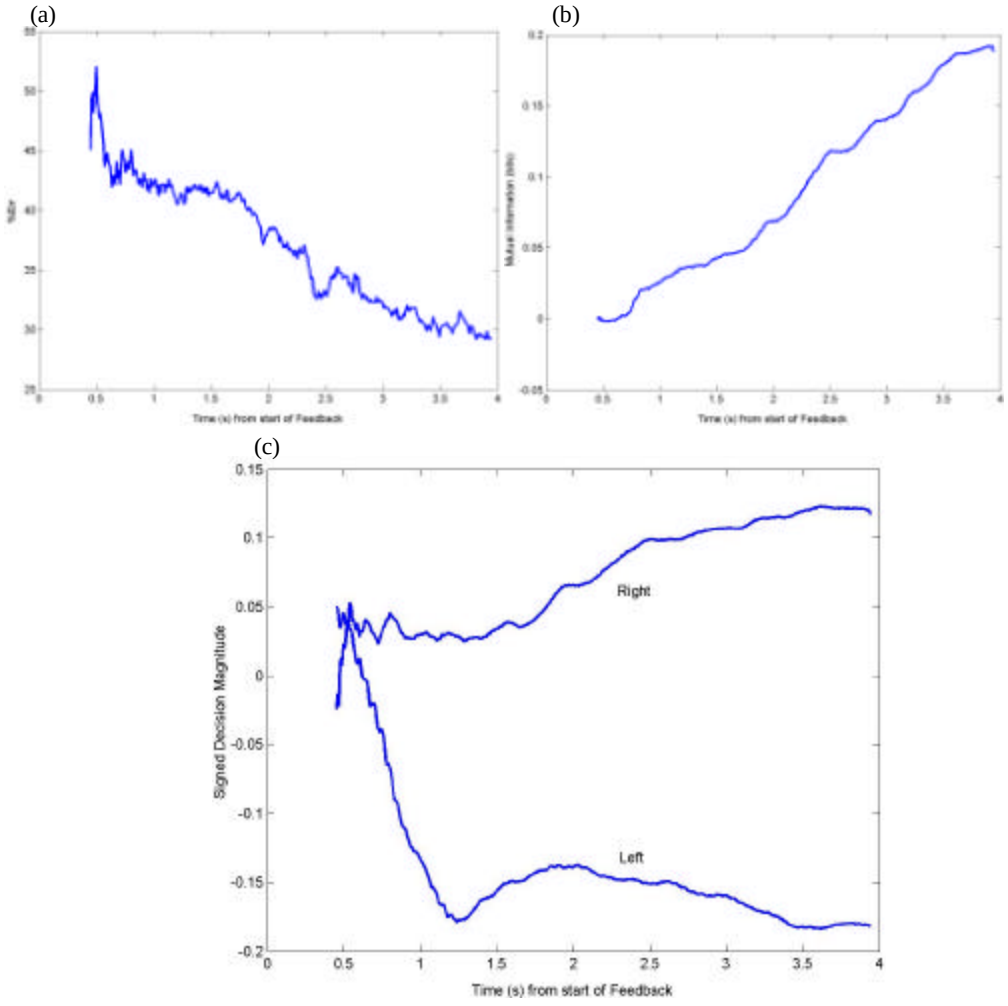

Figure 5-5 – SUBJECT A3: (a) Time course of percent error from 10-fold cross validation for the TEA approach using wavelet features, (b) the time course of the mutual information between the decision output and classes (c) time course of left and right decision outputs

Chapter 6

AAR vs. Wavelet: Comparison, Complimentary Information, and Fusion

In the previous two chapters we present two methods of motor-imagery analysis: wavelet and adaptive autoregression. Recall that both feature sets are classified using the TEA framework discussed in section 4.3.2. Herein, we refer to the wavelet method as WAV and the adaptive autoregressive method as AAR; these terms encompass both the time-frequency feature extraction method as well as the TEA classification framework, unless otherwise specified. In this chapter we compare the performance of the two methods for all three subjects. The strengths and weaknesses of each algorithm are highlighted with reference to results. Based

on these findings and empirical evidence of complimentary information, we propose a fusion scheme that improves performance.

6.1 A Comparison of Performance

For convenience and ease of comparison the results for AAR and WAV from Chapter 4 and Chapter 5 are presented on the same plots in Figure 6-1, Figure 6-2, and Figure 6-3, for subjects C1, B2, and A3, respectively. WAV outperforms AAR in terms of accuracy and mutual information for subjects C1 and B2. Not only does WAV achieve a lower minimum error, but also the error drops off more quickly. Similarly, the mutual information increases more quickly. This has significant implications in BCI applications where the system should ascertain the intention of the user as quickly as possible. Despite better performance for WAV for these two subjects, its Chernoff bound is higher than AAR (see (c) of Figure 6-1 and Figure 6-2). Recall that the Chernoff plots in these figures are representative of the data prior to application of the TEA framework since the bound is used to weight decisions made over time. The fact that the classes in WAV are less separable than AAR *without* TEA (indicated by the Chernoff bound) and more separable than AAR *with* TEA (indicated by the mutual information) suggests that WAV benefits from TEA considerably more.

For subject A3, AAR outperforms WAV. In general the mutual information for subject A3 is considerably lower than the other two subjects. The signal-to-noise ratio (SNR) is highly related to the mutual information (see Appendix B). An interesting notion that these results may support is that for noisier signals, AAR outperforms WAV because of its inherent robustness to noise.

In general AAR seems to have larger average decision magnitudes as depicted in (d) of Figure 6-1, Figure 6-2, and Figure 6-3. This characteristic is desirable, however, the mutual information for AAR is lower for subjects C1 and B2 indicating that the signed decision magnitude for AAR has high variability, which is undesirable.

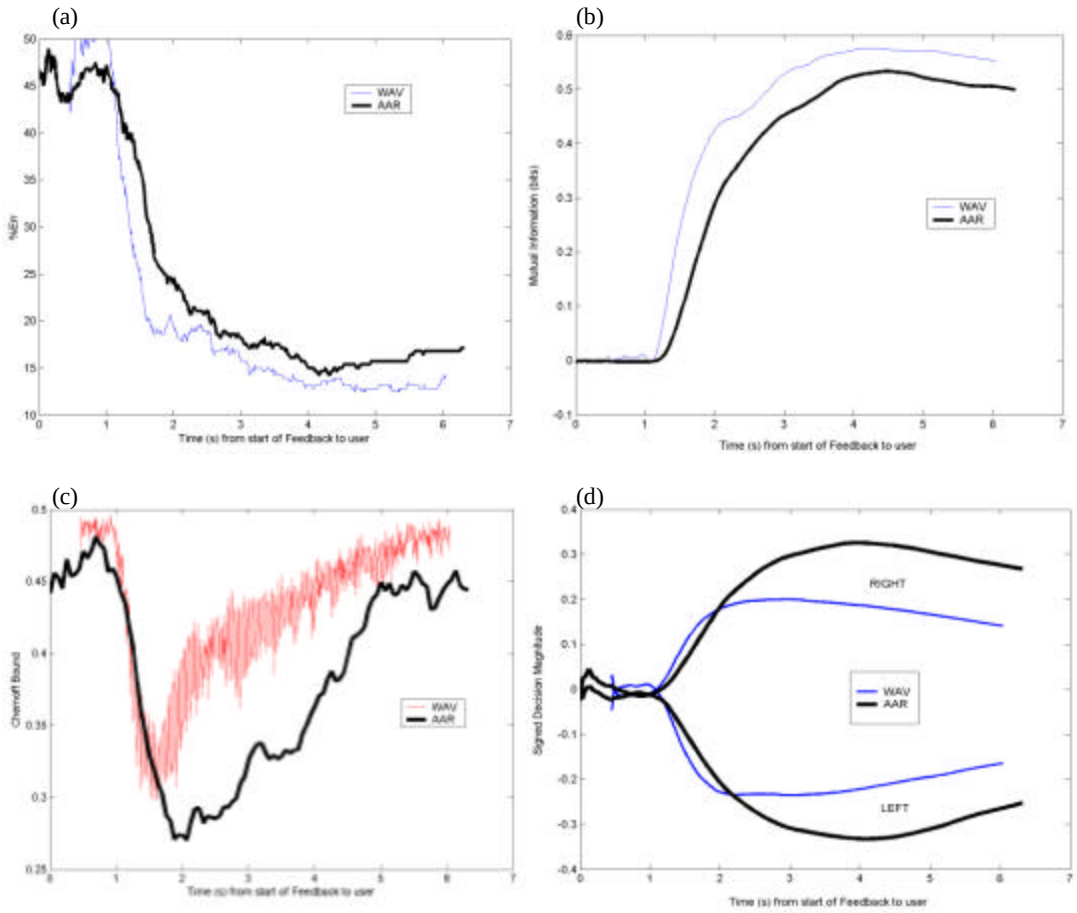

Figure 6-1 – SUBJECT C1: (a) (a) Time course of percent error from 10-fold cross validation for the AAR (thick line) and WAV (thin line), (b) the time course of the mutual information between the decision output and classes for AAR (thick line) and WAV (thin line), (c) time course of the Chernoff bound for AAR (thick line) and WAV (thin line) used for TEA decision weighting, (d) the signed distance to classification boundary for AAR (thick line) and WAV (thin line) for Left (negative) and Right (positive)

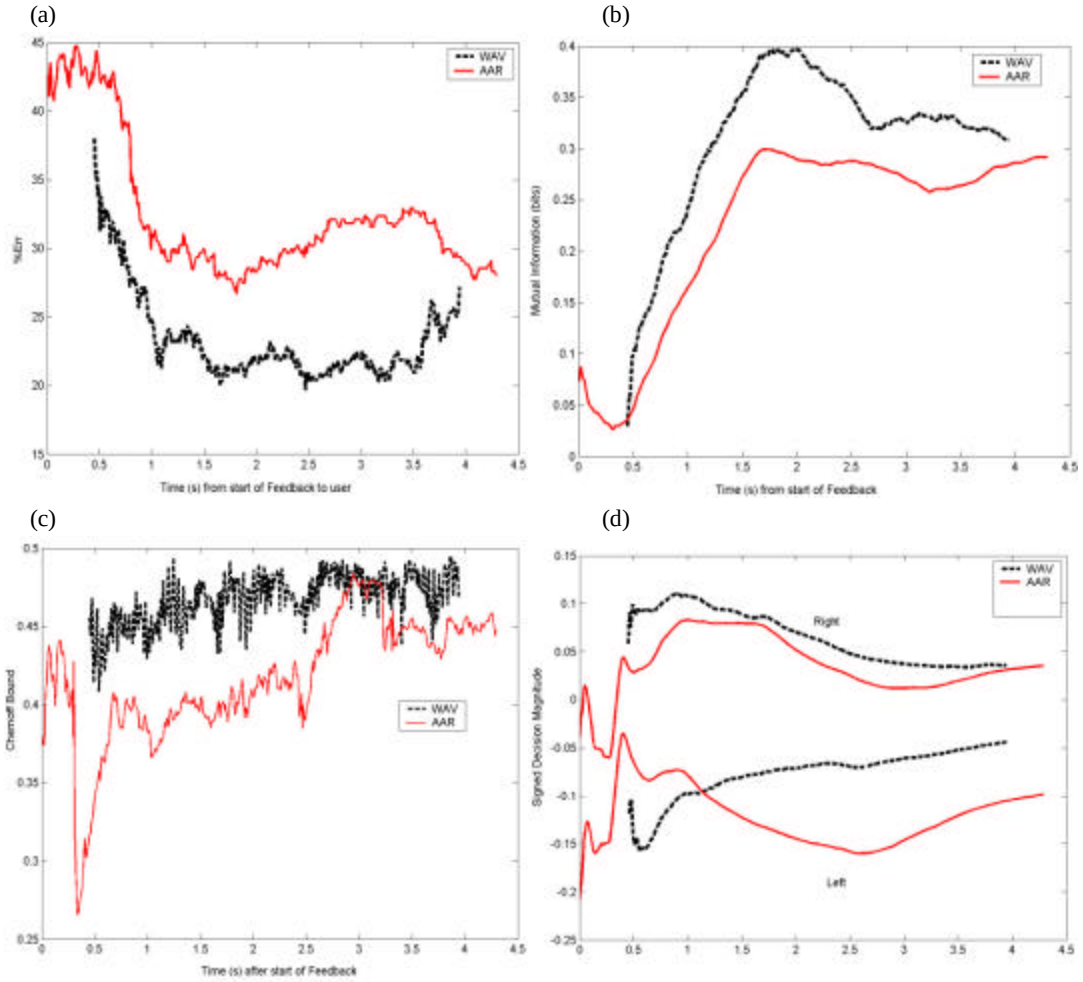

Figure 6-2 – SUBJECT B2: (a) (a) Time course of percent error from 10-fold cross validation for the AAR (thick line) and WAV (thin line) feature sets, (b) the time course of the mutual information between the decision output and classes for AAR (thick line) and WAV (thin line) methods, (c) time course of the Chernoff bound for AAR (thick line) and WAV (thin line) used for TEA decision weighting, (d) the signed distance to classification boundary for AAR (thick line) and WAV (thin line) for Left (negative) and Right (positive)

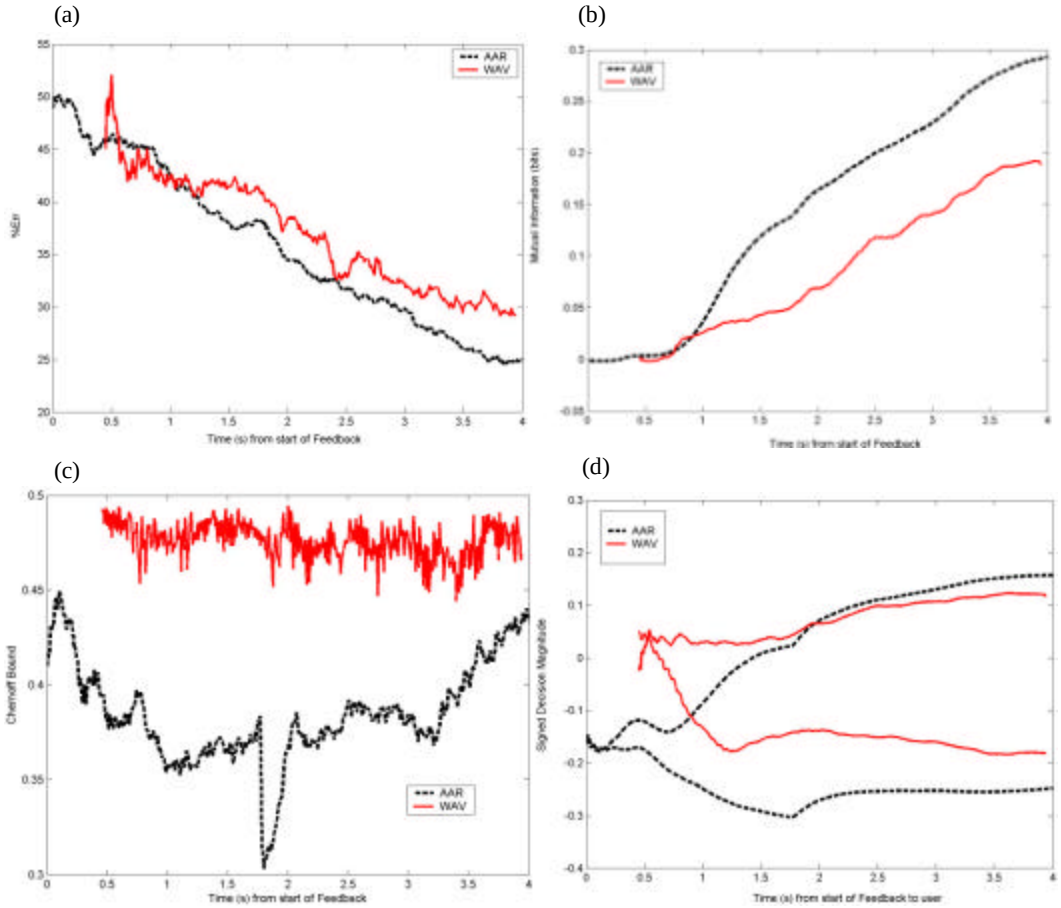

Figure 6-3 – SUBJECT A3: (a) (a) Time course of percent error from 10-fold cross validation for the AAR (thick line) and WAV (thin line) feature sets, (b) the time course of the mutual information between the decision output and classes for AAR (thick line) and WAV (thin line) methods, (c) time course of the Chernoff bound for AAR (thick line) and WAV (thin line) for TEA decision combining, (d) the signed distance to classification boundary for AAR (thick line) and WAV (thin line) for Left (negative) and Right (positive)

6.2 Complimentary Information in the Feature Sets

The idea of combining information provided by wavelet and AAR features is introduced in section 2.2. In section 3.1, we informally eluded to some general theoretical motivations for considering fusion. In this section, we seek experimental evidence that each feature set contributes some information about the signal that the other does not. As an initial investigation into this idea, we assess complimentary classification. In particular, we consider what percentage of misclassified trials in one feature set the other classifies correctly. Figure 6-4 depicts this analysis throughout the duration of the trial for each subject.

For the initial part of the trial classification for both feature sets is not much better than random, so the percentages during this time give little meaning. After this time there is consistent indication that a percentage of misclassified data is correctly classified by the other feature set. It is not surprising that this percentage tends to be lower for the higher performing method (WAV for C1 and B2, AAR for A3). The period of particular interest is just after the better-than-random classification. This earlier time slot is of greater interest because of two reasons: (1) we are primarily interested in improving performance near the beginning of the trial

to decrease the response time of the system, and (2) the algorithms are achieving lower accuracies during this time and the percentages of correctly classified trials by the other method, if exploited, translate to significant improvements in accuracy.

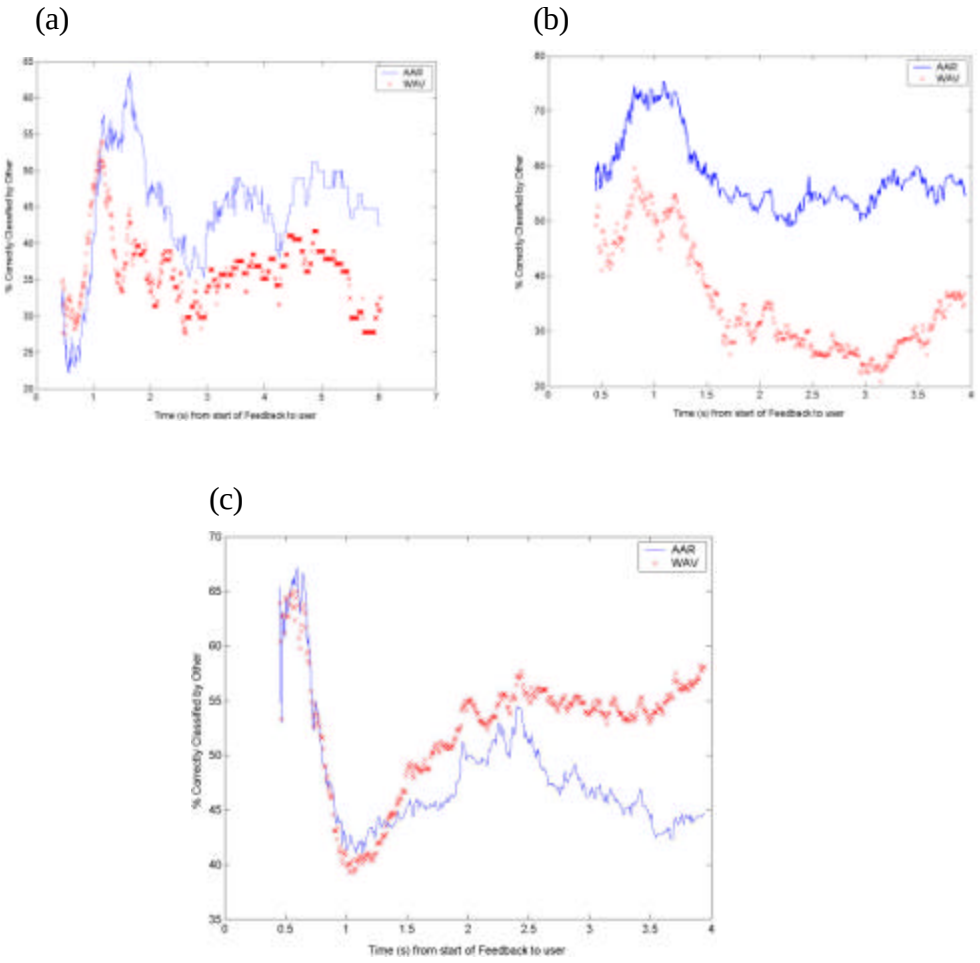

Figure 6-4 - % of misclassified data that was correctly classified by the other feature set for each of AAR and Wavelet: (a) subject C1, (b) subject B2, and (c) subject A3

6.3 Motivation for a Fusion Approach

Multi-resolution time frequency analysis is an important tool to address the principle of uncertainty between frequency and time [60]. This is the primary motivation for wavelet analysis. In contrast, the AAR method does not adequately address the problem of time-frequency resolution. In section 4.2.3, we discuss the connection between time-frequency resolution and the AAR model through two Kalman Filtering parameters: (1) the AAR model order *p* and (2) the update coefficient *uc*. In the formulation of the AAR estimation algorithm employed in this study, these parameters are constant throughout the entire trial. Therefore, there is only one time-frequency resolution in this approach and it has the limitation of not being optimally set for both frequencies of interest, a and ß.

An advantage of the AAR model is its robustness with regard to noise. As a noise driven model, it has an inherent quality of extracting information even in noisy environments. In this study we see evidence of this in the Chernoff bound of Figure 6-1 (c), Figure 6-2 (c), and Figure 6-3 (c). Although superior in overall performance for subjects C1 and B2, WAV is not robust to the high variability of the signal without the TEA framework. In contrast the AAR Chernoff bound is slow and steady inferring certainty more consistently.

Another strong indication of the potential for fusing the two feature sets is the difference in the misclassified trials described in section 6.2. The sum of all of these factors offers good reason for investigating fusion. In the rest of this chapter we propose a fusion approach and assess its performance for all three subjects.

6.4 Proposed Method of Fusion

The goal of the fusion method is to combine information provided by both the AAR and WAV feature sets. Since the TEA framework effectively extracts information from the features over time, we continue to employ the framework within the fusion scheme. Thus, each feature set AAR and WAV are classified using the TEA method as described in section 4.3.2. The output of each TEA classifier is the signed magnitude distance $D(n)$ (see equation (4.27)). Therefore, the output not only reflects the classifier's decision but also the confidence in its decision. This information may prove to be useful in fusion, since each feature set has different strengths to offer at different times during the trial. By training a classifier to learn when and how to trust the decisions by these two experts, we may enhance the results. The idea is illustrated in Figure 6-5.

For the fusion classifier we infer a 2-D Gaussian distribution on the output of the two classifiers. Note that the Gaussian assumption for $D(n)$ is consistent with

previous assumptions made in section 4.4.1 when evaluating the mutual information for the classifier output. Therefore, in the fusion stage we employ the probabilistic classifier defined in section 4.3.2, more specifically in equations (4.23) and (4.24).

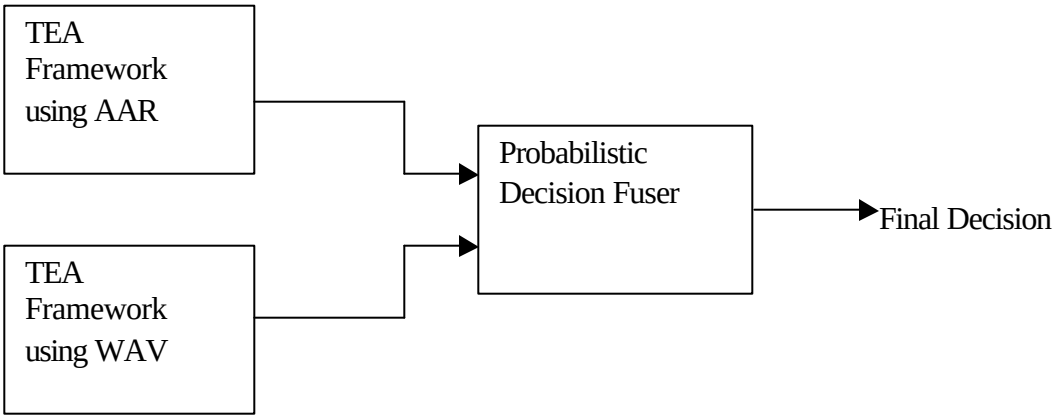

Figure 6-5 – Proposed fusion scheme

6.5 Results

To measure the performance of the fusion method, referred to as FUS herein, we employ the mutual information, percent error, and signed decision magnitude - all using 10-fold cross-validation. The results for subjects C1, B2, and A3 are depicted in Figure 6-6, Figure 6-7 and Figure 6-8, respectively.

For all subjects FUS improved the mutual information; subject C1's mutual information increased by 20%. Recall from section 4.4.1 the significance of this measure. It has strong implications for the overall reliability and consistency of the BCI. Furthermore, notice the rather large increase in the signed decision magnitude for all of the subjects using FUS. The large improvement in both of these measures indicates that the BCI output is more relevant to the classes. For this reason, there are benefits in fusing WAV and AAR in terms of the overall reliability and consistency of the system. However, the percent error for FUS in subjects C1 and B2 is similar throughout the trial to WAV (see (b) of the figures). It is important to recognize that the error is calculated from cross-validation and is specific to this data set. The fact that the mutual information for FUS is significantly higher implies that, in general, FUS should perform as good as or better than the best of WAV and AAR.

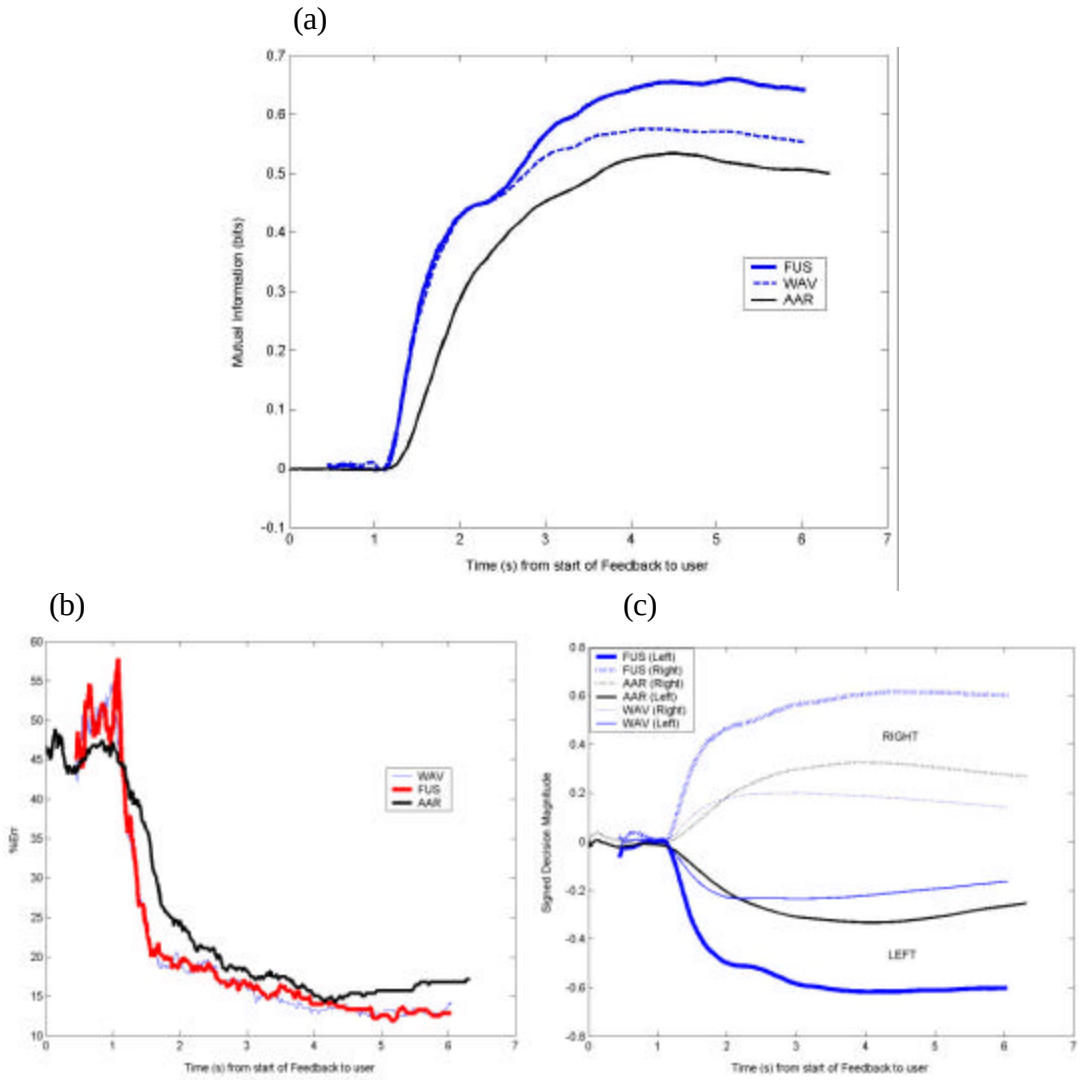

Figure 6-6 – SUBJECT C1: comparison of time-course performance for AAR, WAV, and FUS methods for (a) the mutual information, (b) the percent error, and (c) the signed decision magnitude.

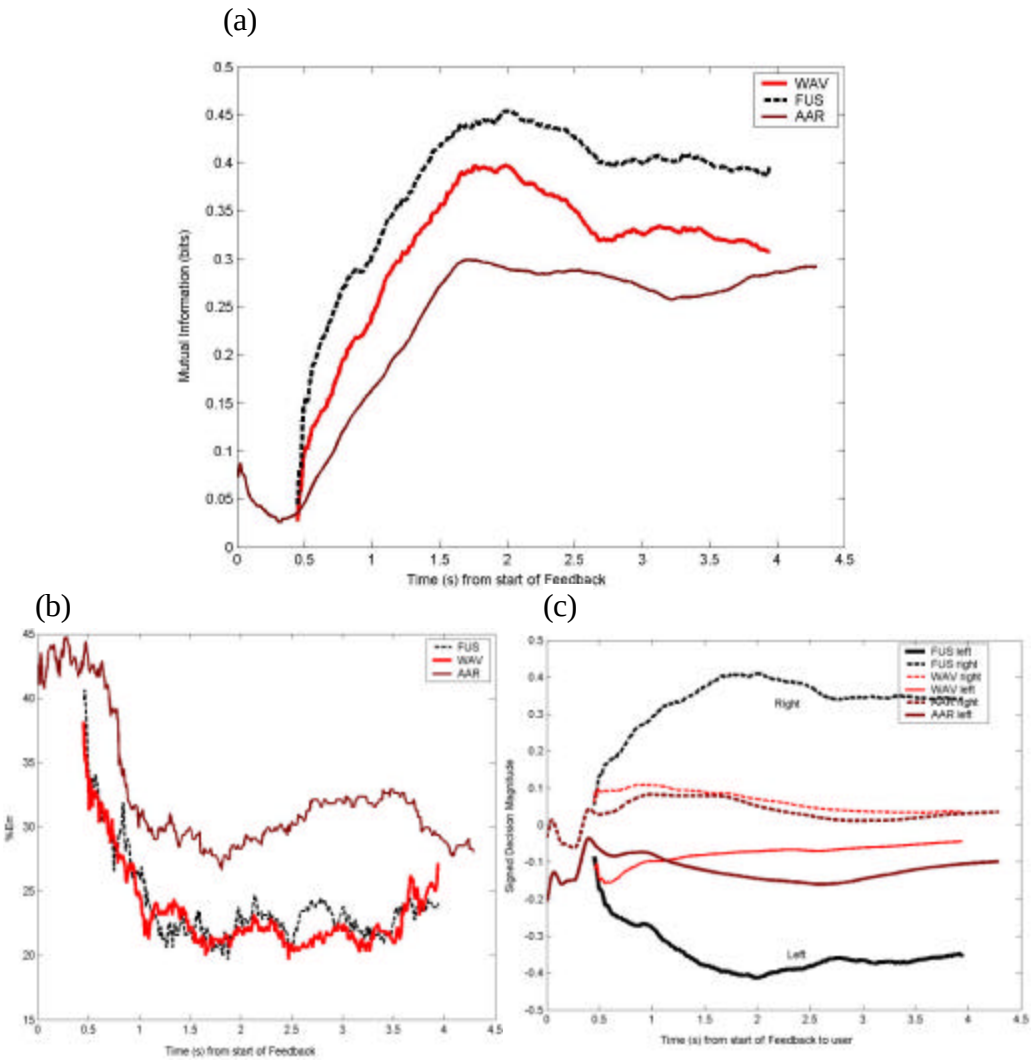

Figure 6-7 – SUBJECT B2: comparison of time-course performance for AAR, WAV, and FUS methods for (a) the mutual information, (b) the percentage error, and (c) the signed decision magnitude.

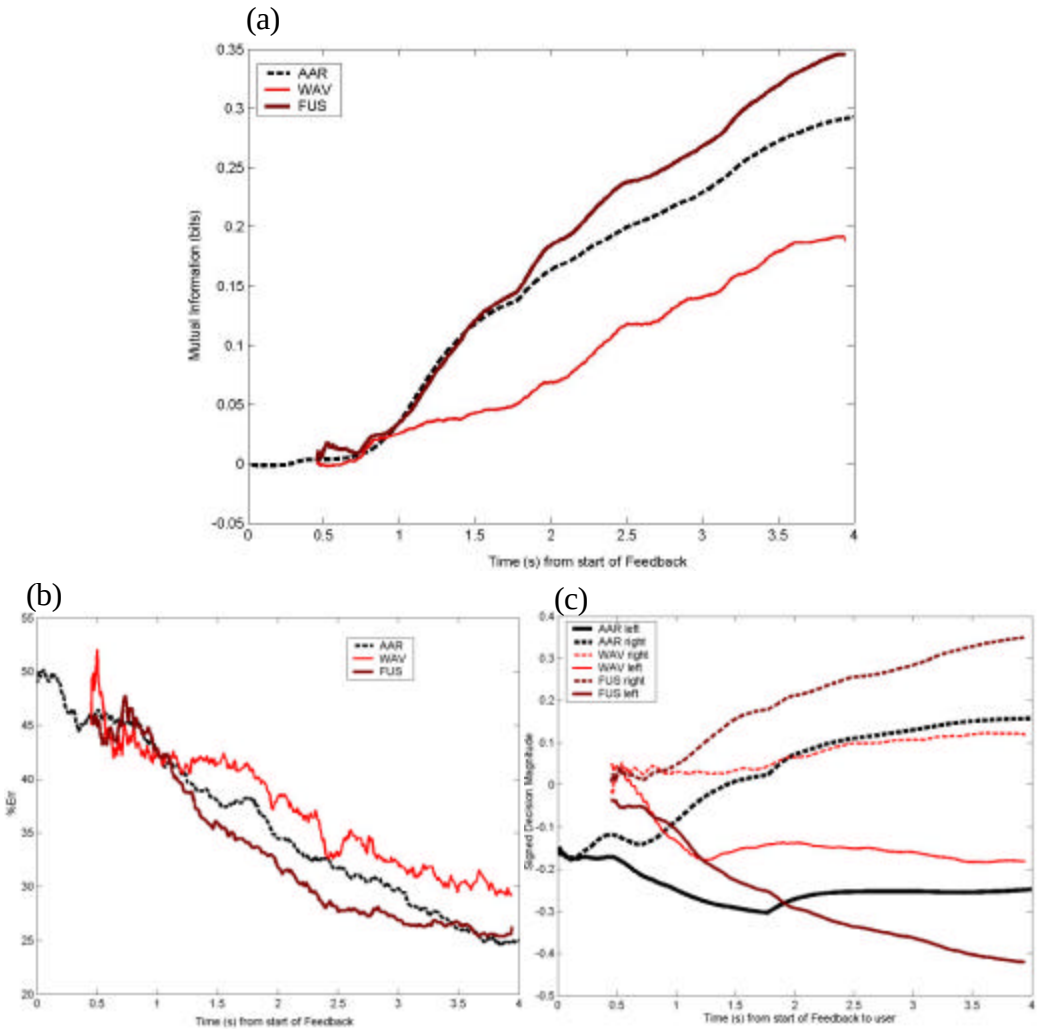

Figure 6-8 – SUBJECT A3: comparison of time-course performance for AAR, WAV, and FUS methods for (a) the mutual information, (b) the percent error, and (c) the signed decision magnitude.

Chapter 7

Conclusion

In this chapter we summarize the findings and contributions in this research. Furthermore, the implications of the work in the field of BCI at large are discussed. Finally, we consider possible future work for extending and further investigating the contributions in this thesis.

7.1 Contributions to Motor-Imagery Analysis in BCI

There are three major contributions in this work to the field of EEG motor-imagery analysis. They are each discussed below.

TEA Framework and AAR features

We proposed the TEA framework for AAR features and in doing so demonstrated considerable performance improvement over conventional classification approaches. The key to this approach is its ability to ascertain certainty in decisions made throughout the trial and to consider decisions made in the past accordingly. Unlike conventional classification techniques, TEA makes use of prior knowledge obtained from the training data to ignore ambiguous parts of the trial that would otherwise degrade performance.

AAR vs. Wavelet

With regard to EEG motor-imagery analysis there has been very little research to compare AAR and WAV under the same study. In particular, to the best knowledge of this research they have never been compared within the TEA framework, as application of TEA to AAR features is proposed in this work. We apply TEA to both feature sets and demonstrate that WAV outperforms AAR for two out of the three subjects. Interestingly, the subject for whom AAR outperformed WAV had lower signal-to-noise ratio; this implies that AAR is more robust to noise than WAV.

Fusion of Wavelet and AAR

We presented compelling reasons for fusing WAV and AAR to improve the performance of a BCI; there are two major reasons: (1) WAV offers multi-resolution time-frequency analysis to better address the principle of uncertainty; and (2) as a noise driven model, AAR handles the heavily corrupted EEG quite well. We propose a scheme that applies the TEA framework to each feature set separately and uses a third classifier to fuse these decisions. This approach results in higher mutual information between the BCI output and the classes. Therefore, we may conclude that fusion of WAV and AAR produces more relevant output to the motor-imagery classification task. This in turn yields a more reliable and consistent BCI.

7.2 Future Work

Although in this research we have been able to improve the mutual information of the BCI using fusion, there is potential to take more advantage of complimentary information provided by WAV and AAR. This is perhaps most notable in the earlier part of trials: when classification is significantly better than random and there is a large proportion of correctly classified data that the other method misclassifies, yet there is little improvement in the performance of FUS. This suggests that more research is worthwhile into more sophisticated fusion

schemes that take full advantage of the available complimentary information. Another approach worth investigating is fusing the strengths of wavelet analysis and autoregressive analysis at a more fundamental level. If AAR's strength is its robustness to noisy signals and the method of wavelet analysis excels in its multi-resolution approach to time-frequency investigation, perhaps the formulation of a multi-scale AAR approach could be a good alternative. In such an approach, the AAR features would be extracted in several time-frequency resolutions to get more accurate spectral information.

In this work we have assumed Gaussian distributions for the feature sets for computational convenience and because it has been used with some success in previous work. However, it is expected to be erroneous to some extent because EEG amplitudes are bounded and for many subjects marginal distributions have been found to be asymmetric [58]. Perhaps a more characteristic distribution can be inferred using a Gaussian Mixture Model (GMM) and estimating its parameters using the expectation-maximization (EM) algorithm. In such an approach, the derivation of the Chernoff bound (the measure of certainty) used in this work would not be applicable. Some other measure of certainty would have to be derived to use the TEA framework.

Asynchronous communication protocols have obvious benefits in BCI; they

allow the user to interact with more spontaneity and flexibility. Other future work should consider the methods proposed in this work in an asynchronous paradigm. In such an approach a motor-event could be detected using methods in [39], followed by classification using algorithms in this research.

References

[1] Y. Daabaj, "An evaluation of the usability of human-computer interaction methods in support of the development of interactive systems," in *Proc. IEE Int. Conf. System Sciences,* Hawaii, pp. 1830-1839, 2001.

[2] J.A. Landay, "Informal user interface for natural human-computer interaction," *IEEE Intell. Syst.,* vol. 13, pp. 14-16, 1998.

[3] V.I. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 19, pp. 677-695, July 1997.

[4] T. Ebrahimi, J. Vesin and G. Garcia, "Brain-Computer Interface in Multimedia Communication," IEEE Signal Processing Magazine, vol. 20, no.1, pp. 14-24, 2003.

[5] R.C. Ficke, *Digest of data on persons with disabilities,* Washington, DC: US Department of Education, National Institute on disability and Rehabilitation research, 1991.

[6] R. Scherer, G.R. Muller, C. Neuper, B. Graimann, G. Pfurtscheller, "An asynchronously controlled EEG-based virtual keybard: improvement of the spelling rate," IEEE Trans. Neural Sys. And Rehab. Eng., vol 12, no. 2, pp.

979-984, June 2004.

[7] J. d. R. Millian, F. Renkens, J. Mourino, and W. Gerstner, "Non-invasive brain-actuated control of a mobile robot," IEEE Transactions on Bomedical Engineering, vol. 51, no. 6, pp 1026-1033, 2004.

[8] A. Pino, E. Kalogeros, E. Salemis, G. Kouroupetroglou, "Brain Computer Interface Cursor Measures for Motion-impaired and Able-bodied Users," Proceddings of HCI International 2003: The 10th International conference on Human-Computer Interaction, June 22-27, 2003 Crete, Greece, pp. 1462-1466.

[9] C. V. Brewer. *The Organization of the Central Nervous System*. London: Heinemann Inc., 1961.

[10] F. Reik, D. Warland, Rob de Ruyter van Steveninck and W. Bialek, *Spikes – Exploring the neural code*, Cambridge: MIT Press, 1997.

[11] M. Steriade, P. Gloor, R. Llinas, Lopes da F. Silva, and M. Mesulam, "Report of IFCN Committee on Basic Mechanisms. Basic mechanisms of cerebral rhythmic activities," *Electroenchephalogr Clin Neurophysiol.*, vol. 76, no. 6, pp. 481-508, 1990.

[12] H. Berger, "Uber das Elektroenzephalogramm des Menschen," Arch. Psychiat. Nervenkr, vol. 87, pp. 527-570, 1929.

[13] J. Evans and A. Abarbanel, *Quantitative EEG and Neurofeedback*, Toronto: Academic Press, 1999.

[14] J.J. Vidal, "Real-time detection of brain events in EEG," *Proc. IEEE,* vol. 65, pp. 633-664, May 1977.

[15] T. P. Jung, S. Makeig, C. Humphries, T.W. Lee, M.J. McKeown, V. Iragui, and T.J. Sejnowki, "Removing electroencephalographic artifacts by blind source separation,"*Psychophisiology*, vol. 37, pp. 163-178, 2000.

[16] S. Vorobyov and A. Cichoki, "Blind noise reduction for multi-sensory signals using ICA and subspace filtering, with application to EEG analysis," *Biol. Cybern.,* vol. 86, pp. 293-303, 2002.

[17] H.H. Jasper, "The Ten-Twenty electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 371-375, 1958.

[18] D. Scot, *Understanding EEG*, London: Gerald Duckworth & Co. Ltd., 1976.

[19] C. Epstein and M. Epstein, *Introduction to EEG and Evoked Potentials*, New York: J. B. Lippincott Company, 1983.

[20] J. Vidal, "Towards direct brain-computer communications," *Annu. Rev. Viophys. Bioeng.* Vol 2, pp. 157-180, 1973.

[21] E. Sutter, "The brain response interface: communication through visually,-induced electrical brain responses" *Journ. Of Microcomputer Appl.,* vol. 15, pp. 31-45, 1992.

[22] M. Middendorf, G. McMillan, G. Calhoun, K. Jones, "Brain-computer interfaces based on steady-state visual evoked resonse," *IEEE Trans.*

Rehabil. Eng., vol. 8, pp. 211-213, 2000.

[23] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked correlates of stimulus uncertainty," *Science*, vol. 150, pp. 1187-1188, 1965.

[24] E. Donchin and D. Smith, "The contingent negative variation and the late positive wave of the average evoked potential," *Electroenceph. Cline. Neurophys.*, vol. 29, pp. 201-203, 1970.

[25] E. Donchin, K. Spence, R. Wijesinghe, "The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehabil. Eng.* Vol. 8, pp. 174-179, 2000.

[26] H. Serby, E. Yom-Tov, and G. Inbar, "An Improved P300-Based Brain-Computer Interface," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 13, no. 1, pp. 89-98, 2005.

[27] B. Roder, F. Rosler, E. Hennighausen, F. and Nacker, "Event-related potentials during auditory and somatosensory discrimination in sighted and blind human subjects," *Brain Res. Cogn. Brain Res.*, vol. 4, pp. 77-93, 1996.

[28] N. Birbaumer, T. Elbert, A. Canavan, and B. Roch, "Slow cortical potentials: their origin, meaning, and clinical use", I*Brain and Behaviour past, present and future*, Tilburg: Tilburg University Press, pp. 25-39, 1997.

[29] N. Birbaumer, A. Kubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor, "The thought translation device (TTD) for completely paralyzed patients," *IEEE Trans. Rehabil. Eng.*, vol. 8, pp. 190-192, 2000.

[30] Fisch B., *Fisch and Spehlmann's third revised and enlarged EEG Primer*, Amseterdam: Elsevier, 1999.

[31] G. Pfurtscheller and A. Arabibar, "Evaluation of even-related desynchronization preceding and following voluntary self-paced movement," *Electroenceph. Clin. Neurophysiol.*, vol. 46, pp. 138-146, 1979.

[32] S. Roberts and W. Penny, "EEG-based communication: a pattern recognition approach," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 56-61, 2000.

[33] S. Roberts and W. Penny, "Real-time Brain Computer Interfacing: p preliminary study using Bayesian learning," *Medical & Biological Engineering and Computing*, vol. 38, no. 1, pp. 56-61, 2000.

[34] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication" *Proceeding of the IEEE,* vol. 89. no. 7, pp. 1123-1134, July 2001.

[35] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A.Schloegl, B. Obermaier, and M. Pregenzer, "Current Trends in Graz Brain-Computer Interface (BCI) Research," *IEEE Trans. Rehab. Eng.,* vol. 8, no. 2, pp. 216-219, 2000.

[36] J. Wolpaw, D. McFarland, G. Neat, C. Forneris, "An EEG-based brain-computer interface for cursor control," *Electroenceph. Clin. Neurophys.*, vol. 78, pp. 252-259, 1991.

[37] J. Wolpaw, D. McFarland, T. Vaughan, "Brain-computer interface research at the Wadsworth Center," *IEEE Trans. Rehabil.*, vol. 8, pp. 222-225, 2000.

[38] A. Schloegl, C. Neuper, G. Pfurtscheller, "Estimating the Mutual Information of an EEG-based Brain-Computer Interface," *Biomedizinische Technik*, vol. 47, pp. 3-8, 2002.

[39] S. G. Mason and G. E. Birch, "A Brain-Controlled Switch for Asynchronous Control Applications," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 10, pp. 1297-1307, 2000.

[40] G. Townsend, B. Graiman, and G. Pfurtscheller, "Continuous EEG Classification During Motor Imagery – Simulation of an Ansynchronous BCI," *IEEE Trans. Neural Sys. Rehabil. Eng.*, vol. 12, no. 2, pp. 258-265, 2004.

[41] R. Scherer, G. Muller, C. Neuper, B. Graimann, and G. Pfurtscheller, "An Asynchronously Controlled EEG-Based Virtual Keyboard: Improvement of the Spelling Rate," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 979-984, 2004.

[42] E. Curran, P. Sykacek, M. Stokes, S. J. Roberts, W. Penny, I. Johnsrude, and A. M. Owen, "Cognitive Tasks for Driving a Brain-Computer Interfacing System: A Pilot Study," *IEEE Trans. Neural Sys. Rehabil. Eng.*, vol. 12, no. 1, pp. 48-54, 2003.

[43] L. S. Lustick, B. Saltzberg, J. K. Buckley, R. G. Heath "Autoregressive model for simplified computer generation of EEG correlation functions," *Proceedings of the IEEE annual conference on engineering in medicine and biology*, vol. 10, New York, pp. 78-94, 1968.

[44] P.B. Fenwick, P. Mitchie, J. Dollimore, G. W. Fenton, "The use of the autoregressive model in EEG analysis," *Electroencephalogr of Clinical Neurophysiology*, vol. 29, no. 3, pp. 319-327, 1970.

[45] P. B. Fenwick, P. Michie, J. Dollimore, G. W. Fention, "Mathematical simulation of the electroencephalogram using an autoregressiver series," *International Journal of Biomedical Computing*, vol. 2, no. 4, pp. 281-307, 1971.

[46] W. Gersch, *Spectral analysis of EEG by autoregressive decomposition of time series," Mathematical Bioscience*, vol. 7, pp. 205-22, 1970.

[47] A. Schloegl, *The electroencephalogram and the adaptive autoregressive model: theory and applications*, Aachen: Shaker Verlag, Germany, 2000, Sections 1 and 2.

[48] M. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley and Sons, Hoboken, NJ, 1996, Ch 4.

[49] J. Y. Lai. *Technical Memorandum: Wake Vortex Velocity Tracking via Kalman Filtering*. Radar Systems Laboratory, Electrical and Computer Engineering, Clemson University, Slemson, SC.

[50] A. Schloegl, D. Flotzinger, G. Pfurtscheller, "Adaptive Autoregressive Modeling use for Single-trial EEG Classification," *Biomedizinische Technik*, vol. 42, pp. 162-167, 1997.

[51] A. Schloegl, K. Lugger, G. Pfurtscheller, "Using Adaptive Autoregressive Parameters for a Brain-Computer Interface Experiment," *Proceeding of the*

19th IEEE International Conference on Engineering in Medicine and Biology, pp. 1533-1535, Chicago, USA, 1997.

[52] K. Lugger, D. Flotzinger, A. Schloegl, M. Pregenzer, G. Pfurtscheller, "Feature extraction for on-line EEG classification using principal components and linear discriminants," *Medical & Biological Engineering & Computing*, vol. 36, pp. 309-314, 1998

[53] C. Neuper, A. Schloegl, and G. Pfurtscheller, "Enhancement of Left-Right Sensorimotor EEG Differences During Feedback-Regulated Motor Imagery," *Journal of Clinical Neurophysiology*, vol. 16, no. 4, 373-382, 1999.

[54] A. Schloegl and G. Pfurtscheller, "Considerations on Adaptive Autoregressive Modelling in EEG Analysis," *Proceedings of First International Symposium on Communication Systems and Digital Signal Processing*, 1998.

[55] G. Pfurtscheller, C. Neuper, A. Schloegl, and D. Lugger, "Separability of EEG Signals Recorded During Right and Left Motor Imagery Using Adaptive Autoregressive Parameters," *IEEE Transactions on Rehabilitation Engineering*, vol. 6, no. 3, pp. 316-325, 1998.

[56] A. Schlogl, C. Neuper, G. Pfurtscheller, "Estimating the Mutual Information of an EEG-based Brain-Computer Interface," *Biomediziniche Technik*, vol. 47, 2002, pp. 3-8, 2002.

[57] R. Hari and R. Salmelin, "Human cortica oscillations: A neuromagnetic view through the skull," *Trends Neurosci.*, vol. 20, pp. 44-49, 1997.

[58] S. Lemm, C. Schafer, G. Curio, "BCI Competition 2003 – Data Set 2: Probabilistic Modeling of Sensorimotor μ Rhythms for Classification of Imaginary Hand Movements," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1077-1080, 2004.

[59] S. Haykin and B. Van Veen, *Signals and Systems*, New York: John Wiley & Sons, Inc..

[60] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets Transforms, A Primer*, Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[61] R. Duda, P. Hart, and D. Stork, *Pattern Classification 2nd Edition*, New York: Wiley, 2001.

[62] C. W. Chen, M. S. Ju, and C. K. Lin, "Real-Time Identification of μ wave with Wavelet Neural Networks," *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering*, pp. 218-220, Capri Island, Italy, 2003.

[63] B. Graimann, J. E. Huggins, S. P. Levince, and G. Pfurtscheller, "Toward a Direct Brain Interface Based on Human Subdural Recordings and Wavelet-Packet Analysis," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 954-962, 2004.

[64] B. Blankertz, K. R. Muller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schloegl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroder, and N. Birbaumer. "The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials," *IEEE*

Transactions on Biomedical Engineering, vol. 51, no. 6, pp. 1044-1051, 2004.

[65] M. Varsta, J. Heikkonen, J. R. Millan, and J. Mourino, "Evaluating the performance of three feature sets for brain-computer interfaces with an early stopping MLP committee," 15th International Conference on Pattern Recognition, pp. 907-910, Barcelona, Spain, 2000.

[66] C. Babiloni, F. Carducci, F. Cincotti, P. M. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni, "Human movement-related potentials vs desynchronization of EEG alpha rhythm: A high-resolution EEG study," *NeuroImage*, vol. 10, pp. 658-665, 1999.

[67] H. Ramoser, J. Muller-Gerking, and G. Pfuurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transaction on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441-446, 2000.

[68] G. Dornhege, B. Blankertz, G. Curio, and K. R. Muller, "Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and muti-class paradigms," *IEEE Transactions on Biomedical Engineering* vol. 51, no 6, pp. 993-1002, 2004.

[69] S. Haykin, *Adaptive Filter Theory, 4th edition*, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[70] M. Arnold, W. Miltner, H. Witte, R. Bauer, C. Braun, "Adaptive AR Modeling of Non-stationary Time Series by means of Kalman Filtering," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 5, 553-562,

1998.

[71] A. Schloegl, S. Roberts, and G. Pfurtscheller, "A criterion for adaptive autoregressive models," *Proceedings of the 22nd IEEE International Conference on Engineering in Medicine and Biology*, pp. 1581-1582.

[72] S Chiappa and S. Bengio, "HMM and IOHMM Modeling of EEG Rhythms for Asynchronous BCI Systems," *European Symposium on Artificial Neural Networks*, pp. 985-992, 2004.

[73] C. E. Shannon, "The mathematical theory of communication", Bell Systems Technical Journal, vol 27, pp. 79-423, 623-656, 1949.

Appendix

A The Chernoff Bound

The material in this section is based on Chapter 2 of [61]. We may quantify the uncertainty of a classification system by estimating its probability of error. Figure A1 depicts the probability of error for a two-class classification problem with Gaussian distributions. Notice that the classification boundary minimizes the probability of error by always selecting the class with greater posterior; this is known as the Bayes optimal classifier. The posterior is given as

$$P(c_i \mid \mathbf{z}) = \frac{p(\mathbf{z} \mid c_i)P(c_i)}{p(\mathbf{z})}$$

and the decision rule is: *decide* c_1 if $P(c_1 \mid \mathbf{z}) > P(c_2 \mid \mathbf{z})$, *otherwise decide* c_2.

There are two ways in which a classification error may occur: (1) an observation $\mathbf{z}$ falls into the region *R1* and it belongs to *C2* or (2) an observation $\mathbf{z}$ falls into the region *R2* and belongs to *C1* (see Figure A1). For simplicity the figure illustrates the 1-D case but the idea extends to multi-dimensional problems. The probability of error may be given as

$$\begin{aligned} P(error) &= P(\mathbf{z} \in R_2, c_1) + P(\mathbf{z} \in R_1, c_2) \\ &= P(\mathbf{z} \in R_2 \mid c_1)P(c_1) + P(\mathbf{z} \in R_1 \mid c_2)P(c_2) \\ &= \int_{R_2} p(\mathbf{z} \mid c_1)P(c_1) + \int_{R_1} p(\mathbf{z} \mid c_2)P(c_2) \end{aligned} \qquad \text{(A1)}$$

The calculation for the probability of error, equation (A1), is difficult and computationally intensive. However, for the two-class case, we may calculate an

upper bound on the error known as the Chernoff bound. Given that another way to state the probability of error is

$$P(error \mid \mathbf{z}) = \min[\, P(c_1 \mid \mathbf{z}), P(c_2 \mid \mathbf{z})] \qquad \text{(A2)}$$

we may derive a bound on the error using the following inequality:

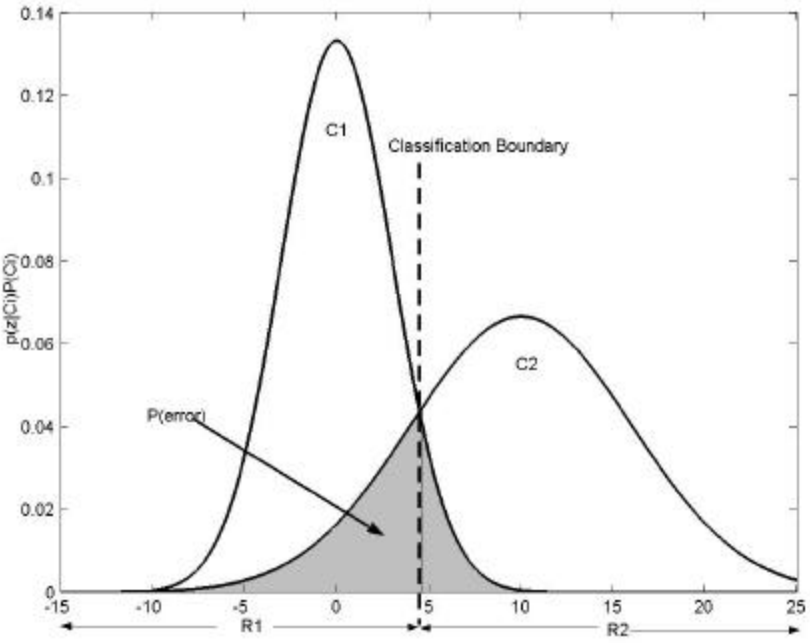

Figure A-1 – Probability of error (shaded region) for a two-class problem with Gaussian distributions.

$\min[\, a,b] \le a^b b^{1-b}$ for $a,b \ge 0$ and $0 \le b \le 1$. We may apply this inequality to equation A2:

$$p(error \mid \mathbf{z}) \le P^b(c_1) P^{1-b}(c_2) \int p^b(\mathbf{z} \mid c_1)\, p^{1-b}(\mathbf{z} \mid c_2) d\mathbf{z} \qquad \text{(A3)}$$

In the experiments of this research each class is equally likely to occur: $P(c_1) = P(c_2) = 0.5$. For the case where the conditional distributions are Gaussian, as they are throughout this research, the integral in equation A3 may be evaluated analytically, yielding

$$\int p^{b}(\mathbf{z}\mid c_1)p^{1-b}(\mathbf{z}\mid c_2)d\mathbf{z} = e^{-k(b)} \qquad \text{(A4)}$$

where

$$k(b) = \frac{b(1-b)}{2}(m_2 - m_1)^T[b\Sigma_1 + (1-b)\Sigma_2]^{-1}(m_2 - m_1).....$$

$$+\frac{1}{2}\ln\frac{|b\Sigma_1 + (1-b)\Sigma_2|}{|\Sigma_1|^{b}|\Sigma_2|^{1-b}} \qquad \text{(A5)}$$

The Chernoff bound on *P*(error) may be found by finding the value of ß that minimizes $e^{-k(ß)}$ (analytically or numerically) and substituting the result into equation (A3). A powerful feature of this approach is that the optimization of ß is in one-dimension regardless of the dimensionality of the feature space.

In general, the bound tends to be looser for extreme values of ß (ß→0, ß→1). Another well known bound on the probability of error is the Bhattacharyya bound [61], which is equivalent to simply setting ß = 0.5 in equation (A3). Since the optimal setting for ß tends to be mid-ranged values between 1 and 0 this is a computationally efficient guess. However, since the minimum value of $e^{-k(ß)}$ does not necessarily occur at ß = 0.5 the Chernoff bound is considered to be a tighter bound on *P*(*error*).

B Information Theory

The Shannon entropy of a random variable is a measure of the uncertainty and variability in the data and, therefore, quantifies its information capacity [73]. It measures the average amount of possibilities in the data; the more possibilities the more information that can be conveyed by the random variable. For a random variable x with probability density $p(x)$ the entropy is defined as

$$H(x) = -\int_{-\infty}^{\infty} p(x)\log_2(p(x))dx \tag{B1}$$

In this research we assume the random variables have Gaussian distributions,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{\frac{-(x-m)^2}{2\sigma^2}\} \tag{B2}$$

and the entropy formula may be reduced by the following

$$\log_2(p(x)) = \log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \log(\exp\{\frac{-(x-m)^2}{2\sigma^2}\})$$

$$\log_2(p(x)) = -\frac{1}{2}\log(2\pi\sigma^2) + \frac{-(x-m)^2}{2\sigma^2}\log(e)$$

Substituting the above logarithm into equation (B1) we get

$$H(x) = \frac{1}{2}\log(2\pi\sigma^2)\int_{-\infty}^{\infty} p(x)dx + \frac{\log(e)}{2\sigma^2}\int_{-\infty}^{\infty} p(x)(x-m)^2 dx$$

$$H(x) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log(e)}{2}$$

$$H(x) = \frac{1}{2}\log(2\pi\sigma^2 e). \tag{B3}$$

For the task of assigning a random variable x to a class c, we may quantify the relevance of x to c using the mutual information. This is a measure of the reduction in entropy between x and the class conditional x. It is defined as

$$I = H(x) - H(x \mid c) \tag{B4}$$

where $H(x|c)$ is the conditional entropy and is defined as

$$H(x \mid c) = \sum_{c} \int_{x} p(x,c) \log_2 (p(x \mid c) dx \tag{B5}$$

In this application we infer a Gaussian distribution on p(x|c) and have two classes L (left) and R (right) that are equally likely. Using a derivation similar to the one used to obtain equation (B3) the mutual information defined in (B4) can be expressed as

$$I = 0.5 * \log_2 \{ \frac{s_x^2}{0.5(s_{L_L}^2 + s_R^2)} \} . \tag{B6}$$

where s_L^2 and s_R^2 are the class conditional variances of x for class L and R, respectively.

Signal-to-Noise Ratio

For a given process with noise, such as the signed decision magnitude $D(n)$, it can be decomposed into a signal process y and a noise process v such that

$$D(n) = y(n) + v(n) \tag{B7}$$

$$y = N(m_y, s_y^2) \tag{B8}$$

$$v = N(m_v, s_v^2) \tag{B9}$$

$$\sum_{n} (y - u_y)(v - u_v) = 0 \tag{B10}$$

For independent Gaussian distributions the variance s_D^2 of D is

$$s_D^2 = s_y^2 + s_v^2 \tag{B11}$$

We may define the signal-to-noise ratio as

$$SNR = \frac{s_y^2}{s_v^2} = \frac{s_D^2}{s_v^2} - 1 \tag{B12}$$

The noise variance s_v^2 is the average within-class variance and, therefore, has the following relationship with mutual information according to equation (C6)

$$I = 0.5 * \log_2 \{1 + SNR\} . \tag{B13}$$

Abbreviations

AAR – Adpative Autoregression: this term is not only used to describe the signal processing technique, but also the Brain-Computer Interfacing approach that uses AAR parameters as features.

BCI – Brain Computer Interface

CONV – Conventional: refers to a method that is commonly used for classification of AAR parameters in BCI and is defined in section 4.3.1

FT – Fourier Transform

LDA – Linear Discriminant Analysis

FUS – Fusion: refers to the BCI method proposed in this thesis that involves fusion of AAR and Wavelet features

WAV – Wavelet: refers to the BCI method that uses wavelet features

WT – Wavelet Transform