Towards Adaptive User Interfaces using Real Time fNIRS

A dissertation

submitted by

Audrey Girouard

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

August 2010

Advisor: Prof. Robert J.K. Jacob

Abstract

Enhancing user experience is a constant goal for human computer interaction (HCI) researchers, and the methods to achieve this goal are widespread, from changing the properties of the interface to adapting the task to the user's ability level. By sensing user's cognitive states, such as interest, workload, frustration, flow, we can adapt the interface immediately to keep them working optimally. This new train of thought in the brain computer interfaces community considers brain activity as an additional source of information, to augment and adapt the interface in conjunction with standard devices, instead of controlling it directly with the brain.

To obtain measures of brain activity, I adopt the relatively less-explored brain sensing technique called functional near-infrared spectroscopy (fNIRS), a safe, non-invasive measurement of changes in blood oxygenation. This dissertation presents a body of technologies and tools that enable the use of real time measures of cognitive load for adaptive interfaces, to support the thesis that fNIRS is an input technology usable in conventional HCI contexts, especially when applied to the general, healthy public as an additional input.

First, I discuss the practicality and applicability of the technology in realistic, desktop environments. Our work shows that fNIRS signals are robust enough to remain unaffected by typing and clicking but that some facial and head movements interfere

with the measurements. Then, I investigate the use of fNIRS to obtain meaningful data related to mental workload. My studies progress from very controlled experiments that help us identify centers of brain activity, to experiments using simple user interfaces, showing how this technique may be applied to more realistic, complex interfaces. Our first study distinguishes levels of workload and interaction styles, and our second differentiates levels of game difficulty. Statistical analysis and machine learning classification results show that we discriminate well between subjects performing a mentally demanding task or resting, and distinguish between two levels with some success. Finally, I present a real time analysis and classification system that can communicate user cognitive load information to an application. I categorize adaption of interfaces with brain as an input, and propose a series of adaptations possible using our system.

Acknowledgements

There are many people that have helped me complete this thesis, and I thank them for all their support.

To my advisor Rob Jacob, for constant encouragements, guidance, insights, and advice, and for keeping me on track, especially every time I overthought problems;

To Francois Lalonde, wonderful husband and partner, for supporting me in everything I do, and for telling the world that my research consists of reading minds (which it is not);

To Erin Solovey, for all the wonderful collaboration and friendship throughout the last few years, for always understanding me and my ideas;

To Krysta Chauncey, great friend and collaborator, for many coffee shop work-dates that have let me to complete this work;

To Eddie Aftandilian and Noah Smith, for moral support and because graduate school wouldn't have been the same without you;

To my family, Maman, Marc-André, Laurence, Francois Olivier, Papa, Hélène, Guy, Marc-André, for supporting me from a distance, trying to understand what I do, and doing a good job at pretending otherwise;

To the HCI group at Tufts, colleagues and alumni, in particular Evan Peck, Juan Carlos Montemayor Elosua, Doug Weaver, Margarita Parasi, Francine Lalooses, Wyatt Newport, Leanne Hirshfield, Rebecca Gulotta, Kelly Moran, Hadar Rosenhand, Orit Shaer, Michael Horn, and Michael Poor, for many brainstorming sessions, discussions, support and encouragements;

To Wendy Mackay, for many great conversations that made me think in depth about my work and HCI in general and to the In|Situ| research group, Michel Beaudoin-Lafon, Olivier Chapuis, Nicolas Masson, Aurélien Tabard, Olivier Chapuis, Olivier Bau, Stéphane Huot, Emmanuel Pietriga, for welcoming me into your research family and sharing good times;

To the biomedical engineering group at Tufts, Angelo Sassaroli, Sergio Fantini, Yunjie Tong, for your collaboration and for sharing your knowledge about fNIRS;

To the machine learning group at Tufts, Roni Khardon, Carla Brodley, Rachel Lomasky, Umaa Rebbapragada, and D. Sculley for their help with initial machine learning aspects of our work;

To Desney Tan at Microsoft Research for sparking a few ideas, and for his helpful inputs and encouragement;

To the members of my committee for helpful comments and suggestions, specifically to Roni Khardon and Tim Bickmore for your detailed review of my work.

In addition, I would like to acknowledge and thank the National Science Foundation (NSF Grant Nos. IIS-0414389 and BES-93840) and the Natural Sciences and Engineering Research Council of Canada for financially supporting this research.

Table of Contents

List of Figures

List of Tables

Chapter 1:

Introduction

Imagine a device embedded in a hat, or a cap, that could wirelessly transmit the user's cognitive state to their computer. How can it make use of this new information? What kind of change in the interface could that lead to? There are many types of interfaces that can use such information, and many ways to adapt them. For example, entertainment interfaces (such as games) could make use of the subject's affective and cognitive state by adapting the interface to keep the user engaged, and to elicit specific emotional responses.

Enhancing user experience is a constant goal for human computer interaction (HCI) researchers, and the methods to achieve this goal are widespread, from changing the properties of the interface to adapting the task to the user's ability level. Ideally, those modifications are done automatically, in real time, to obtain maximum benefit. By sensing different user properties, such as interest, workload, frustration, flow, we can adapt the interface immediately to keep them working optimally.

Although we can accurately measure task completion time and accuracy, measuring cognitive factors such as distraction, surprise or mental workload are typically limited to qualitatively observing users or administering subjective surveys to them. These surveys are often taken after the completion of a task, potentially missing valuable insight into the user's changing experiences throughout the task. They fail to capture internal details of the operator's mental state, and they are not available in real time to allow interface adaptation. Monitoring performance data could address some of these issues. However, user performance measures may miss context, as they don't reflect all of the user's activities, on or off the computer.

Therefore, new measurements and evaluation techniques that monitor user experiences are increasingly necessary. To address these issues, much current research focuses on developing objective techniques to measure in real time user states such as emotion, workload, and fatigue (Gevins & Smith, 2003; John, et al., 2004; Marshall, Pleydell-Pearce, & Dickson, 2003). Although this ongoing research has advanced user experience measurements in the HCI field, finding accurate, non-invasive tools to measure computer users' states in real working conditions remains a challenge.

Brain sensing and imaging techniques, primarily developed for clinical settings, have been powerful tools for understanding brain function as well as for diagnosing brain injuries or disorders. More recently, these devices have found uses outside of hospital and laboratory settings, and HCI researchers have begun to employ them to understand more about the user's cognitive state relative to the task at hand (e.g. Chen, Hart, & Vertegaal, 2008; Grimes, et al., 2008; Sjölie, et al., 2010). Technological advances and lower costs associated with the devices have opened a new research area, brain computer interfaces. This field is blooming: the ACM Conference on Human Factors in Computing Systems CHI 2008 workshop on Brain-Computer Interfaces for HCI and Games (Nijholt, et al., 2008) and the CHI 2010 workshop on psychophysiological user interaction called Brain Body and Bytes (Girouard, et al., 2010b) are evidence of that.

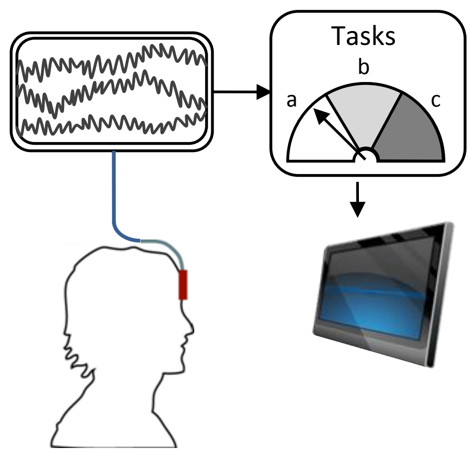

Figure 1-1. In traditional brain computer interfaces, brain activity is converted into predicted tasks, and is the only input to the interface.

Brain computer interfaces (BCI) are designed to use brain activity as an input for interfaces. Most of the current work in the field focuses on letting disabled patients communicate with their caretakers and their environment with the sole use of electroencephalography (Krepki, et al., 2007; Millán, et al., 2004; Wolpaw, et al., 2002)

(Figure 1-1). The resulting interfaces usually let the user select binary choices (e.g. yes/no), type or move a mouse, typically by comparing two brain signals together.

Communicating through traditional BCI systems is currently time consuming and mentally demanding. This paradigm requires a great deal of training from the user, to learn which type of signals to produce, and from the system, to learn which actions yield which signal. The interface is often slow to respond, especially in comparison with traditional input technologies (mouse and keyboard). Open research challenges in BCI concern the accuracy of such BCI systems (systems often misinterpret a user's intentions) and the information transfer rate of such systems, which are often lacking for use in real world settings.

A new train of thought in the BCI community considers brain activity as an additional source of information, to augment and adapt the interface instead of controlling it directly with the brain. The new methodology focuses on a broader group of users—the general population—for whom current BCIs are impractical because of their slow speed of transfer. Passive BCIs are designed to use brain activity as a new input modality, allowing the adaptation of the interface in real time according to the user's mental state (Cutrell & Tan, 2008), in conjunction with standard devices such as keyboards and mice (Figure 1-2). This type of BCI can capture intentional commands, but is best designed for implicit communication (Zander, et al., 2010).

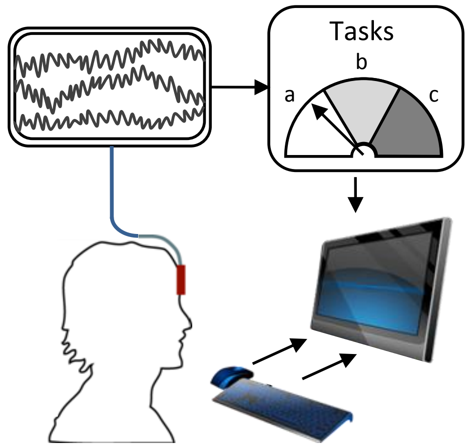

Figure 1-2. New brain computer interfaces uses brain activity as an additional input, in addition to mouse and keyboard.

While most BCI research is done in fields such as psychology and biomedical engineering, the study of passive BCIs could gain from the knowledge and expertise of the field of human computer interaction. HCI studies how to evaluate and improve the connection between human and computer, to create seamless interaction. I hope to contribute to this effort using the brain. Minnery and Fine (2009) point out in a recent *interactions* article that "only a small percentage of current neuroscience research is explicitly aimed at understanding aspects of HCI". With this thesis, I attempt to bridge part of the gap between two fields.

Neural signals can act as a complementary source of information when combined with conventional computer inputs such as the mouse or the keyboard. Work in this thesis illustrates this direction in BCI and shows how to move from controlled experiments exploring task-specific brain activity to a more general framework using mental activity to guide interface response. My work, grounded in the field of human-computer interaction, suggests the practicality and feasibility of using normal untrained brain activity to inform interfaces.

The advantages of using brain sensing to adapt interfaces are numerous (Allanson & Fairclough, 2004). Brain activity is continuously available and does not intrude onto the operator's task, while behavioral triggers may be discrete and intermittent (Wilson & Russell, 2007). Measuring it passively doesn't require the user to perform additional tasks, and they are continuously available. Finally, there are many aspects of user state that are covert, "within the user which can only be detected with weak reliability by using overt measures" (Zander, et al., 2010), for instance using brain activity sensing.

The design challenges for such an unobtrusive, passive, real-time brain interface are considerable. As I seek improved interaction for all users, rather than only disabled users for whom brain input is a viable alternative to otherwise unavailable arm, leg, or other inputs, the goal is to design user interfaces that treat the brain activity as an additional input channel, rather than as the primary input. For example, the user operates a conventional interface with a mouse, and the interface responds not only to the explicit mouse inputs but also to the information measured from the user's brain, letting only critical emails through should the user be in a state of flow. In this case the challenge is to design a user interface that makes judicious, subtle, "lightweight" use of brain input, rather than using it to, for example, directly drive a cursor. I believe the present work points to the ultimate feasibility of such real time input in HCI.

In this thesis, I associate passive brain computer interfaces and healthy users. However, there are situations where non-disabled users might be interested active, or direct BCI, for instance to perform hands free operations. I recognize such situations, but I believe that integrating the brain as a passive input in user interfaces covers an explored region of the BCI research space.

While most BCIs use the electroencephalogram to measure brain activity, I adopt the relatively less-explored technique of functional near-infrared spectroscopy (fNIRS), a non-invasive measurement of changes in blood oxygenation, used to extrapolate levels of brain activation (Chance, et al., 1998; Izzetoglu, et al., 2004a). The fNIRS tool is safe, portable, non-invasive, and can be implemented wirelessly, allowing for use in real world environments (Izzetoglu, et al., 2004a). One of the main benefits of fNIRS is that the equipment imposes few physical or behavioral restrictions on the participant (Hoshi, 2009), as illustrated in Figure 1-3.

Figure 1-3. A participant wearing one fNIRS probe under a sports band.

Overall, fNIRS output offers potential as an additional parallel, lightweight, continuous input channel for users. This additional information from the brain could be used to improve the efficiency, effectiveness, or intuitiveness of the user's interaction with the machine as well as to provide new access methods for both healthy and disabled users. Previous work using fNIRS for BCI has explored the basic technology and demonstrated the feasibility of distinguishing mental state using such signals (Hirshfield, et al., 2009b; Izzetoglu, et al., 2004b; Luu & Chau, 2009). In this thesis I take this research program to a more advanced setting, developing method to analyze the signal in real time and showing how this can be used in an HCI setting.

I believe that signals pertaining to the user's high level cognitive functions are most useful for a passive adaptation: the knowledge of the user's frustration levels would prove more useful as an additional signal than the knowledge of basic visual signals. In this research, I focus on mental workload to improve the interface. I investigate ways to obtain workload information the user naturally gives off when using the computer by acquiring brain patterns, to automatically enhance their experience.

1.1 Thesis statement

This dissertation presents a body of technologies and tools that enable the use of real time functional near infrared spectroscopy measures of cognitive load for adaptive interfaces. This work is designed to support the following thesis:

> *Functional near infrared spectroscopy is an input technology usable in conventional human computer interaction contexts, especially when applied to the general, healthy public as an additional input.*

I identify three research questions that either shape the body of work presented in this thesis or motivate it. (1) What kind of cognitive states can we measure using fNIRS that can be useful in HCI contexts? (2) Can this technology be adapted to identify them in real time? (3) How should we use this information as input to an adaptive user interface? I also have a subgoal parallel to these questions, to find an accurate method for classifying multivariate sequential data we obtain from fNIRS.

To address my first question, I start by discussing practicality and applicability of the technology in realistic, desktop environments (Chapter 3). Ideally, for HCI research, the fNIRS signals would be robust enough to remain unaffected by physical activities, such as typing, occurring during the participant's task performance. I will then describe studies investigating the use of fNIRS to obtain meaningful data related to mental workload (Chapter 4 and 5). My studies progress from very controlled experiments that help us identify centers of brain activity, to experiments using simple user interfaces, showing how this technique may be applied to more realistic interfaces. In contrast to most previous fNIRS studies which only distinguish brain activity from rest, I also focus on distinguishing multiple states. Throughout all studies in this thesis, I show the use of novel machine learning techniques applied to fNIRS, to classify and use the brain activity information. My hypothesis is that useful features extracted from fNIRS data combined with machine learning models can accurately determine workload levels that the user was experiencing when completing a task in HCI.

To answer my second question, I transformed the offline processing analyses of fNIRS data and present a real time analysis and classification system (Chapter 6). Machine learning algorithms were changed to work with incoming data streams, and the predicted classification is used in real time interfaces to modify properties according to the user's cognitive load.

My third question focuses on creating new interactive, real-time user interfaces, which can adapt behavior based on brain measurements. This question serves as motivation throughout the thesis and I attempt to answer it in Chapter 6. The design challenge is to use this information in a subtle and judicious way, as an additional, lightweight input that could make a mouse or keyboard-driven interface more intuitive or efficient. Specifically, I am exploring situations and interfaces that can be adapted slowly, in a manner that is subtle and unobtrusive to the user, which could increase productivity and decrease frustration. I discuss prototypes of user interfaces that can adapt to the user's workload profile or other brain activity in real time.

The motivation for using fNIRS and other brain sensors in HCI research is to pick up cognitive state information that is difficult to detect otherwise (Lee & Tan, 2006). It should be noted that some changes in cognitive state may also have physical manifestations (overt user state). For example, when someone is under stress, his or her breathing patterns may change. It may also be possible to make inferences based on the contents of the computer screen, or on the input to the computer. However, since these can be detected with other methods, I am less interested in picking them up using brain sensors. Instead, I are interested in using brain sensors to detect information that does not have obvious physical manifestations, and that can only be sensed using tools such as fNIRS (covert state).

1.2 Thesis overview

The dissertation begins with an exploration of previous work and background knowledge that form the foundation of the current research. I describe different types of brain computer interfaces and measurements tools to sense brain activity and issues to consider when designing them. I follow with analysis methods for fNIRS data, and discuss previous real time work. I then focus on mental workload and techniques used to measure it. Finally, I address current work in BCI with an HCI point of view.

Because functional near-infrared spectroscopy eases many of the restrictions of other brain sensors, it has potential to open up new possibilities for HCI research. In Chapter 3, I identify several considerations and provide guidelines for using fNIRS in realistic HCI laboratory settings. Chapter 3 attempts to answer the second part of question one by exploring brain sensing in HCI contexts. I empirically examine whether typical human behavior (e.g. head and facial movement) or computer interaction (e.g. keyboard and mouse usage) interfere with brain measurement using fNIRS. Based on the results of my study, I establish which physical behaviors inherent in computer usage interfere with accurate fNIRS sensing of cognitive state information, which can be corrected in data analysis, and which are acceptable. With these findings, I hope to facilitate further adoption of fNIRS brain sensing technology in HCI research.

Chapter 4 and 5 explore brain signals methods with fNIRS and introduce two studies that distinguish different levels of mental workload. They both work towards solving my first research question. First, I distinguish levels of user workload and interaction styles. I look at four cognitive loads through a color counting task, both on graphical and physical objects. I use machine learning to analyze the data.

The following chapter distinguishes between levels of game difficulty. It describes a study designed to lead to adaptive interfaces that respond to the user's brain activity in real time. Subjects played two levels of the game Pacman while their brain activity was measured using fNIRS. Statistical analysis and machine learning classification results show that the system can discriminate well between subjects playing or resting, and distinguish between the two levels of difficulty with some success.

The last chapter creates a passive adaptive lightweight interface. I have developed a software system that allows for real time brain signal analysis and machine learning classification of affective and workload states measured with functional near-infrared spectroscopy, called the Online fNIRS Analysis and Classification system (OFAC). My system reproduces successful offline procedures, adapting them for real time input to a user interface. My first evaluation compares a previous offline analysis with my real time analysis. The second study demonstrates the online features of OFAC through the real time classification of two tasks, and the adaptation of an interface according to the predicted task. With OFAC, I have created the first working real time passive BCI using fNIRS, opening the door to building adaptive user interfaces. This chapter answers the second research question, and presents a high level discussion of the third question.

Chapter 2:

Background and Related Work

There are many components that interact in brain computer interfaces research. This multidisciplinary work ties fields such as neuroscience, brain anatomy, biomedical engineering and computer science. This chapter presents background knowledge and related work about brain computer interfaces, brain measurements, analysis methods, mental workload and human computer interaction.

2.1 Brain Computer Interfaces

A brain computer interface can be loosely defined as an interface controlled directly or indirectly by brain activity of the user. The most common types of brain computer interfaces use intentionally generated brain activity as the primary input device. They are called *active BCIs*, but they can also be labeled as *direct BCIs* or *BCIs for control*. Active BCIs are how most researchers define the general term of BCI, for instance by

Wolpaw et al. (2002). The original motivation (and conventional view) for such BCIs is to provide assistive technology for users with severe physical disabilities, such as paralyzed or “locked in” patients, to interact with their environment by translating their brain activity into specific device control signals (Adams, Bahr, & Moreno, 2008; Moore, 2003; Wolpaw, et al., 2002). This technology provides a new channel of communication that allows users to answer simple questions, control their environment, conduct word processing, or control prosthetics devices (Schalk, et al., 2004).

In addition, active BCIs often require the user to be trained to generate specific brain states which are interpreted as explicit input. These input behavior are not always related to the specific output action, for instance performing motor imagery of the left hand to move the cursor up, and motor imagery of the right foot to move it down (Mappus, et al., 2009). Daly et al. (2008) state that direct brain computer interfaces are unintuitive because of that inconsistency between input and output.

Active BCIs in contrast with *passive BCIs*, which detect brain activity that occurs naturally during task performance for use as an additional input, in conjunction with standard devices such as keyboards and mice (Cutrell & Tan, 2008). Passive BCI can detect voluntary input as active BCI, but their use is maximized when detecting signals such as emotions, language, and workload as passive BCIs are designed not to require the user’s full attention.

The terms active and passive can be used in other manner within BCI contexts. We define them as brain activity input to interfaces: active BCIs when brain signals are the only input activating the interface; passive BCIs when the interface reacts to other modalities as well as brain activity. This association follows the work by Cutrell and Tan (2008). However, Zander (2010) proposed a classification of brain computer interfaces according to the type of mental activity measured: active, passive or reactive. BCIs can measure active brain signals—generated intentionally—, passive—spontaneously generated states—, or reactive—states generated automatically upon the perception of certain stimuli.

These two paradigms also apply to physiological computing (Allanson & Fairclough, 2004; Fairclough, 2009). Physiological computing measures psychophysiological signals such as heartbeat, respiratory patterns, galvanic skin response, electroencephalography, electromyography, and uses them for purposes of biofeedback or interface adaptation. As Fairclough (2009) notes, "the physiological computing approach provides one means of monitoring, quantifying and representing the context of the user to the system in order to enable proactive and implicit adaptation in real-time.” Although Allanson and Fairclough (2004) use the expression “brain-computer interface” with a strict definition leading only to interfaces directly controlled by the brain, we extend their physiological computing principles to BCIs.

Many traditional brain-computer interfaces, designed for disabled users, require the user to be trained to control his or her brain activity, and this brain signal is used explicitly as the primary input to the system (Millán, et al., 2004). More recently, it has been suggested that untrained users may benefit from systems that use pattern recognition and machine learning to classify signals users naturally give off when using a computer system (Hirshfield, et al., 2009b; Lee & Tan, 2006). The system would use brain sensors to automatically discover aspects of the user’s cognitive state and use this information as passive or implicit input to a system, augmenting any explicit input from other devices, and increasing the bandwidth from humans to computers. Ju and Leifer (2008) present an interesting framework for implicit interactions.

There are many issues to consider when designing brain computer interfaces. We identify five main categories of BCI characteristics: recording technologies, physiological indicators, mental states and experimental strategies, feedback and adaptation, and users (Figure 2-1). Each issue will be addressed in this chapter or in the thesis in more detail, but they are presented here as a high level preview of what to keep in mind when exploring the field of brain computer interfaces.

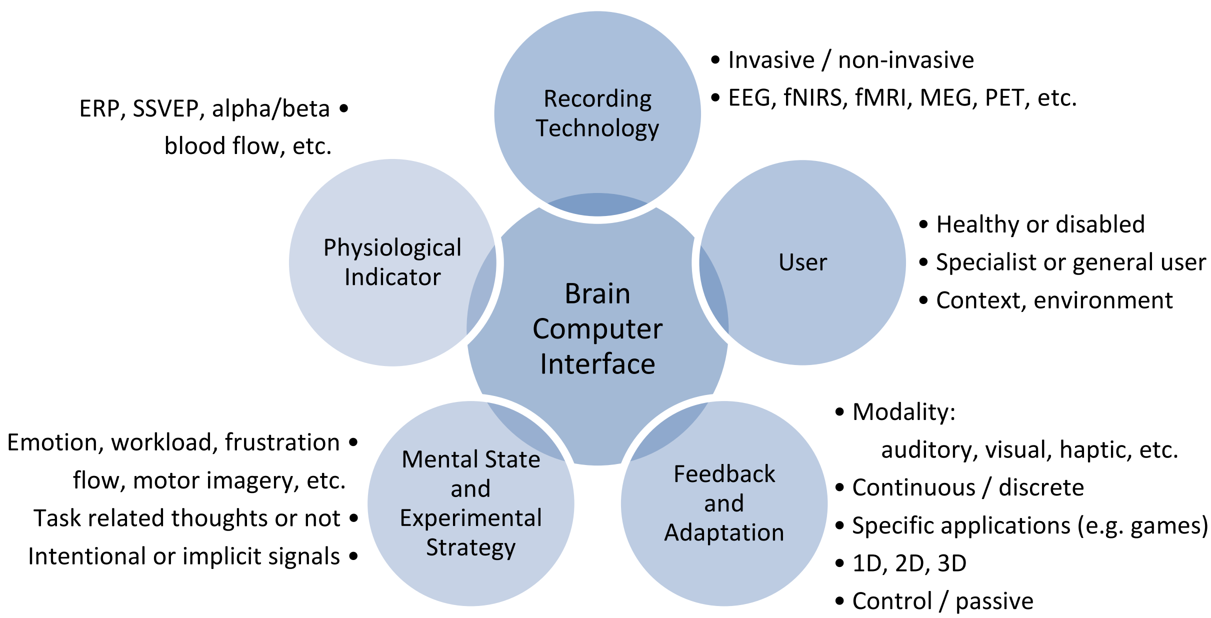

Figure 2-1. Issues to consider when designing a brain computer interface

Other authors have proposed BCI classifications, or elaborated the different components of a BCI, such as Pfurtscheller et al. (2007) who identified five components: brain signal, type of recording, experimental strategy, mode of operation and feedback.

The brain computer interface paradigm is analogous to that of eye movement-based interaction. Jacob (1993) found two axes to categorize such interaction, one in the nature of the eye movements and the other is in the response, with both axes going from natural interaction to unnatural, where there is no real-world counterpart. He recognized that most of the work up to that point was done with the disabled population, and linked unnatural eye movement to unnatural responses, but identified the benefit of research that combined natural eye movement with unnatural responses for healthy users. This is the direction I take with passive BCIs for healthy users in this research.

2.2 Brain Computer Interface Measurements

A myriad of brain imaging techniques have been utilized in BCI systems. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), magneto-encephalography (MEG), electroencephalography (EEG), electrocorticography (ECoG), and functional near-infrared spectroscopy (fNIRS) have been widely used to learn about human brain anatomy and activity. Coffey et al. (2010) provide a detailed table comparing the characteristics of fMRI, MEG, EEG and fNIRS in terms of principle of operation (e.g. hemodynamics or electromagnetic), signal characteristics (e.g. temporal and spatial resolution), and portability and comfort.

Although these techniques provide valuable insight into brain activity, some are very costly (mainly MEG, fMRI and PET) and invasive in ways that make them unsuited to ordinary HCI settings (ECoG) (Lee & Tan, 2006). For instance, they require motionless subjects in constricted positions, and they often expose subjects to hazardous radioactive materials (PET) or to loud noises (fMRI) (Izzetoglu, et al., 2004a; Raz, et al., 2005). In particular, the high magnetic field required for fMRI makes it difficult to introduce computer displays or input devices into the room for interaction as it prevents any metal in the vicinity. It permits largely passive situations, in which a subject can view projected images, but must remain extremely still, and interaction is difficult. Electrocorticography implants electrodes directly on the cortex, an invasive technique that is not practical for the general population.

Scerbo et al. (2001) provide a detailed list of the characteristics of a number of physiological and brain measures according to the criterions of sensitivity, diagnosticity, ease of use, current real world or real time feasibility, cost and intrusiveness, and their results are in line with my observations. In a comparison table, Coffey et al. (2010) arrive at similar conclusions, identifying EEG and fNIRS as better BCI candidates for space applications. Therefore, other techniques are not suitable for evaluating the brain activity of subjects under normal working conditions.

2.2.1 Electroencephalography

Because it is less invasive than other brain monitoring techniques, EEG has thus far been the tool of choice for researchers looking to measure user's brain activity non-invasively, for both brain computer interfaces and other areas such as to learn about neural correlates (Lee & Tan, 2006). EEG measures the electrical activity occurring in the brain when neurons fire. There are many approaches to using the EEG signal, especially in real time BCI, which include event related potential (ERPs), state visually evoked potentials (SSVEP), P300-based BCI, motor imagery, event related synchronization and slow cortical potentials (Krepki, et al., 2007). The two currently most common paradigms are SSVEP and P300. State visually evoked potentials are brain electrical signals occurring at the same frequency of the visual stimuli (between 3.5Hz and 75Hz) (Sutter, 1992). P300 based BCIs use the positive peak in the signal that occurs around 300ms after infrequent visual or auditory stimuli. Wolpaw et al. (2002) present a strong review of EEG based BCI systems.

The fine temporal resolution, ease of use, portability, and low set-up cost have made it the most commonly used BCI technique. EEG provides a fast measurement and seems appropriate for monitoring both instantaneous (<1s), and short term activity or states (<1min). BCI researchers have used machine learning algorithms on EEG data to increase system accuracy, system transfer rate, and to transition the burden of translating brain activity from the user to the computer. In an attempt to deploy EEG in environments beyond the laboratory, Gevins and Smith (2003) designed EEG hardware and data processing algorithms that comes closer to being usable in working environments such as airplane cockpits.

Unfortunately, there are a few drawbacks of using EEG. The technology is sensitive to subject movement and to interference from electronic devices near the EEG. It is susceptible to noise because fluid, bone, and scalp shield the electrodes from actual brain activity. For instance, eye blinking produces a large artifact in the signal. Its setup time is non-negligible, around 45-60 minutes in most systems, and it also requires placing the electrodes and gel across a user's scalp (Figure 2-2).

Figure 2-2. The setup for EEG requires placing each electrode individually, after applying gel to each location.

2.2.2 Functional Near-Infrared Spectroscopy

fNIRS can complement, and in some cases overcome technical/practical limitations of EEG and other brain monitoring techniques. While EEG measures electrical activity, fNIRS measures blood flow through hemoglobin concentrations and tissue oxygenation in the brain (Chance, et al., 1998; Chance, et al., 1993; Maki, et al., 1995; Meek, et al., 1995; Villringer & Chance, 1997; Villringer, et al., 1993).

fNIRS uses light sources placed on the scalp to send near-infrared light (in the wavelength range 650 - 850 nm) into the head. Biological tissues are relatively transparent at these wavelengths, so the light attenuation through tissues is sufficiently low to allow for tissue imaging at depths up to 2-4 centimeters (Bunce, et al., 2006). Deoxygenated and oxygenated hemoglobin, present in the blood, are the main absorbers of near-infrared light in tissues, and they provide relevant markers of hemodynamic and metabolic changes associated with neural activity in the brain. Therefore, fNIRS researchers can estimate hemodynamic changes connected to brain activation by using light detectors sensitive to the diffusively reflected light that has probed the brain cortex (Izzetoglu, et al., 2004b; Sassaroli, et al., 2006).

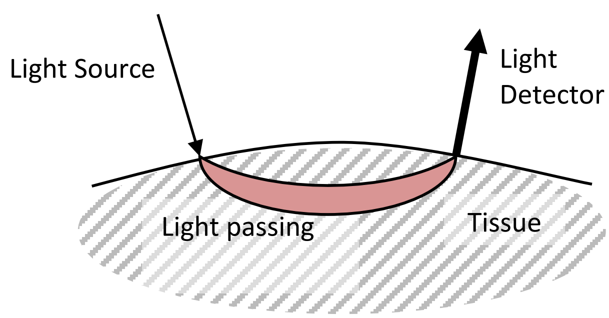

Figure 2-3. Light path in tissue, between source and detector.

By measuring the light sent at two wavelengths, we can calculate oxygenated and deoxygenated hemoglobin concentrations. Figure 2-3 illustrates the path taken by the light. Source fibers deliver light into the tissue. As a result of light scattering, some of the light travels through the tissue back to the surface and is collected by the detector. The light intensity measured at different source-detector distances is sampled at a frequency of a few Hz and translated into concentration changes of oxy and deoxy-hemoglobin, a measure of blood oxygenation. This results in several time series (one for each source-detector channel), providing a multivariate time series dataset that corresponds to blood oxygenation. Bunce et al. (2006) and Rolfe (2000) provide good overviews of the fNIRS technology applied to the brain and the photospectrometry principles inherent to this technology.

While fNIRS provides high temporal resolution, with data points measured in the order of tenths of milliseconds, the slow hemodynamic changes measured by fNIRS occur in a time span of 6-8 sec (Bunce, et al., 2006). Hence, fNIRS is appropriate to measure short term states but not instant ones. The spatial resolution of fNIRS is approximately five millimeters; the area measured being the one right below the sensor. However, it can only measure the cortical surface of the brain (Schroeter, et al., 2006; Tanida, Katsuyama, & Sakatani, 2007). In comparison, fMRI has a low temporal resolution but allows whole-brain imaging, including both cortical and subcortical structures, and EEG can gather information from electrodes placed all over the scalp, with a high temporal resolution. fNIRS causes less environmental stress than fMRI since it has no head constraints (Tanida, et al., 2007). It is also easy to setup, taking less than a minute to install the sensor (Figure 2-4) under a sports headband, which is an advantage over EEG.

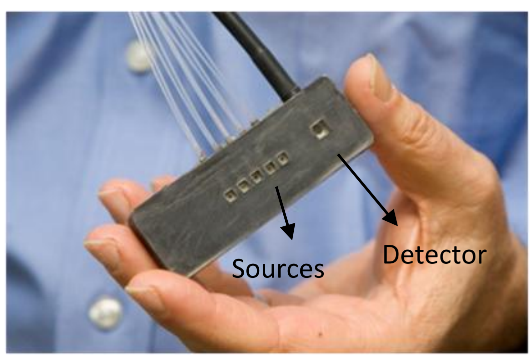

Figure 2-4. A linear array probe.

fNIRS sensors can take many forms, the only constraint being the presence of light sources and detectors. With multiplexed light sources, the same detector can be used with many sources, and many detectors can pick up the light from a single source in specific arrangements. Geometrical arrangements of light detectors and sources in an fNIRS device vary from linear arrays (Figure 2-5A) to circle configurations (Figure 2-5B), with varying degree of complexity depending on the number of light sources and detectors. Linear arrays allow the measure of different depth of the same brain area: sources placed further from the detector measure deeper tissues. Circle configurations are designed to probe a larger cortex area, at a fixed cortex depth.

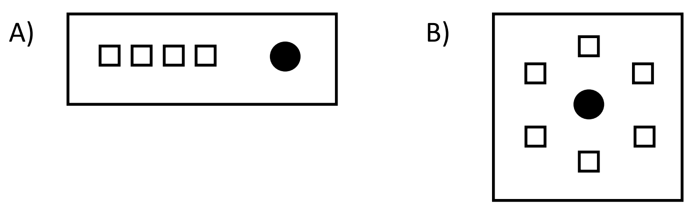

Figure 2-5. Possible geographical arrangements of fNIRS sensors.

White squares indicate sources, and black circles illustrate detectors.

While there are many brain imaging techniques, each with advantages and disadvantages, we believe fNIRS to be a suitable brain sensing technology for HCI research because it is safe, non-invasive, easy to use, and relatively impervious to user movement, as compared to other brain techniques (see Lee and Tan (2006), Scerbo et al. (2001) and Coffey et al. (2010) for brain sensing and imaging techniques comparisons). Chapter 3 explores potential sources of noise and artifacts and investigates in detail the influence of typing, clicking, head and facial movements on brain signals. fNIRS also removes many of the physical restrictions on the subject, in theory making naturalistic human-computer interaction possible. The technology is even portable (Hoshi, 2009).

2.2.3 Combining Technologies

There are few technological limitations to combining multiple brain sensing systems, such as EEG and fNIRS (Hirshfield, et al., 2009a; Kennan, et al., 2002; Merzagora, et al., 2009). The two types of sensors can be placed around the scalp in alternating patterns, with minimal interference, provided the EEG electrode gel is kept away from the fNIRS sensors. In fact, the combination of both signals can provide complimentary information about the user, and machine learning analysis will be particularly helpful in combining the rather disparate data. The use of two measurements can counter balance their disadvantages (Zander, et al., 2010), for instance in terms of time and spatial resolutions.

Brain sensing can also be combined with other modalities. Noel et al. (2005) showed that the combination of EEG with physiological signals such as cardiac, ocular, and respiration measures accurately classifies workload levels on multiple subjects and days, while classification was previously unsuccessful by looking at datasets separately. In turn, Vilimek and Zander (2009) combined EEG and eye gaze to build a mouse click system that used the eye movements to determine the object to be selected, and the brain signal to select it. Their system produced fewer errors than with eye gaze only, using each tool for their main benefit.

2.3 fNIRS Data Processing and Analysis Methods

fNIRS is still a new methodology, and as such it lacks well-established preprocessing and analysis methods (Butti, et al., 2007; Hoshi, 2009; Kondo, Dan, & Shimada, 2006). For example, Matthews et al. (2008) identified more than a dozen different signal processing techniques used in fNIRS to remove noises and motion artifacts. Physiological noise present in fNIRS signals includes cardiac signals (0.5-2 Hz), respiration (0.2-0.4 Hz) and Mayer wave (spontaneous low frequency oscillations without clear origins; 0.1 Hz) (Coyle, Ward, & Markham, 2004), and we can find many signal processing techniques to remove each of the source of noise (Coyle, et al., 2004; Izzetoglu, et al., 2005a; Matthews, et al., 2008; Robertson, Douglas, & Meintjes, 2010).

To analyze the data, each researcher is currently left to his or her better judgment to find a method that works best. Some researchers choose to do a visual inspection of the data to determine patterns (Nishimura, et al., 2007), while most use some sort of statistical analysis of the data, with no real consensus on how to perform this analysis.

2.3.1 Statistical Analysis

fNIRS data are multivariate time series, as we have many measurements at once. We identify two main types of statistical analysis approaches. Analyses of variance can be performed by comparing the different curves together (to identify whether different curves are activated differently), or by comparing each time point to the first point (baseline) of the curve (to identify whether the time series contains areas of activation). The analyses are also identified as single trial analysis or folding average, respectively. The second comparison, however, offers little generalizability over multiple measurements and subjects. Data reduction done with fNIRS for statistical analysis also varies: reducing the number of time points, or the number of sources of data (Figure 2-6). Reducing the number of points sometimes result in a single averaged point per curve.

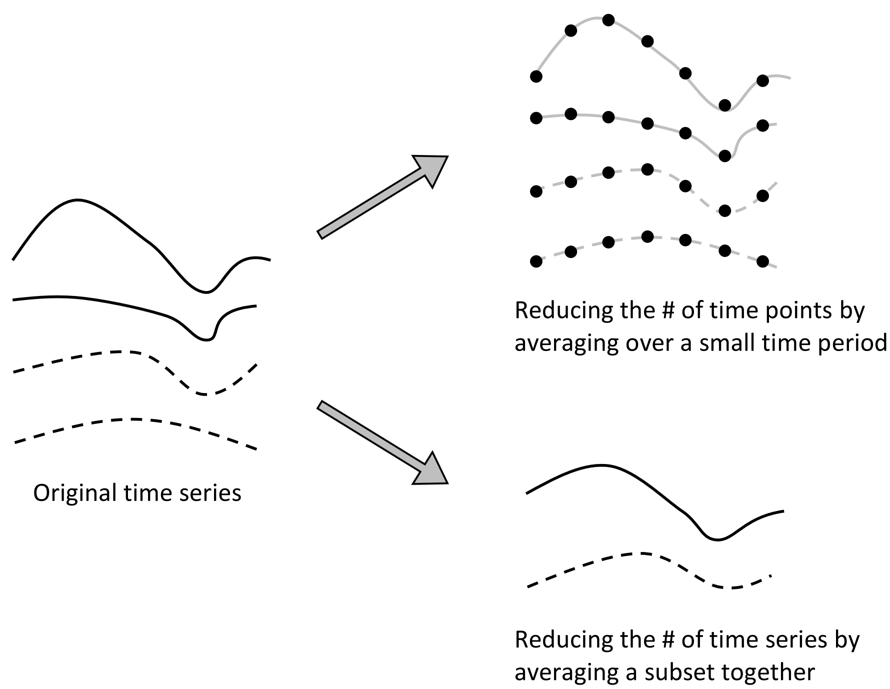

Figure 2-6. Illustration of the fNIRS data reduction in statistical analyses.

Practically, these approaches result in different analysis techniques. Many researchers perform paired t-test on averaged concentration change for each trial (Matsuda & Hiraki, 2005; Sakatani, et al., 2006), or analyses of variance on hemoglobin concentration (Izzetoglu, et al., 2004b; Matsui, et al., 2007), sometimes averaging channels together to obtain data about a certain region of the brain (Kono, et al., 2007). Others average all the trials at each time point and performs t-test to compare each point with a baseline point (Hirshfield, et al., 2009b; Nagamitsu, et al., 2006), or use a regressor model for analysis (Sitaram, et al., 2007). Additional analyses include using a generalized linear model (Butti, et al., 2007; Schroeter, Zysset, & von Cramon, 2004), an analysis of variance to compare multiple sessions (Kono, et al., 2007), or using the Pearson product-moment correlation coefficient between behavioral and fNIRS data (Izzetoglu, et al., 2007; Kono, et al., 2007).

2.3.2 Machine Learning Classification

Additionally, a small number of researchers perform machine learning classification and clustering on fNIRS data. Machine learning algorithms can provide more powerful data classification than traditional statistical techniques for real time analysis. Classification algorithms can perform well in the presence of noise and suboptimal data, which is common in measurements of brain activity. They can be applied offline as well as in real time, which is an important quality for adaptive brain computer interfaces.

Machine learning allows us to train classifiers without having domain expertise of the fNIRS output. For HCI researchers, the primary interest is in accurately detecting user state, rather than understanding why or how the fNIRS device produces its output. This added level of abstraction gives us flexibility to classify a myriad of user states. As fNIRS is an emerging technology, it is not clear how reliable the data will be, and whether the noise in the data will be prohibitive. We hope to show that fNIRS data analysis using machine learning can achieve high levels of accuracy in classifying the task a user is performing, making it a valuable, non-invasive tool for human-computer interaction research.

We find time series research in many domains such as meteorology, astrophysics, geology, multimedia, and economics (Sakurai, Yoshikawa, & Faloutsos, 2005), and there is a long history of time series analysis, including Fourier transforms, Hidden Markov Models, recurrent neural networks, and dynamic time warping (Berndt & Clifford, 1996). However, the EEG and fNIRS data analysis goals differ from those of other domains, as researchers are mostly interested in classifying current datasets, instead of making a future prediction, as is customary in fields like market economy.

In addition, the analysis of biological data collected using non-invasive methods such as EEG and fNIRS does present unique challenges. The complexity of the fNIRS dataset makes it a difficult problem to solve. The data is multivariate (many measures at once), and time series. This high dimensional data also usually suffers from a limited amount of examples. Depending on the classification technique, the cross validation may need to be performed in blocks, though this is not a problem unique to fNIRS, as pointed out by Lee and Tan (2006). Finally, the difficulty of cross-subject data classification for this type of data is an identified problem (Noel, et al., 2005).

Because fNIRS is a relatively new technique, machine learning algorithms have not been widely applied to data collected with this tool. There are many approaches that can be taken to classify time series data, such as comparing the distance from a sequence to another using Euclidian distance or dynamic time warping. One can observe the data in either the time or frequency domain, with a Fourier transform. fNIRS classification techniques explored include support vector machines and hidden Markov models to separate between left and right motor imagery (Sitaram, et al., 2007). Abdelnour and Huppert (2009) used an adaptive version of the general linear model that works in real time. Ayaz et al. (2009) compares classification with k-nearest-neighbor versus naïve Bayes to classify data from a second day of measurement using data from the first day, both for rest tasks and activation task separately. k-nearest-neighbor performed better at classifying rest tasks than activation tasks, but naive Bayes performed equally well (over 95% accuracy) for both classifications.

Some researchers have also attempted clustering. If the goal is to evaluate interface using brain signal, the use of clustering will provide a relevant tool to compare the different signals to see similarities (Hirshfield, et al., 2009b). Son et al. (2005) used K-means to cluster fNIRS data with moderate success. Their dataset was small and they analyzed each sensor's data separately, lacking generalizability. Hirshfield et al. (2009b) did clustering by preselecting the most relevant channels, and performed hierarchical clustering with unweighted average Euclidian distance similarity matrix. Their results compared which tasks were more similar to provide insights as to the mental similarities.

Research with other biological and physiological measures has achieved successful classification with machine learning. Wilson and Fisher (1995) achieved 86% correct classification of 14 tasks using EEG signals using principle components analysis with stepwise discriminant analysis. Anderson and Sijercic (1996) used two and three layer feed forward neural networks to classify five cognitive tasks from data collected using a six-channel EEG. They achieved 38%-71% accuracy, depending on the subject. Millan (2003) also used neural networks with success. Lee and Tan (2006) converted the data into a time independent dataset and achieves classification of three mental tasks at 84% accuracy, using a Bayesian network. Using a similar method, Grimes et al. (2008)

explored different feature generation for classification. Their results indicate that a larger window size produces higher accuracies, more training data leads to better results, although not by much, and more than two EEG channels produces similar results. Other methods include using Bayesian classifiers (Keirn & Aunon, 1990), artificial neural networks (Wilson & Russell, 2003) and clustering algorithms (Anderson & Sijercic, 1996).

2.4 Real Time fNIRS Brain Computer Interfaces

Many BCI systems and tools operate in real time, processing EEG data streams, and controlling interfaces (Krepki, et al., 2007; Pfurtscheller, et al., 2007; Schalk, et al., 2004; Wolpaw, et al., 2002). Delorme (2010) reviews more than a half-dozen existing BCI tools, and Schlögl (2007a) lists many open source packages. There are, however, very few fNIRS BCI real time systems available.

A common experimental protocol is to generate two brain states (either two types of activation, or one activation state and a rest state), and attempt to differentiate the two. Translated to a real time system, this protocol usually leads to binary decisions, where the user is asked to perform an activating task to indicate intent, and to rest otherwise.

The most common outcomes of such binary decisions are direct control of interfaces, active BCIs. For instance, Coyle et al. (2007) presented a real time fNIRS system that allowed participants to select a colored box by performing a mental rotation when the preferred target was highlighted. We found this simple interface to be one on the first example demonstrating the ability for the fNIRS signal to be analysed in real time. Their state distinction is done using a threshold, which requires domain expertise and the selected settings may not be easily reused from one participant to another.

In a more complex (but preliminary) interface by Mappus et al. (2009), users drew a line on a two dimensional plane using activated periods to do straight lines, and rest periods to curve the line. In both studies, participants were instructed what brain task to perform in order to use the system. Another fNIRS system, proposed by Nishimura et al. (2010), uses the hemoglobin activation value to control the movements of a swimming dolphin (up or down). The continuous feedback is engaging to participants as they try to move the dolphin in order to eat fish placed at different heights.

Using a different paradigm than the activation-rest one, Luu and Chau (2009) compared two different activated brain signals to indicate drink preference. To my knowledge, this is the first example of a real time system that distinguished two activation states without specific instructions. Their analysis simply compared the signals and identified the signal with maximal amplitude as the drink of choice.

While the studies mentioned are using direct brain input, we believe fNIRS to be better suited for passive BCIs. The relatively slow signal response of fNIRS doesn’t lend itself to be the best technology for rapid communication, direct input, especially when designed for the general public. In addition, researchers use basic techniques such as selecting the signal with the highest amplitude to make the binary decisions. We believe there are more powerful and better suited techniques.

2.5 Measuring the Brain with fNIRS

Much fNIRS research until now focused on iteratively designing the tool and running feasibility studies to show that it measures brain activity with accuracy levels comparable to more well-established brain imaging techniques (Sassaroli, et al., 2006). As compared to other brain imaging devices which have been around for a long time, the fNIRS device is still in relative infancy (Lee & Tan, 2006). The extensive applications conducted with other brain imaging techniques such as EEG, have yet to be implemented. None the less, current research shows optimism to reach this stage in the near future.

Brain activity measurements with fNIRS are directly linked to the sensor's location. There are many possible placements of probes, allowing the study of multiple brain regions. The basic technology is common to all systems, and the measured signal depends on the location of the probe and the amount of light received.

The most common placements are on the frontal lobe, including the motor cortex, and the prefrontal cortex (PFC), although other regions have also been explored such as the visual cortex (Herrmann, et al., 2008a) (Figure 2-7). The frontal lobe plays a part in memory, problem solving, judgment, impulse control, language, motor function, sexual behavior, socialization and spontaneity. It also assists in planning, coordinating, controlling and executing behavior.

Sensing the motor cortex allows the detection of both motor tasks, such as moving a limb (Sitaram, et al., 2007), or motor imagery, where the movement is thought but not executed (Coyle, et al., 2003; Sitaram, et al., 2007). Motor imagery produces a smaller signal than motor tasks, but has a greater potential with the disabled, paralyzed population of users.

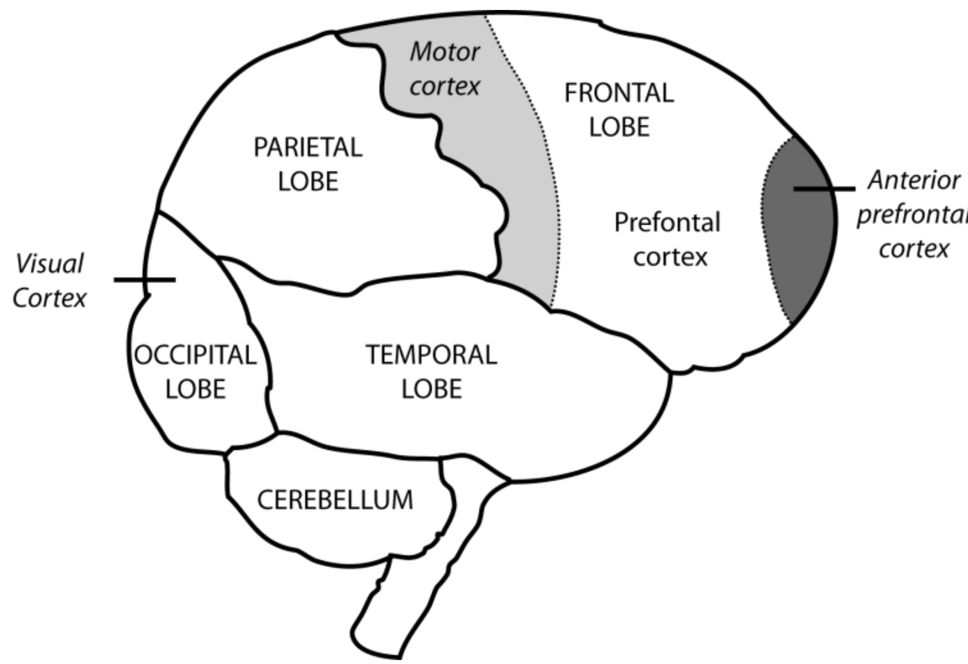

Figure 2-7. Cerebral lobes and the anterior prefrontal cortex.

While Matthews et al. (2008), note that the “motor cortex activation is the most common mental strategy for fNIRS-BCI control” researchers have shown that by placing the light sources and detectors on a subject’s forehead, fNIRS provides an accurate measure of activity within the prefrontal lobe of the brain (Quaresima, et al., 2005). We believe these prefrontal cortex signals to be of great potential to HCI, more so than measurements of the motor and visual cortex. If the participant is using a computer, the system is aware of the core of their movements (though keyboard and mouse input), as well as what is in their visual field, reducing the usefulness of the motor and visual cortex measurements.

The prefrontal cortex has been the source of a large number of studies. Emotions were investigated through alertness (Herrmann, et al., 2008b), and general arousal and valence levels generated by showing pictures (Leon-Carrion, et al., 2006; Yang, et al., 2007). Stress has also been shown to increase oxyhemoglobin (Tanida, et al., 2007), as well as anxiety, where anticipation of a shock produces high activation (Morinaga, et al., 2007). Both Bunce et al. (2005) and Tian et al. (2009) successfully investigated the detection of intentional deception in adults. Mappus et al. (2009) studied language production in Broca's area. Kobuta et al. (2006) researched the prediction of false memory, which occurs when subject recognize a previously unstudied word semantically related to a group of words memorized.

Finally, a large number of articles study the more general concept of mental workload through tasks like a warship control task (in a command and control environment), navigation into hyperspace, auditory ordering of letters, preference (Hirshfield, et al., 2009b; Izzetoglu, et al., 2004b; Izzetoglu, et al., 2005b; Luu & Chau, 2009; Son, et al., 2005), although some research have studied specific components of it. Hirshfield et al. (2009b) attempted to separate syntax and semantics of interfaces and succeeded in identifying the syntactic elements.

Within the prefrontal cortex, we chose to study specifically the anterior prefrontal cortex (aPFC), also called the frontal poles, an active region that deals with high-level processing, such as working memory, planning, problem solving, inhibition, memory retrieval and attention (Burgess, Quayle, & Frith, 2001; Horn, et al., 2003; Koechlin, et al., 2000; Ramnani & Owen, 2004; Simons, et al., 2005). The aPFC region is located under the forehead (Figure 2-7), and is identified by the Brodmann area 10p. It was selected because of location specific neural correlates, and because of easy access. As the forehead is hairless, we can use simple, comfortable sensors, and we can access it on everyone. This is a major benefit of our setup.

In most fNIRS studies, researchers identify the difference between two states only: activation and rest. Activation occurs when subjects perform a specific task for a few seconds up to a few minutes, such as mental rotations, arithmetic, or language production. Rest periods are produced by telling the user to think of nothing and stare into an empty screen. These studies are mostly designed to identify which types of activity are present in specific locations. They omit the exploration of finer details of levels of activation.

2.6 Mental Workload

The aPFC is rich is high level processes, and I concentrate my research on mental workload. Mental workload is a concept used by many, and yet researchers cannot all agree on a single definition of the term (Hacker, 2006). Nevertheless, they do agree that mental workload is multidimensional, influenced by a wide variety of elements, such as visual perception, selection, memory (storing and recall), comprehension and processing, data entry, reasoning, and motor movements (Iqbal, et al., 2005). Mental workload is composed of both conscious and unconscious efforts to perform a task (Alty, 2003). It is well understood that a reliable measure of user workload could have a positive impact in many real life interactions (Guhe, et al., 2005; Iqbal, Zheng, & Bailey, 2004; John, et al., 2004).

Performance as a function of mental workload can be illustrated with a Gaussian curve (Figure 2-8), where low and high mental workload are associated with a reduced performance. There is an optimal mental workload level associated with the optimal performance. Low mental workload, also called underload, is often observed by operators monitoring automated systems for long period with little intervention (Hancock & Chignell, 1988). Overload (high mental workload) may happen when novices must perform a highly difficult task, in a small amount of time, for example. For a common task, associated mental workload depends on the experience of the operator.

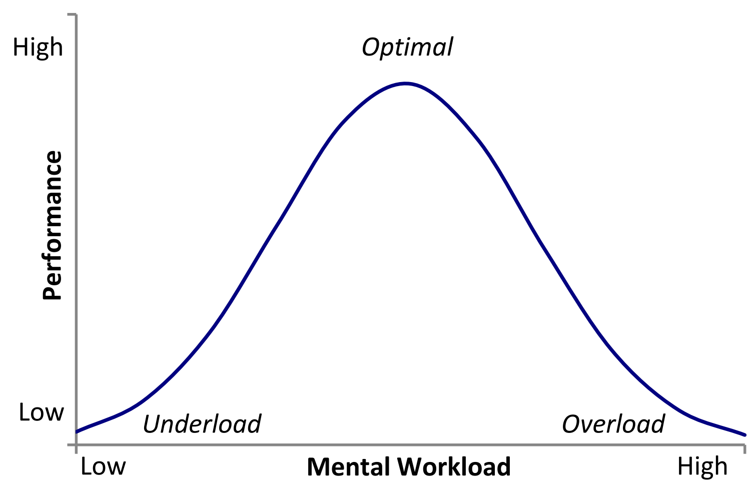

Figure 2-8. Performance according to Mental Workload.

Some researchers have associated mental workload with effort. Hancock and Chignell (1988) evaluate mental workload with the formula: $W = 1/et^{s-1}$, where e is the effort required for the task, t the time available to perform it, and s the skill level (low for novices, high for experts). In this case, mental workload is inversely proportional to the effort. However, evaluating effort is just as difficult as evaluating workload, leaving this formula unused.

Mental workload also usually varies over the course of a task. Complex and/or long tasks are composed of subtasks, each of which results in different levels of mental workload (Iqbal, et al., 2005). Further, there is also a small lag between the task demand and the mental workload level (Hancock & Chignell, 1988). Figure 2-9 illustrates this chronological fluctuation.

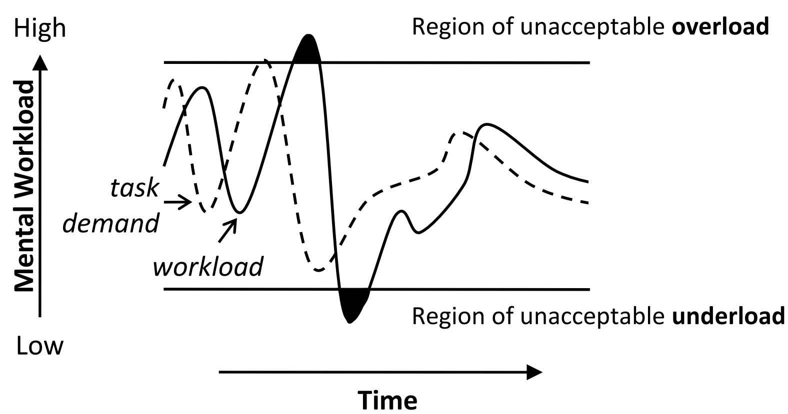

Figure 2-9. Change in level of mental workload as function of chronological progression. The level of workload displays a small lag following the task demand. Black areas represent regions of unacceptable load (Hancock & Chignell, 1988).

2.6.1 Assessing Mental Workload

Performance, physiological and psychophysiological measurement, subjective assessment and secondary tasks performance can be used to measure mental workload, all presenting advantages and drawbacks. Reliable measures of user workload can have a positive impact on performance (Guhe, et al., 2005; Iqbal, et al., 2004).

Numerous physiological measures have been proven to reflect the current mental workload of the user, such as change in heart, respiration, blink rate and pupillary response (Chenier & Sawan, 2007; Iqbal, et al., 2004), change in body temperature, galvanic skin response (John, et al., 2004; Tao, et al., 2005), facial features (Guhe, et al., 2005), to name a few. While those measures are objective and obtained in real time, their main drawback is that they are external manifestations of the cognitive state.

Measuring user workload with psychophysical measures such as electroencephalogram (Gevins & Smith, 2003; Grimes, et al., 2008; Kok, 1997; Lee & Tan, 2006), functional near-infrared spectroscopy (Izzetoglu, et al., 2003; John, et al., 2004; Son, et al., 2005), and facial electromyography (Fuller, et al., 1995) have also been a topic of much research recently. These measures provide an objective assessment of mental and physical responses to a particular task. They also allow real time measurements, used to create real time adaptive systems. However, these physiological and psychophysiological measures rely on equipment that could be difficult to use properly (hard to place correctly on the body, for example), and they impose physical constraints on the user (Wickens & Hollands, 1999). Within this group of measures, we believe fNIRS to have the advantages associated with a direct, objective, and potentially real time load measures, while having fewer constraints, and being easier to setup than most.

Subjective assessment tools provide simple methods to evaluate the load imposed on the user for a particular system. One of the first measures of mental workload because they do not require any special equipment, these measures do not influence the task itself since they are performed after the fact. However, they are self-observation, subjective by nature, and the data cannot be collected in real time.

There are many subjective tools available on the market. Three common multidimensional assessments technique are the NASA Task Load Index (NASA-TLX), the subjective workload assessment technique (SWAT) and the Workload Profile (WP). The NASA-TLX (Hart & Staveland, 1988) measures workload on six different dimensions (mental demands, physical demands, temporal demands, own performance, effort, and frustration), and adds weights to balance each value per task, to calculate the amount and type of mental workload a user experiences during task performance. SWAT (Rubio, et al., 2004) uses the conjoint measurement technique to combine ratings on three different dimensions of workload (time load, mental effort load, and stress load). WP (Tsang & Velazquez, 1996) compares the proportion of attention resources of users of four workload dimensions (stage of processing, code of processing, input, and output) measured after the completion of all experiments. Rubio et al. (2004) compared these three methods and recommended using the Workload Profile when the goal is to compare two or more tasks with different levels of difficulty. They advise NASA-TLX for predicting the performance of an individual at a task. Finally, when an analysis of cognitive demand is required, WP is the better choice, followed by SWAT.

In an attempt to combine the real time nature of physical measurement with the self-examinatory quality of subjective assessment, Pickup et al. (2005) developed the Integrated Workload Scale (IWS), a one-dimension eleven point scale, that prompts users to categorize their current mental workload every couple of minutes. The authors showed a correlation between the IWS measure and the task demand, showing that mental workload could be measured by IWS. This measure shows potential for combining multiple types of assessment to produce an accurate method, but it interrupts the user constantly.

Finally, secondary tasks performance can provide a reliable measure of mental workload (Hockey, et al., 2003). Consider the situation where a user is given instructions to perform a first task correctly, in priority, and to perform a second task when possible. The performance of the second task will be an indicator of the effort put into maintaining his performance at the first task. Just as physiological assessment, secondary task assessment provides a real time measure of the mental workload of the user. However, performing two tasks at the same time is harder than one, leading to possible secondary task contamination, where a second task actually influences the performance of the first task. Koechlin et al. (2000) found anterior prefrontal cortex activation for dual tasks, especially during branching, where subjects remember a primary goal while processing secondary tasks.

2.6.2 Game Play

The mental workload framework encompasses many types of tasks, and I narrow my focus on game play. These multidimensional tasks include time constraints, sub goals, planning, visual perception and motor movements and they should lend themselves well to brain sensing. Indeed, game play has been measured using psychophysiological signals. For instance, Chen et al. (2008) used two physiological measures (heart rate variability and electromyogram) to measure the interruptibility of subjects in different tasks, including a game, and found a high correlation between those measures and the self-report of interruptibility. Chanel et al. (2008) successfully differentiated between three emotional states (boredom, engagement and anxiety) using galvanic skin response, blood pressure and respiration, and suggest game adaptation based on those states.

Several fNIRS studies evaluating gaming environments reported a significant variation in hemoglobin concentration in the prefrontal cortex in comparison to resting in many studies. Using fNIRS, Nagamitsu (2006) observed a significant increase in the hemoglobin concentration of the prefrontal cortex in adult subjects while playing an arcade game (Donkey Kong). Matsuda and Hiraki (2005, 2006) reported a decrease in oxygenated hemoglobin in the prefrontal cortex when playing video games, both in adult and children. Their subjects played a shooting game, a rhythm action game, a block puzzle and a dice puzzle. Hattahara et al. (2008) investigated the influence of expertise on fNIRS measured brain activity. They report that novices produce strong deactivations in the prefrontal cortex, but that the response is inversed with experts. However, this result was obtained with a very limited number of subjects (three subjects for each of three levels), and the generalizability of their work is unclear. Their subsequent experiment comparing one subject’s brain during four measurement sessions also reaches inconclusive results.

Studies with other brain measurements corroborate the activation of the prefrontal cortex when playing games. A functional magnetic resonance imagery (fMRI) study by Saito et al. (2007) demonstrated that they could differentiate between playing and not playing a computer game. Their study compared three video games: Space Invaders, Othello and Tetris. Others have measured the brain during game play using EEG and demonstrated the ability to distinguish the user resting, exploring the game environment or playing the video game (Lee & Tan, 2006). Nijolt, Bos and Reuderink (2009) present a comprehensive survey of EEG games research, showing the success in measurements, and potential in use.

2.7 Brain Sensing in Human-Computer Interaction

To conclude this related work chapter, it is imperative to discuss brain sensing research within human computer interaction. Gevins and Smith (2003) identified four qualities of cognitive load monitoring methods necessary for HCI settings: the tools should be "robust enough to be reliably measured under relatively unstructured task conditions, sensitive enough to consistently vary with some dimension of interest, unobtrusive enough to not interfere with operator performance and inexpensive enough to eventually be deployable outside of specialized laboratory environment."

Researchers have taken two main paths with brain sensing: either to evaluate interfaces, or to adapt them. The core of the work has been done in interface adaptation (or towards interface adaptation), although we believe that usability and user experience evaluation is a growing field.

2.7.1 Usability and User Experience Evaluation

Using fNIRS brain sensing, Hirshfield et al. (2009b) explored separating syntactic and semantic components of a user interface following Shneiderman's theory (2005). Hypothesizing that the overall mental effort required performing a task using an interactive computer system is composed of a portion attributable to the difficulty of the task itself plus a portion attributable to the difficulty of operating the user interface of the interactive tool, they successfully identified syntax components, which can be used to redesign interfaces.

In an ACM CHI 2010 Conference workshop entitled BELIV'10: BEyond time and errors: novel evaLuation methods for Information Visualization (Bertini, Lam, & Perer, 2010), participants discussed the use of physiological measures to evaluate information visualization tools (Riche, 2010). We believe this is a trend that will extend to brain measures.

2.7.2 Interface Adaptation

The term *adaptive interface* relates to the automatic modification of the interface without explicit user directives to optimize a certain property (e.g. performance). According to Kuikkaniemi et al. (2010), “adaptation refers basically to systems which collect data on user or use-context and adapt their functionality according to some algorithm”. Wilson and Russell (2007) define a subset of adaption, called adaptive aiding, which is designed specifically to help the user accomplish their current task. The goal of adapting aiding is to “dynamically match the momentary cognitive capabilities of the operator with the demands of the task”. Allanson and Fairclough (2004) define the biocybernetic loop as interfaces that adapt based on the real-time measurement of psychophysiology. Coyle et al. (2009) proposes a limited theory on how to adapt interfaces, mainly to reduce the intrinsic cognitive load, which is how difficulty the new material or task is to learn.

Adaptation and adaptive aiding can be done using many measures. Specifically, adaptive aiding is a method of providing assistance to the operator by introducing automation only when required (Parasuraman, Mouloua, & Molloy, 1996). Parasuraman et al (1996) identified five strategies to implement aiding in systems, based on critical environmental events, operator workload, performance, or physiology, and performance modeling. They evaluated model-based and performance-based adapting, and showed that both provide significant improvement, with no basis for a choice of method. The adaptive method should be selected using other considerations, such as user preference or availability. Finally, hybrid methods that combine a subset of the other five techniques might also improve the performance. As adaptation may not always be the best strategy to help the user, Wilson and Russell (2007) explored different aiding techniques in a task with high cognitive load, one of which adapted the interface through brain and physiological sensing and one without aiding. They found an overall improvement in performance by aiding, and that adaptive aiding is better than random aiding.

In a survey of the physiological computing, Allanson and Fairclough (2004) identified significant findings in biocybernetic adaptation. They found increased performance and engagement when the adaptation was sustained for long periods. They also observed that biocybernetic adaptation leads to “increased performance and reduced subjective mental workload”.

We find many examples of interface adaptation in the literature, mainly done with EEG as it is the most commonly used technology. We identify a few successful examples of the diverse results achievable, both with active or passive BCIs.

Games are an application of choice for BCI researchers. In a state-of-the-art survey of BCI for games, Nijholt, Bos and Reuderink (2009) point out two axes of ways to use brain signals in games: one axis for the type of action, either game control or game adaptation; the second axis for the type of signal, either as internally or externally evoked signals, which is equivalent to active and passive signals as defined by Zander (2010). For instance, the user could do a mental calculation to externally evoke a game command, or the game could adapt to the user's boredom (internally invoked).

An fNIRS active BCI was created by Nishimura et al (2010). They proposed a dolphin trainer game that allows participants to control their brain signal to move a dolphin up and down to eat fish. The application can generate fish of different colors, each of which could be associated with unique tasks. For example, each fish could trigger the move of a different board game piece using a robotic arm.

Recently, Yuksel et al. (2010) used the common P300 electroencephalography paradigm to select physical objects by placing them on an interactive multi-touch table. This extension of the P300 paradigm, typically used to spell words, fits well into an HCI context.

Finally, by programming the behavior of a domestic robot using a commercially sold device that measures bioelectric signals (OCZ Peripherals, 2010), Saulnier, Sharlin and Greenberg (2009) have shown a simple example of an application of brain activity in day to day tasks. While they investigated the direct control of the speed of the robot with emotional states, they found behavioral control to be more reliable and appreciated by the participants. In this case, the robot would clean when the person was stressed, while it behaved more like a pet, sitting near the user when s/he was relaxed. However, they found the commercial system very limited, and their experience showed only muscle tension was measured reliably.

2.8 Summary

The work presented in this thesis will address some of the lacunas of the discussed prior work with fNIRS. We identified and summarized the main issues with the related work.

Most of the fNIRS studies compare an activated state with a rest state, which omits the exploration of levels of activation. The analysis of those studies is almost always performed offline, lacking any applicability for real time brain computer interfaces. The few fNIRS systems working in real time compare two states which lead to a binary decision. The response always controls the interface, and never leads to passive adaptation. Additionally, these real time studies omit discussing a meaningful integration of the inherent fNIRS signal delay into the interface response (which cannot be instantaneous). The techniques used to perform analyses require fNIRS domain expertise, which limits their use by non-experts, who might knowledgeable in other fields. Finally, we found not work that explicitly explores the impact of typical computer artifacts in data, although this work would have a high impact for any real-world applicability.

Chapter 3:

Using fNIRS in Realistic HCI Settings[1]

To be valuable in human computer interaction (HCI) settings, brain sensors should collect useful information while ideally allowing normal interaction with the computer, such as looking at the screen, or using the keyboard and the mouse. In addition, the measurements should have a quick set up time, be comfortable, place few (or no) postural constraints, and provide continuous, real time measures.

Because most brain imaging and sensing devices were developed for clinical settings, they often have characteristics that make them less suitable for use in realistic HCI

[1] The work in this chapter was originally described in Solovey, et al. "Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines" in the proceedings of the ACM UIST'09 Symposium on User Interface Software and Technology, (2009) p.157-166. This was joint work with Erin Solovey.

settings. For example, although functional magnetic resonance imaging (fMRI) is effective for functional brain imaging, it is extremely susceptible to motion artifacts, and even slight movement (more than 3mm) will corrupt the image. In addition, the strong magnetic field prohibits all metal objects from the room, making computer usage impractical. Even the most common technology used for brain-computer interfaces, electroencephalography (EEG), poses some obstacles for HCI, as it is susceptible to artifacts from eye and facial movements, requires gel in the participant's hair, takes some time to set up properly, and is subject to noise from nearby electronic devices.

Figure 3-1. The use of fNIRS in typical computer settings.

We believe that functional near-infrared spectroscopy (fNIRS) overcomes some of those constraints, and is well-suited for use in HCI, in part because the fundamental technology and the sensors do not constrain the user (Figure 3-1). fNIRS has been used in previous HCI studies because it has many characteristics that make it suitable for use outside of clinical settings (Hirshfield, et al., 2009b; Mappus, et al., 2009). Benefits include ease of use, short setup time, and portability, making it a promising tool for HCI researchers.

While we intend to use fNIRS to pick up psychophysiological data, we do not expect that the participant is physically constrained while using the computer. Yet, common behaviors such as head and eye movements are currently restricted during most fNIRS experiments.

In most studies using any type of brain sensors, researchers control these problems by expending great effort to reduce the noise picked up by the sensors. Typically, participants are asked to remain still, avoid head and facial movement, and use restricted movement when interacting with the computer. In fMRI, subjects are even physically restrained by soft pads to prevent movements from disrupting the measurements (Raz, et al., 2005). The experiments are often held in soundproofed rooms to prevent environmental noise and electrical interference with the measures. In addition, many factors simply cannot be controlled, therefore researchers sometimes throw out data that may have been contaminated by environmental or behavioral noise, or they develop complex algorithms for removing the noise from the data. By doing this, the researchers hope to achieve higher quality brain sensor data, and therefore better estimates of cognitive state information.

However, it is not clear that all of these factors contribute to problems in the case of fNIRS or that these restrictions improve the signal quality. Ideally, for HCI research, the fNIRS signals would be robust enough to be relatively unaffected by other non-mental activity occurring during the participant's task performance. In fact, one of the main benefits of fNIRS is that the equipment imposes very few physical or behavioral restrictions on the participant (Hoshi, 2009). Thus, we would like to establish which physical behaviors inherent in computer usage interfere with accurate fNIRS sensing of cognitive state information, which can be corrected in data analysis, and which are acceptable.

We felt it was important to identify and examine empirically considerations necessary for appropriate use of fNIRS in realistic HCI laboratory settings. Based on the results of our study, we will provide guidelines clarifying which behavioral conditions need to be controlled, avoided, or corrected when using fNIRS, and which factors are not problematic. With this information, researchers can better take advantage of fNIRS brain sensing technology.

3.1 fNIRS Considerations

With the introduction of any new technology, there are considerations that should be made for its proper use. For this reason, we use our previous experience with fNIRS as well as a literature review to recognize characteristics specific to fNIRS sensors that are relevant for HCI, and develop paradigms for using fNIRS properly in HCI research. In particular, we identify below potential sources of noise and artifacts in the fNIRS signal when used in typical HCI laboratory settings.

As mentioned in Chapter 2, we have selected the brain region of the anterior prefrontal cortex as location of our measures. Hence, our considerations below are intended for researchers measuring the anterior prefrontal cortex, as the impact of the human behavior and typical interactions will vary depending on the measured region of the brain. However, we expect our results to be valid for other experimental setups and contexts that use the prefrontal cortex area.

3.1.1 Head Movement

Several fNIRS researchers have brought attention to motion artifacts in fNIRS sensor data, particularly those from head movement (Devaraj, et al., 2004; Matthews, et al., 2008). Matthews et al. (2008) explains that “motion can cause an increase in blood flow through the scalp, or, more rarely, an increase in blood pressure in the interrogated cerebral regions.” In addition, they point out that “orientation of the head can affect the signal due to gravity’s effect on the blood.” They note that these issues are significant if the head is not restricted, and even more so in an entirely mobile situation. However, other researchers indicate that fNIRS systems can “monitor brain activity of freely moving subjects outside of laboratories" and note that “measurements with less motion restriction in the daily-life environment open new dimensions in neuroimaging studies” (Hoshi, 2009). While fNIRS data may be affected by head movements, this should be contrasted with fMRI where movement over 3mm will blur the image. Because of the lack of consensus in the community, we chose to investigate the artifacts associated with head movements during typical computer usage to determine their effect on fNIRS sensor data in a typical HCI setting.

3.1.2 Facial Movement

fNIRS sensors are often placed on the forehead, and as a result, it is possible that facial movements could interfere with accurate measurements. Coyle, Ward, and Markham

(2004) point out that “slight movements of the optodes on the scalp can cause large changes in the optical signal, due to variations in optical path. It is therefore important to ensure robust coupling of optodes to the subject’s head”. These forehead movements could be caused by talking, smiling, frowning, or by emotional states such as surprise or anger, and many researchers have participants refrain from moving their face, including talking (Chenier & Sawan, 2007). However, as there is little empirical evidence of this phenomenon, we will examine it further in the experiment. We selected frowning for testing as it would have the largest effect on fNIRS data collected from the forehead.

Eye movements and blinking are known to produce large artifacts in EEG data which leads to the rejection of trials including such an artifact (Izzetoglu, et al., 2004b). However, fNIRS is less sensitive to muscle tension and researchers have reported that no artifact is produced in nearby areas of the brain (Izzetoglu, et al., 2004b). It would also be unrealistic to prevent eye blinks and movement in HCI settings. Overall, we conclude eye artifacts and blinks should not be problematic for fNIRS, and we do not constrain participants in this study.

3.1.3 Ambient Light

Because fNIRS is an optical technique, light in the environment could contribute to noise in the data. Coyle, Ward, and Markham (2004) advise that stray light should be prevented from reaching the detector. Chenier and Sawan (2007) note that they use a black hat to cover the sensors, permitting the detector to only receive light from the fNIRS light sources.

While this is a concern for researchers currently using raw fNIRS sensors that are still under development, we feel that future fNIRS sensors will be embedded in a helmet or hat that properly isolates them from this source of noise. Therefore, in this chapter, we do not further examine how the introduction of light can affect fNIRS data. Instead we just caution that excess light should be kept to a minimum when using fNIRS, or the sensors should be properly covered to filter out the excess light.

3.1.4 Ambient Noise

During experiments and regular computer usage, one is subjected to different sounds in the environment. Many studies using brain sensors are conducted in sound-proof rooms to prevent these sounds from affecting the sensor data (Morioka, Yamada, & Komori, 2008). However, this is not a realistic setting for most HCI research. Wakatsuki et al. (2009) demonstrated that environmental noise (construction sounds) did not have an influence on brain activation in the PFC unless they were at high volume. Therefore, we conducted this study in a setting similar to a normal office. It was mostly quiet, but the room was not soundproof, and there was occasional noise in the hallway, or from heating and air conditioning systems in the building.

3.1.5 Respiration and Heartbeat

The fNIRS signals picks up artifacts from respiration and heartbeat, by definition, as it measures blood flow and oxygenation (Coyle, et al., 2004; Matthews, et al., 2008). These systemic noise sources can be removed using known filtering techniques. For a discussion of the many filtering techniques, see Matthew et al. (2008) and Coyle et al. (2004).

3.1.6 Muscle Movement

In clinical settings, it is reasonable to have participants perform purely cognitive tasks while collecting brain sensor data. This allows researchers to learn about brain function, without any interference from other factors such as muscle movement. However, to move this technology into HCI settings, this constraint would have to be relaxed, or methods for correcting the artifacts must be developed. Fink et al. (2007) discussed the difficulty of introducing tasks that have a physical component in most brain imaging devices, explaining that they may "cause artifact (e.g. muscle artifacts in EEG or activation artifacts due to task-related motor activity in fMRI) and consequently reduce the number of reliable (artifact-free) time segments that can be analyzed". In addition, they note that the test environment of fMRI scanners also makes it difficult for any physical movement. Their solution was to have subjects think about their solutions during brain measurements, and to provide it after the measurement, which does not seem to be a likely solution for real world settings.

One of the main benefits of fNIRS is that the setup does not physically constrain participants, allowing them to use external devices such as a keyboard or mouse. In addition, motion artifacts are expected to have less of an effect on the resulting brain sensor data. In this study, we examine physical motions that are common in HCI settings, typing and mouse clicking, to determine whether they are problematic when using fNIRS.

3.1.7 Slow Hemodynamic Response

The slow hemodynamic changes measured by fNIRS occur in a time span of 6-8 seconds (Bunce, et al., 2006). This is important when designing interfaces based on fNIRS sensor data, as the interface would have to respond in this time scale. While the possibility of using event-related fNIRS has been explored (Herrmann, et al., 2008a), most studies take advantage of the slow response to measure short term cognitive state, instead of instantaneous ones.

3.2 General Experimental Protocol

Understanding how the potential noise sources described above affect fNIRS data during cognitive tasks is critical for proper use of fNIRS in HCI research. Thus, we devised a study to empirically test whether or not several common behavioral factors interfere with fNIRS measurements. Specifically, we selected typical human behaviors (head and facial movement) and computer interaction (keyboard and mouse usage), to determine whether each of them needs to be controlled, corrected, or avoided at all cost. This will help us determine whether standard interfaces can be used along with fNIRS in real brain-computer interfaces.

We will call each of the examined physical actions *artifacts*, since they are not the targeted behavior we would like to detect with fNIRS. Using fNIRS, we measured brain activity as these artifacts were introduced while the participant was otherwise at rest, as well as while the participant was performing a cognitive task. We then compared these results to signals generated while the participant was completely at rest with no artifact, as well as to when the participant performed the cognitive task without the artifact. This allowed us to determine whether the artifact had an influence on the signal generated in a rested state, as well as if it has an impact on the signal during activation.

For each artifact, there were four conditions tested as described above: (A) a baseline with no cognitive task or artifact; (B) the cognitive task alone with no artifact; (C) the artifact alone with no cognitive task; and (D) the cognitive task along with an artifact (see Figure 3-2).

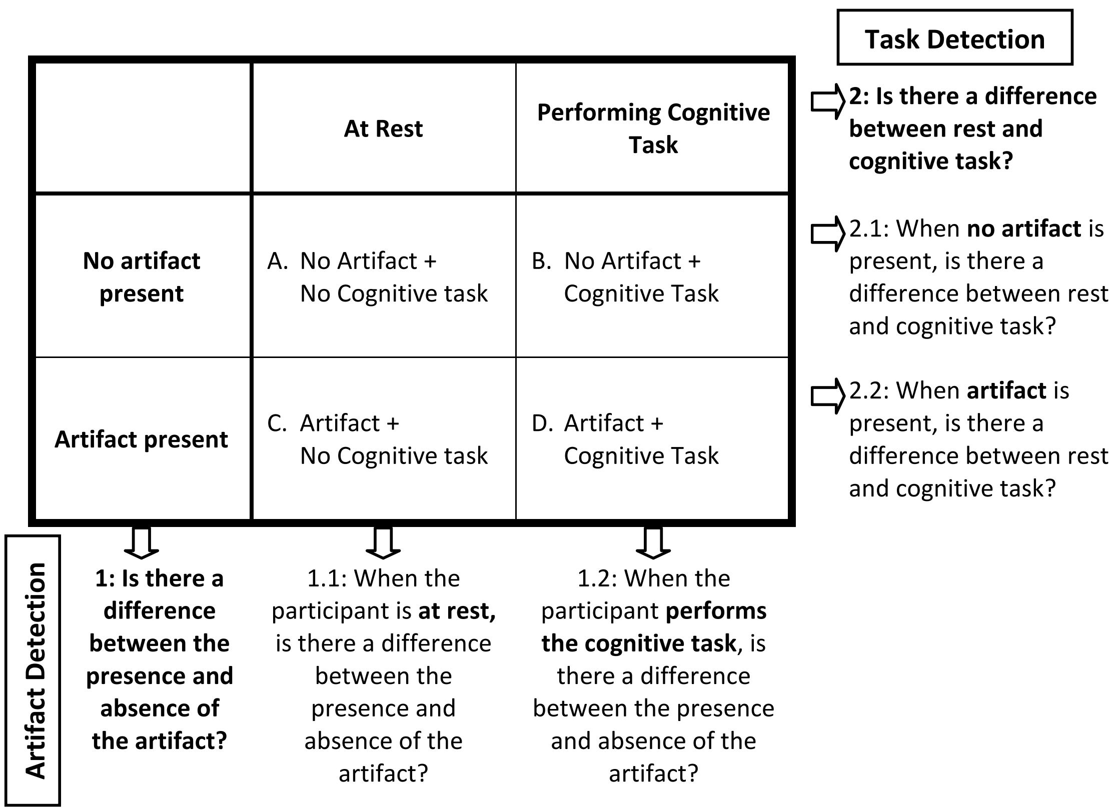

Figure 3-2. Letters A, B, C, and D show the conditions tested. The numbered questions indicate the comparisons between the conditions done in the analysis.

Our goal in designing the protocol for each artifact was to reproduce realistic occurrences. As these artifacts do not necessarily happen often, we tried to balance conservatism (i.e. highly exaggerated artifact) with optimism (i.e. minute occurrence of artifact), and chose a reasonable exaggeration of the artifact, maximizing the possibility of measuring the artifact if it can be measured, yet keeping the conditions somewhat realistic.

3.2.1 Participants

Ten participants took part in this study (mean age = 20.6, std = 2.59, 6 females). All were right-handed, with normal or corrected vision and no history of major head injury. They signed an informed consent approved by the Institutional Review Board of the university, and were compensated for their participation.

All participants completed the five experiments described below in one sitting. They were given small breaks between each part, while wearing the probes. The study is within subject (each participant did all the experiments and conditions), and was counterbalanced to eliminate bias due the order of the experiments, and the conditions.

3.2.2 fNIRS Apparatus

We used a multichannel frequency domain OxiplexTS from ISS Inc. (Champaign, IL) for data acquisition (Figure 3-3). We used two probes on the forehead to measure the two hemispheres of the anterior prefrontal cortex (see Figure 3-4). The source-detector distances are 1.5, 2, 2.5, 3cm respectively. Each distance measures a different depth in the cortex. Each source emits two light wavelengths (690nm and 830nm) to pick up and differentiate between oxygenated hemoglobin ([HbO]) and deoxygenated hemoglobin ([Hb]). The sampling rate was 6.25Hz. We use the term *channel* to define a source-detector distance.

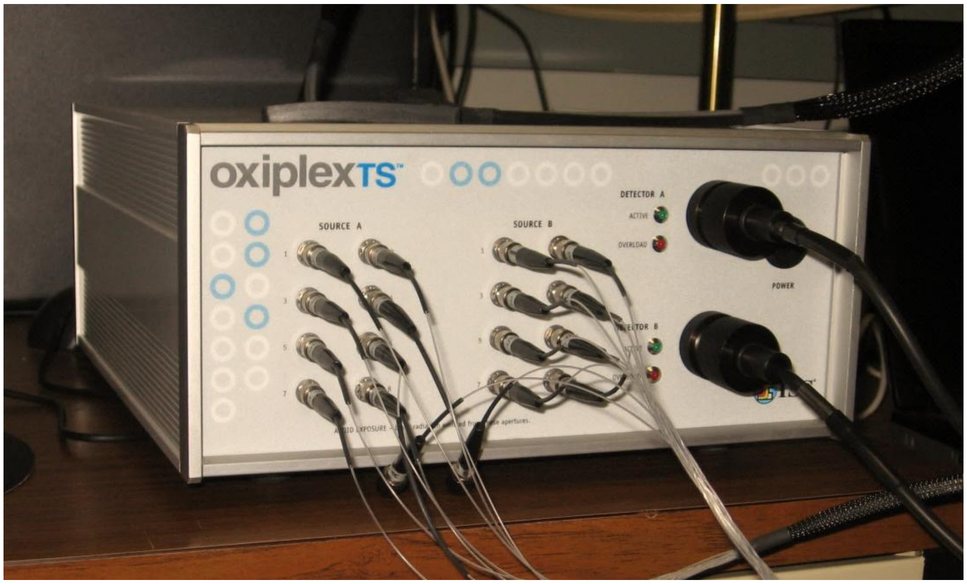

Figure 3-3. fNIRS Equipment.

The two optical probes were placed on the middle of the forehead of participants on either side by use of an elastic headband to keep contact between the fibers and the scalp, as shown Figure 3-1. Note that the discomfort associated with wearing the probes across one's forehead is minimal. Our probe is made of rubber, offering a comfortable sensor that isolates well the ambient light.

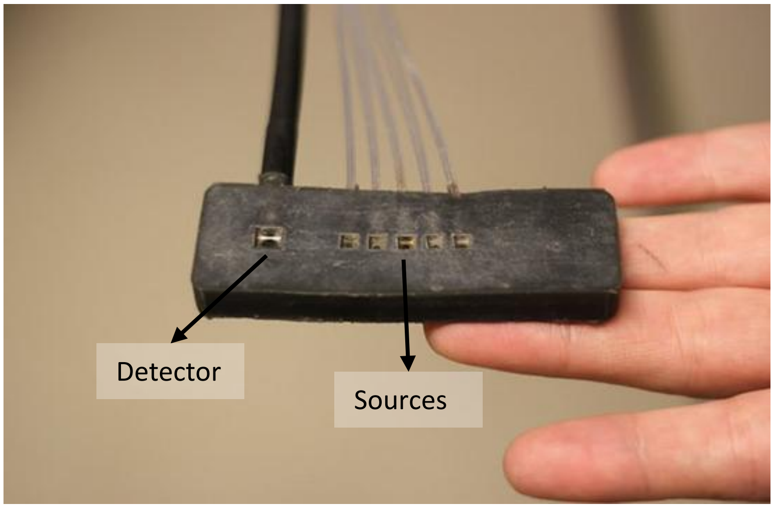

Figure 3-4. A picture of the left probe. A probe includes a detector and light sources.

In previous studies using a similar, linearly arranged probe, researchers have chosen to use data from the furthest two channels only, in order to guarantee that the depth of the measurement reached the cortex. While it is likely that the shallower channels pick up systemic responses, or other noise sources, we decided to keep the data from all four source-detector distances measured as they might help separate out artifacts from task activation.

In all the experiments, the participants were at a desk with only a small lamp (60 W) beside the desk turned on, and they were sitting at a distance of roughly 30” from a 19” flat monitor. The room was quiet, but was not soundproof and noise from the hallway outside the laboratory could be heard occasionally. The participants were instructed to keep their eyes fixated on one point on the screen, and to refrain from speaking, frowning or moving their limbs, unless instructed otherwise.

3.2.3 Procedure and Design

There were five different experiments conducted with each participant, all in one session. These corresponded with the four artifacts being studied (keyboard input, mouse input, head movement, and facial movement), plus the tasks without any artifact present. In between each experiment, the participant could take a break. Although the descriptions below are numbered as Experiments 0, 1, 2, 3, 4, the ordering of the experiments was counterbalanced between subjects. The main difference between the experiments was which additional physical artifact, if any, was introduced as the participant performed the two tasks.

3.2.4 Cognitive Task

All five experiments used the same cognitive task. At the beginning of each trial, the participants were shown a 7-digit number on the screen for four seconds. The number then disappeared from the screen, but the participants were instructed to remember it in their head. After 15 seconds, the participants were asked to enter as much of the number as they could remember.

The goal of the cognitive task used in these experiments was to provide a common task that participants would perform in all experiments, which yields a brain signal that could be detected with fNIRS. We choose a simple verbal working memory task because previous fNIRS studies have reported this type of task to produce a clear and consistent brain signal across participants (Ehlis, et al., 2008; Hirshfield, et al., 2009b). Many studies have successfully shown discrimination of two (or more) states, and we believe our results will generalize to those as well.

3.3 Experiment 0: No artifacts

This experiment consisted primarily of the cognitive task and rest periods. No additional artifact was introduced. This experiment was used to verify that we could distinguish the fNIRS data while the participant was at rest from the fNIRS data while the participant performed the cognitive task, when no artifact was present.

First, the researcher read instructions to the participants, explaining the two tasks that they would perform in the experiment. Then the participants were presented with a practice trial which included an example of each task in that experiment, so the participants would know what to expect. The participants then relaxed for one minute, so their brains could be measured at a normal, rested state. During this period, as well as all other rest periods, there was a black screen and participants were instructed to focus their eyes on the focal point and relax, clearing their heads of any thoughts. This was followed by ten trials.

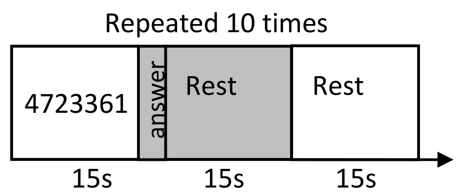

Figure 3-5. Experiment 0 (No artifacts).

The white areas represent the two conditions analyzed. The answer period's length was variable.

A trial contained one 15s condition with the cognitive task, followed by a 15s rest period to allow the participant's brain to return to a rested state. In addition, there was a 15s condition without the cognitive task in which the participant was essentially at rest (see Figure 3-5). These conditions were counterbalanced so that sometimes participants started with the cognitive task, and sometimes they started without the cognitive task.

3.3.1 Preprocessing

The preprocessing step transforms the raw data from the device into hemoglobin values, and smoothes the data to remove any high-frequency noise, as well as heartbeat. We chose to filter the data in these experiments because this is a standard step in fNIRS experiments, and the goal was to determine the influence of interaction techniques and artifacts on a typical fNIRS experiment. We applied a simple preprocessing procedure. We used a non-recursive time-domain band-pass filter, keeping frequencies between 0.01-0.5 Hz (Folley & Park, 2005). The data was then transformed to obtain oxy- ([HbO]) and deoxy-hemoglobin ([Hb]) concentration values, using the modified Beer-Lambert law (Villringer & Chance, 1997). The law governs the influence of light absorption and scattering on optical measurements, and states that the change in light attenuation is proportional to the changes in the concentrations of oxy- and deoxy-hemoglobin. It should be noted that the combination of [HbO] and [Hb] gives a measure of total hemoglobin, which we will refer to as [HbT]. We averaged each trial in two seconds periods, to obtain seven averaged points we call *Time Course*. All ten trials from all subjects were included in the analysis. Figure 3-6 displays a typical example of the data for those two tasks.

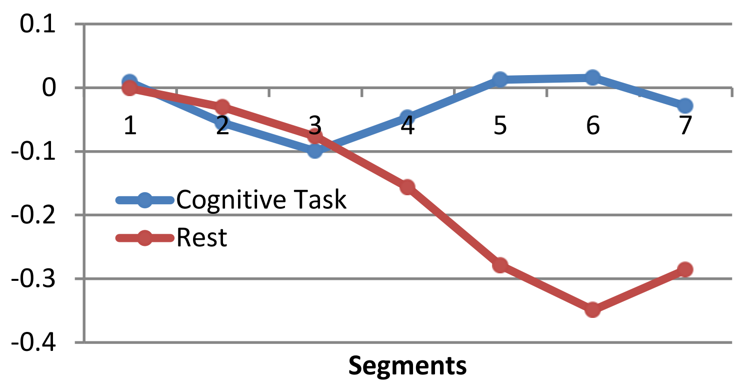

Figure 3-6. 7 data points time series example for typical rest and cognitive load tasks.

3.3.2 Analysis

In this experiment, we wanted to observe whether the cognitive task, on its own, yielded a brain signal that was distinguishable from the signal during a rested state. This result is fundamental to all the other experiments that include the cognitive task. If we were not able to significantly distinguish the cognitive task from rest with no added artifacts, it would have been difficult to distinguish the two when additional noise was introduced into the data.

To evaluate the presence of the cognitive task in the data, we choose to perform a statistical analysis through an analysis of variance. This type of ANOVA is designed to uncover the main and interaction effects of independent variables on a dependent variable. In our case, we have five independent variables: the condition performed (the cognitive task or the rest task), the hemisphere (left or right), the channel (labeled 1 to 4, from the shortest source-detector distance to the furthest), and the time course (7 sequential data points), as well as in some subset of the tests the type of hemoglobin (oxy- or deoxygenated). Our dependent variable is the amount of light measured. In lay terms, the analysis will observe whether any of those factors, or the combination of them, show significance, meaning that there is a difference in the data between the groups. For example, if the factor hemisphere is significant, this means the data shows a difference in values between the left and the right hemisphere. If the interaction of hemisphere and channel is significant, it would indicate that a combination of the two factors is significant, which could mean that the left channel 1 is different than the right channel 3. More combinations of elements can be significant in an interaction, two were given here as an example.

First, this dataset and all reported datasets in this chapter were tested for conformity with the ANOVA assumption of normality by creating a normal probability plot, on which normal data produces a straight or nearly straight line, confirming that the

ANOVA is an appropriate test of significance. We omit the inter-subject variability testing as it is always positive in brain studies.

We did a factorial repeated measures ANOVA on *Cognitive Task* (cognitive task or rest) x *Hemisphere* (left or right) x *Channel* (4) x *Time Course* (7). This identifies differences within each participant, and determines if those differences are significant across participants. This is Comparison 2.1 in Figure 3-2. We ran this analysis with [HbO], [Hb] and [HbT] data separately, as well as together by including *Hemoglobin Type* as a factor. While we did a factorial ANOVA, we are most interested in results that show significant interactions including the *Cognitive Task* factor, since these show significant differences between the signal during the cognitive task and the signal during rest. In this analysis, and all those following, we will only report significant results ($p<0.05$) that are pertinent to current HCI questions. The full statistical results can be found in Appendix A-1.

3.3.3 Results

From these three analyses, the only relevant significant factor found was with [Hb], *Cognitive Task x Channel* ($F(3, 27)= 5.670$, $p= 0.031$). This confirms that levels of [Hb] differ between trials where participants performed a cognitive task, and trials where they simply rested, and that this difference in [Hb] levels varied by channel. Therefore, one might hope that using measurement of different source-detector distances (channels) we can distinguish the cognitive task versus the rest tasks, thus feeding this decision to an HCI system. However, [HbO] and [HbT] did not find this interface significant, indicating that our cognitive task might show weak brain signal differentiation in the region measured. We believe we can go forward with the rest of the analysis because of the positive result obtained with [Hb].

3.4 Experiment 1: Keyboard Input

The keyboard and mouse are the most common input devices for modern computers. We tested keyboard input in Experiment 1 and mouse input in Experiment 2. We hypothesized that keyboard inputs would not be a problem with fNIRS, since most brain activation for motor movement occurs in the motor cortex, an area not probed with our sensors. In addition, we did not believe that the physical act of typing would cause the sensors to move out of place or change the blood oxygenation characteristics in the PFC.

We decided not to have participants type specific words because we were only interested in measuring the influence of the typing motions on the signal, instead of any brain activity associated with composing and typing text. They were instructed to randomly type on the keyboard, using both hands, at a pace resembling their regular typing pace, including space bars occasionally to simulate words.

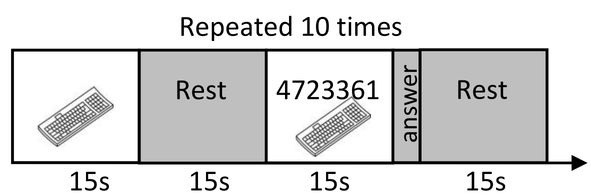

Figure 3-7. Experiment 1 (Keyboard Input).

The white areas represent the two conditions analyzed in the experiment.

The protocol was analogous to Experiment 0. The main difference is that in the task, the participant was also typing randomly as described above (see Figure 3-7). We do not include a condition combining the cognitive task with no artifact as it has been successfully tested in Experiment 0. We reuse those results in the analysis of each artifact.

3.4.1 Analysis

To observe the influence of typing on the brain data, we examined the data in several different ways, corresponding with the numbers in Figure 3-2. Comparison 1 determines whether there is a difference between typing and not typing, **regardless of whether there was cognitive task**. Comparison 1.1 examines whether there is a difference in the fNIRS data between the presence and absence of the typing artifacts when the participant is **at rest**. Comparison 1.2 determines whether there is a difference between the presence and absence of the typing artifacts when the participant **performs the cognitive task.** Comparison 2 determines whether there is a difference between doing a cognitive task and no cognitive task, **regardless of whether the participant was typing**. Comparison 2.2 looks at whether there is a difference between rest and cognitive task **when typing artifacts are present**. Note that the comparison 2.1 was not examined in Experiments 1 to 4, as there are no artifacts present in this condition. We use the results of Experiment 0 for the comparison 2.1 in the analysis of each artifact.

As in Experiment 0, we were most interested in results that showed significant interactions including the *Cognitive Task* factor, since these show significant differences between the signal during the cognitive task and the signal during rest. In addition, we were interested in significant interactions that included the artifact *Typing*, since these show significant differences between when the subject was typing and when the subject was not typing.

Comparison 1, 1.1 and 1.2 used the interaction *Typing* (present or not) x *Hemisphere* (left or right) x *Channel* (4) x *Time Course* (7); Comparison 1.1 uses data from rest tasks; Comparison 1.2 uses data during cognitive tasks; while Comparison 1 uses both datasets. Comparisons 2 and 2.2 used the interaction *Cognitive Task* (cognitive task or rest) x *Hemisphere* (left or right) x *Channel* (4) x *Time Course* (7). Comparison 2.2 used data containing typing while Comparison 2 used data both with and without typing.

Ideally, we would observe the absence of *Typing* as a factor in significant interactions for Comparisons 1, 1.1, and 1.2. For Comparisons 2 and 2.2, ideally we would find *Cognitive Task* as a factor in significant interactions, as this indicates the ability to distinguish the presence or absence of a cognitive task.

For each comparison, we analyze the data for [Hb], [HbO] and [HbT] separately, as was done for Comparison 1 in Experiment 0.

3.4.2 Results

Task Detection—In Comparison 2, we found *Cognitive Task x Hemisphere* to be significant with [Hb] data ($F(1, 9)= 5.358$, $p= 0.046$. This indicates that when typing and not typing tasks are combined, we can determine whether the participant is performing a cognitive task or not using the right hemisphere. In Comparison 2.2, [Hb] yielded significance with *Cognitive Task x Hemisphere* ($F(1, 9)= 5.319$, $p= 0.047$). Comparison 2.2 demonstrates that given typing, we can distinguish whether the participant is also performing a cognitive task or not, specifically using the [Hb] data and looking at both hemispheres.

Artifact Detection—Comparison 1 showed significance for *Typing x Time Course* with [HbO] ($F(6, 54)= 3.762$, $p= 0.034$), meaning that with cognitive task and rest tasks combined, we can distinguish typing using how they change over time (time course). We did not observe any significant interaction that included *Typin*g in Comparison 1.1. We can conclude that at rest, there is no significant difference in the fNIRS signal between typing and not typing. We found that for Comparison 1.2, [Hb] data revealed significance with *Typing x Hemisphere x Channel* ($F(3, 27)= 3.650$, $p= 0.042$). We find *Typing x Hemoglobin Type x Time Course* to be significant ($F(6, 54)= 6.190$, $p= 0.012$). These results show that when the participant is performing a cognitive task, there is a difference whether the participant is also typing or not, as typing shows up in significant interactions.

3.4.3 Discussion

Comparison 1.1 confirmed that the sensors are not picking up a difference between the typing task and rest. However, in Comparison 1.2, we found that typing is influenced by the cognitive task. This is also true in general, as typing tasks are usually related to the current task.

Overall, while typing can be picked up when there is a cognitive task present, we can still distinguish the cognitive task itself (Comparison 2.2 and 2). This confirms our hypothesis and validates that typing is an acceptable interaction when using fNIRS. From this, we can also assume that simple key presses (e.g. using arrow keys) would also be acceptable with fNIRS since it is just a more limited movement than typing with both hands.

3.5 Experiment 2: Mouse Input

We designed a task that tests mouse movement and clicking. We hypothesized that small hand movement such as using the mouse would not interfere with fNIRS signal. The participant was instructed to move a cursor until it was in a yellow box on the screen, and click. The box would then disappear and another one would appear somewhere else. Participants were directed to move at a comfortable pace, not particularly fast or slow, and to repeat the action until the end of the condition. All participants used their right hand to control the mouse.

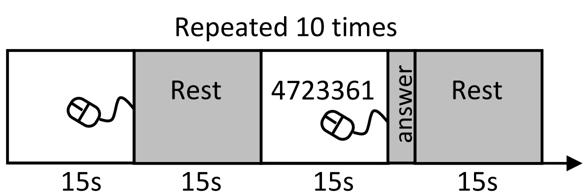

Figure 3-8. Experiment 2 (Mouse Input).

The procedure was identical to Experiment 1, except that the typing was replaced with mouse clicking (see Figure 3-8). We analyzed the data using the same comparisons as in Experiment 1, substituting mouse input for keyboard input.

3.5.1 Results

Task Detection—Comparison 2 yielded no significant interactions, indicating that we cannot distinguish between rest and cognitive task, when the data includes both clicking and not clicking. In Comparison 2.2, we found both *Cognitive Task x Hemisphere x Hemoglobin Type* ($F(1, 9)= 5.296$, $p= 0.047$) and *Cognitive Task x Hemisphere x Hemoglobin Type x Time Course* ($F(6, 54)= 4.537$, $p= 0.036$) to be significant, indicating that even in data containing clicking, we can tell whether the participant is doing a cognitive task or resting.

Artifact Detection—Comparison 1 yielded no significant interactions, indicating that we cannot observe differences between the presence and absence of clicking, when combining data from the cognitive task and rest. In Comparison 1.1, with [Hb], we observe an interaction of *Clicking x Channel* ($F(3, 27)= 4.811$, $p= 0.044$). This shows that we can tell whether someone is clicking when looking at specific channels, with the participant being at rest (Figure 3-9).

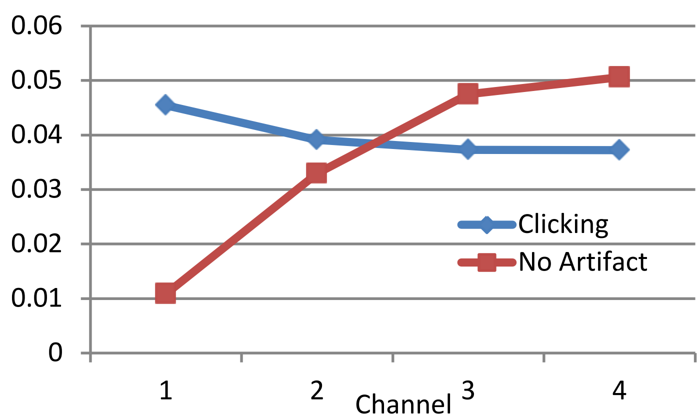

Figure 3-9. Mean Plots of *Clicking x* Channel for [Hb].

In Comparison 1.2, [HbO] data reveals significant interaction with *Clicking x Hemisphere* ($F(1, 9)= 9.599$, $p= 0.013$) and *Clicking x Hemisphere x Time Course* ($F(6, 54)= 4.168$, $p= 0.037$). This indicates the ability to distinguish *Clicking* from no motor activity when the participant is performing a cognitive task, although this effect differs across hemispheres. Finally, we observed significant interactions with *Clicking x Hemisphere* with [HbT] ($F(1, 9)= 6.260$, $p= 0.034$) and *Clicking x Hemisphere x Hemoglobin Type* ($F(1, 9)= 5.222$, $p= 0.048$), which leads to the same conclusion as with [HbO] data only. Overall, we can tell whether someone is clicking depending on the brain hemisphere. Specifically, the left hemisphere is significant at distinguishing the two states, as illustrated in Figure 3-10.

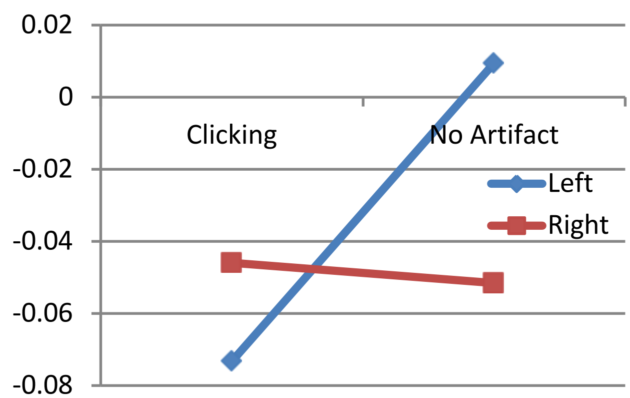

Figure 3-10. Mean plots for *Clicking x Hemisphere* for [HbO].

3.5.2 Discussion

We found that clicking in this experiment might affect the fNIRS signal we are collecting, as Comparison 1.1 yielded interactions with the factor of clicking. This means that when the participant is at rest, there is a difference between the presence and absence of clicking. The difference in activation is not surprising as we did not have a “random clicking” task, but one where subject had to reach targets, which may have activated the anterior prefrontal cortex. However, because Comparison 2.2 still was able to distinguish *Cognitive Task*, the cognitive task of remembering numbers may produce a different signal from clicking.

While the hand movements of clicking and typing are not identical, we also believe the core difference between the clicking experiment and typing experiment is mainly due to the fact that clicking involved some brain activity and typing was random. This explains why did observe the presence of the artifact in rest-only conditions.

Hence, results indicate that when we want to observe a cognitive task that contains clicking, we need to have the rest task contain clicking as well, as Comparison 2.2 found significant interactions, but Comparison 2 did not. In short, we need to know whether the user is clicking in order to distinguish the cognitive task. Luckily, this information is easily obtained by adding mouse events to our analysis. Overall, we believe that clicking is acceptable if the experiment is controlled, confirming in part our hypothesis.

3.6 Experiment 3: Head Movement

General head movements could affect the fNIRS signal, both because of possible probe movement on the skin, and possible change in blood flow due to the movement itself, as was noted earlier. We hypothesize that head movement could be a problem, as this seems to be reported by many researchers.

Many types of head movements can occur, in all directions. We chose a condition that is representative of common movement while using the computer: we simulated looking down at the keyboard and up at the screen. These movements were done in an intermittent manner, similar to head movements that may occur during normal computer usage, three times per 15s trial.

The procedure was identical to Experiment 1 and 2, except that the typing or mouse clicking was replaced by the head movement (see Figure 3-11). We analyzed the data using the same comparisons as in Experiment 1 and 2, substituting head movement for keyboard or mouse input.

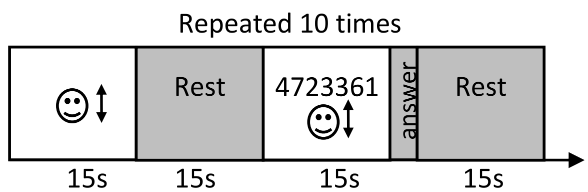

Figure 3-11. Experiment 3 (Head Movement).

3.6.1 Results

Task Detection—We found no significant interactions for Comparison 2, meaning that it is not possible to separate the cognitive task from rest when including both data with head movements and data without head movements. In Comparison 2.2, we find that *Cognitive Task x Hemoglobin Type x Channel x Time Course* is significant (F(18, 162)= 3.915, p= 0.048). With head movements, there is a difference between rest and the cognitive task.

Artifact Detection—We found no significant interactions for Comparison 1, which indicates that it is not possible to distinguish between the presence and absence of head movements when the cognitive and rest data are combined. There were no significant results for Comparison 1.1, indicating that at rest, there is no significant difference in the signal when the participant is moving his or her head or not. Comparison 1.2 showed that with [Hb] data, we can distinguish *Head Movement x Hemisphere x Channel* ($F(3, 27)= 5.363$, $p= 0.028$), and we can significantly observe *Head Movement x Hemoglobin Type x Time Course* ($F(6, 54)= 7.455$, $p= 0.002$), meaning that during the cognitive task, we can tell between the participant moving their head or not.

3.6.2 Discussion

Similar to the clicking results, we found that we require the presence of head movements in both the rest and the cognitive task to distinguish it (Comparison 2.2), which leads us to suggest that head movement should be avoided. However, the movements in this experiment were more exaggerated and frequent than regular moving from keyboard to screen: for example, most subjects could not see the screen when looking at the keyboard. We suggest that participants minimize major head movements, and instead move their eyes towards the keyboard. We found our initial hypothesis correct, although we believe head movement may be minimized and corrected using filtering techniques. This conclusion is based on our experiment and on the work of Matthews et al. (2008).

3.7 Experiment 4: Facial Movement

Forehead facial movement moves the skin located under the probe, which may interfere with the light sent into the brain and its path. We hypothesize that forehead facial movement, e.g. frowning, will have an effect on the data.

In this experiment, participants were prompted to frown for two seconds, every five seconds. Specifically, we asked them to draw the brows together and wrinkle the forehead, as if they were worried, angry, or concentrating.

The procedure was also identical to the other experiments, except that the artifact introduced was head movement (see Figure 3-12). We analyzed the data using the same comparisons as in the other experiments, substituting frowning motion for keyboard or mouse input, or head movement.

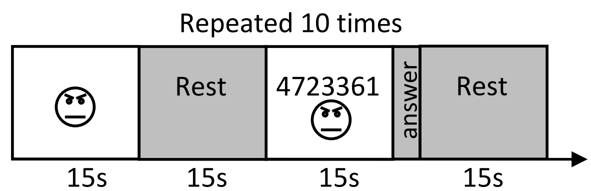

Figure 3-12. Experiment 4 (Facial Movement).

3.7.1 Results

Task Detection—Comparison 2 found *Cognitive Task x Channel x Time Course* to be significant with [HbO] (F(18, 162)= 3.647, p= 0.043). *Cognitive Task x Hemoglobin Type x Channel x Time Course* was a significant interaction (F(18, 162)= 4.130, p= 0.042), both indicating that when frowning data is combined with not frowning, we can tell the cognitive task from rest at some but not all channels. Finally, Comparison 2.2 showed no significance for interactions that included *Cognitive Task*, indicating we cannot distinguish the cognitive task from rest when the subject is frowning.

Artifact Detection—Comparison 1 showed significance with [HbO] for *Frowning x Channel* (F(3, 27)= 5.287, p= 0.035). We found significance with *Frowning x Channel* with

[HbT] (F(3, 27)= 5.343, p= 0.035), *Frowning x Hemoglobin Type x Channel* (F(3, 27)= 4.451, p= 0.046). We see that regardless of whether at rest or doing cognitive task, we can distinguish whether frowning is occurring at some but not all channels (Figure 3-13), which is consistent with previous results.

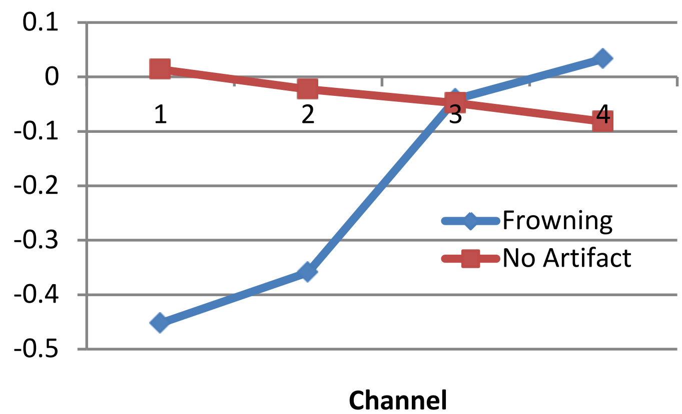

Figure 3-13. Mean Plots in *Frowning x Channel* for [HbO].

In Comparison 1.1, we found that [HbO] data showed *Frowning x Channel* to be significant (F(3, 27)= 5.194, p= 0.037), which we also noticed with both types of hemoglobin (F(3, 27)= 5.191, p= 0.037). When the participant was at rest, we can distinguish whether the participant is frowning or not at some but not all channels (Figure 3-14 plots a typical example of frowning). Comparison 1.2 found *Frowning x Channel* to be significant for [HbO] data (F(3, 27)= 4.862, p= 0.042) and with both types of hemoglobin (F(3, 27)= 4.978, p= 0.041). This indicates that there is a difference in [HbO] levels when participants were frowning or not frowning, and that this difference varied by channel, similarly to Comparison 1.1.

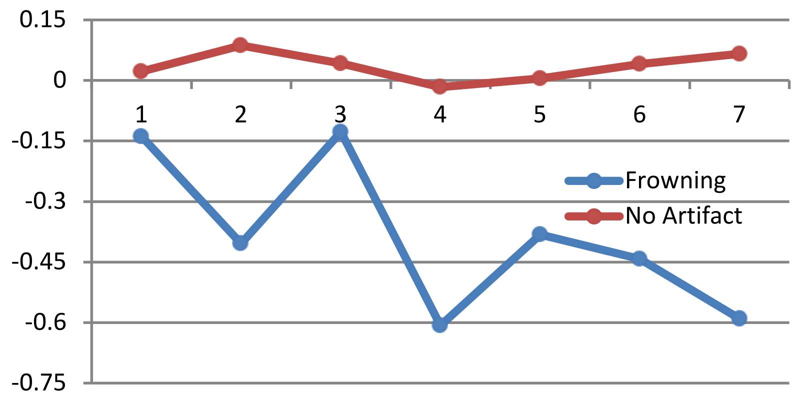

Figure 3-14. Typical example of frowning.

3.7.2 Discussion

We found that frowning data always can be distinguished from non-frowning. We also learned that if all the data includes frowns, then we cannot tell apart the cognitive task from the rest condition. However, we found that if we mix the data that contains frowning and no frowning, we can then discriminate the cognitive task, which shows interesting potential. Those results indicate clearly that frowning is a problematic artifact, and should be avoided as much as possible. This confirms our hypothesis.

3.8 Performance Data

In all five experiments, after each cognitive task, participants entered the 7-digit number that they had been remembering. To obtain the error rate of those answers, we compared each digit entered to the original digit, and found the number of digits correctly answered. Figure 3-15 shows the number of digits correctly answered averaged over all subjects, for each experiment. A repeated measures ANOVA examining the error rate across artifact types revealed no statistical differences between them (F(4,36)= 0.637, p= 0.526). This result indicates that each experiment was of similar difficulty.

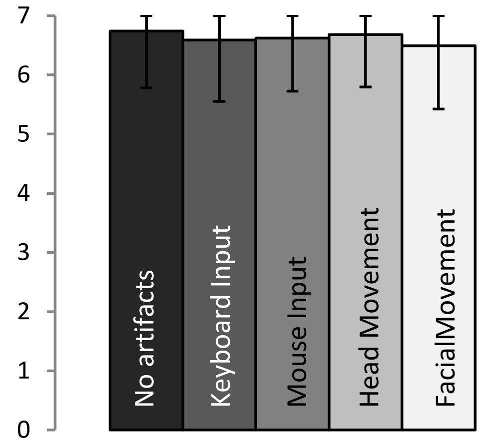

Figure 3-15. Average number of correct digits, with standard deviation.

3.9 Guidelines for fNIRS in HCI

To take advantage of the benefits of fNIRS technology in HCI, researchers should be aware of several considerations, which were identified in this chapter, and summarized in Table 3-1. Our goal was to reveal whether or not several common behavioral factors interfere with fNIRS measurements. We empirically examined whether four physical behaviors inherent in computer usage interfere with accurate fNIRS sensing of cognitive state information. Overall, we found that given specific conditions, we can use typing and clicking in HCI experiments, and that we should avoid or control major head movements and frowns. Through our clicking experiment, we may extrapolate that non-random artifact must be present in rest conditions as well as cognitive tasks, to maximize differentiation.

Table 3-1 . Summary of fNIRS considerations for HCI.

Results Legend: ✓ indicates acceptable, C indicates to correct,

and ✗ indicates to avoid or control.

Considerations	**Result**	**Reference**	**Correction Methods**
Forehead movement	✗	Exp 4	
Major head movement	✗	Exp 3	Use chin rest
Minor head movement	C	Exp 3, (Matthews, et al., 2008)	Filter
Respiration and Heartbeat	C	(Coyle, et al., 2004; Matthews, et al., 2008)	Filter
Mouse Clicking	✓	Exp 2	Collect signal during a clicking only task (rest task)
Typing	✓	Exp 1	
Ambient Light	C	(Chenier & Sawan, 2007)	Wear isolating cap
Hemodynamic Response	✓	(Bunce, et al., 2006)	Expect 6-8s response
Ambient Noise	C	(Morioka, et al., 2008; Wakatsuki, et al., 2009)	Minimize external noise
Eye Movement and Blinking	✓	(Izzetoglu, et al., 2004b)	

Other artifacts, such as minor head movements, heartbeat and respiration may be corrected using filtering. There are many types of filtering algorithms that can help reduce the amount of noise in data (Matthews, et al., 2008). Methods include adaptive finite impulse response (FIR) filtering, Weiner filtering (Devaraj, et al., 2004; Izzetoglu, et al., 2005a), adaptive filtering (Devaraj, et al., 2004) and principal component analysis (Huppert & Boas, 2005; Matthews, et al., 2008; Sitaram, et al., 2007). Matthews et al. (2008) note that FIR can be used in real time if accelerometers are used simultaneously on the head to record head motion. The other methods are mainly offline procedures, making them less practical for real-time systems.

The experimental protocol was designed to reproduce realistic occurrences of artifacts that might be present during typical computer usage in HCI laboratory settings. We purposefully exaggerated the artifacts to make sure they would be measured with fNIRS. So, we need to keep that in mind as the exaggerated artifacts are less likely to happen than in real experiments. Note that this was run in a typical, quiet office space, and not in a sound proof room like most brain sensing studies.

In the future, it would be worthwhile to take these results a step further, to investigate even more realistic settings with multiple potentially interfering sources of noise. In addition, it would be useful to investigate using machine learning to identify the presence of artifacts in fNIRS data. With a database of undesirable artifacts in fNIRS signals, we could feed data from a new experiment to see whether any of the artifacts are found. This could provide a new and objective way to remove examples contaminated by such artifacts, instead of using visual observation.

In conclusion, we have confirmed that many restrictions such as long setup time, highly restricted position, intolerance to movement, and other limitations, that are inherent to other brain sensing and imaging devices are not factors when using fNIRS. However, major head movements and frowning present an unacceptable source of noise in the data. By using the guidelines described above, researchers can have access to the user's cognitive state in realistic HCI laboratory conditions. This is important for adoption in HCI, and we recommend fNIRS as a valuable and effective input technology.

Chapter 4:

Exploring Mental Workload and Interaction Style[2]

We showed in Chapter 3 that fNIRS is a viable tool for HCI settings. The goal of this chapter is to explore its ability to measure a signal with strong potential for HCI. We are also interested in applying machine learning to automatically classify the brain states measured.

[2] The work in this chapter was partially described in Hirshfield, et al. "Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy" in the proceedings of the ACM UIST'07 Symposium on User Interface Software and Technology, (2007). This was joint work with Leanne Hirshfield and Erin Solovey.

Past research shows the potential for fNIRS to measure frontal lobe activity such as workload (Hirshfield, et al., 2009b; Izzetoglu, et al., 2004b; Izzetoglu, et al., 2005b; Luu & Chau, 2009; Son, et al., 2005). We present a study designed to distinguish several discrete levels of workload that users experience while completing a given set of tasks. We chose to evaluate several degrees of load as they are often associated with different tasks, and determining underload or overload situations can be beneficial in many real life interactions (Guhe, et al., 2005; Iqbal, et al., 2004; John, et al., 2004). With this new technique, we hope to provide objective measures of workload instead of the more classic subjective assessments. In the study, we use a standard task with varying workload levels that are cross-validated with an established measure of workload, the NASA-Task Load Index (Hart & Staveland, 1988).

We use machine learning techniques to analyze fNIRS data to classify up to four levels of mental workload. The hypothesis driving the study is that useful features extracted from fNIRS output could be combined with machine learning models to accurately determine workload levels that the user was experiencing when completing a task in HCI. Machine learning classification techniques were selected as they add a level of abstraction to the dataset, permitting researchers without fNIRS domain expertise to extract meaningful user states from the brain data.

Subjects completed thirty tasks where they viewed the top and all sides of a rotating three dimensional (3D) shape comprised of eight small cubes. The sides of the cubes within each shape were colored. In the experiment, cubes could be colored with two, three, or four colors. Possible colors were green, yellow, red and blue, all easily distinguishable for a non-colorblind person. Figure 4-1 illustrates an example of a rotating shape, with four colors.

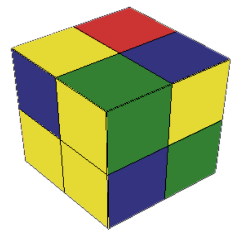

Figure 4-1. A cube made up of eight smaller cubes.

During each task, subjects counted the number of squares of each color displayed on the rotating shape in front of them. The shape rotated three times in increments of 90°, allowing the subject to view each side only once (a 270° rotation) but the top is visible at all times. During the rotation, each side of the cube was displayed for nine seconds. Subjects did not view the bottom of the shape, resulting in a total of twenty visible squares of different colors in each rotation. Rotation time and the layout of the shape were controlled during the experiment.

To vary workload, we changed the number of colors present on the rotating shape. As the number of colors in the shape increased, it was necessary for subjects to keep more items in working memory to remember how many squares of each color had been viewed. There were four workload conditions. In the *workload level 0* condition (WL0), subjects were asked to clear their minds and think of nothing. In the other three workload conditions subjects counted the colors on a rotating shape with two, three, and four colors. We refer to these conditions as WL2, WL3 and WL4. We did not use WL1 (one color) because of its triviality, as the answer would always be 20.

These conditions were chosen because of their potential relevance in the realm of HCI (Figure 4-2). Through pilot studies, we hypothesize that workload level 0 resembles a condition of user underload. Workload level 2 represents a situation when users were experiencing a normal level of workload (almost always producing the correct answer after the task completed). Workload level 4 corresponds to a condition of user overload (subjects usually lost track of their numerical counts and answered incorrectly on these tasks). WL3 conditions produced mixed results in our pilot studies, with subjects answering some WL3 tasks correctly and others incorrectly.

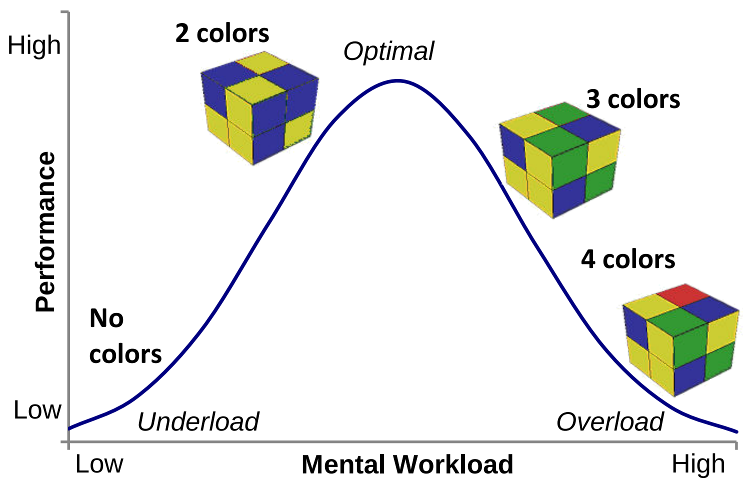

Figure 4-2. Tasks in relation to workload.

The main goal of this experiment was to decide whether fNIRS data is sufficient for determining the workload level of users as they perform tasks. To accomplish this, a graphical user interface (GUI) displayed the rotating shapes described above.

A second goal was to determine whether there is a difference in mental workload when users complete varying spatial reasoning tasks: specifically tasks on a graphical display versus using a physical object, such as a tangible user interface. Prior research on the comparisons between tangible interfaces and graphical user interfaces was the catalyst for inclusion of this condition (Ullmer, Ishii, & Jacob, 2005). One hypothesis is that perhaps the activation might not be located in the same part of the brain. To study this property, we developed physical shapes identical to the three colors graphical shapes (WL3). These physical shapes were rotated for the same amount of time as the graphical shapes on a circular turntable placed in front of subjects. We hypothesized that the WL3 would require less workload with the physical shape than with the graphical shape because humans have some difficulty extracting 3D spatial information from a two-dimensional screen.

Therefore, there were five conditions tested in this experiment, which are outlined in Table 4-1.

Table 4-1. Experimental conditions include workload levels and display type.

Workload Level	*Number of colors*	*Shape*
WL0	0	-
WL2	2	GUI
WL3	3	GUI
WL3 physical	3	Physical
WL4	4	GUI

4.1 Procedure and Participants

Our study was run on five subjects (three females), from 18 to 26 years of age. None of our subjects was colorblind, and four were right handed. We followed the block design used in previous BCI experiments (Keirn & Aunon, 1990; Lee & Tan, 2006): we randomly placed each of the workload conditions into a set (five tasks per set) and each experiment consisted of six sets. Therefore, each subject saw each workload condition six times, one time in each set. The ordering of the conditions was randomized within each set, and per subject.

At the completion of each task, the subject was prompted to state their answer (i.e. "nine blue and eleven yellow"). After stating an answer, the subject was instructed to rest for thirty seconds, allowing his or her brain to return to a baseline state. After completing the tasks, the subject was presented with an additional example of each workload level and asked to fill out a NASA-Task Load Index (TLX) (Hart & Staveland, 1988). NASA-TLX provides a ground truth measurement, a benchmark for comparing and validating fNIRS results. It is a collection of questions relating to the task's mental, physical, and temporal demands on the user, their performance, effort and frustration level when executing the task. We administered the NASA-TLX, commonly used today to subjectively measure user workload, to compare our results with an established measure of workload. This allowed us to validate our workload levels.

4.2 Data Analysis

We collected five datasets, composed of 30 tasks each (six tasks of each workload level), with 16 channel measures at each time point (2 light detectors picking up two types of hemoglobin from four light emitters = 2 detectors x 4 light sources x 2 types of hemoglobin).

4.2.1 Pre-Processing Steps

We used a similar preprocessing technique to that of the previous chapter. We detail the differences in the processing. A Fourier transform was used to offset the trend in the fNIRS sensor readings throughout each task (Akgul, 2005). This trend is composed of very low frequency components (< 3mHz). Data in between tasks was not included in analysis, as participants talked while giving their answer to the task and rested for 30 seconds to allow the blood flow in their brain to return to a baseline state.

We normalized the data using z-normalization (Goldin & Kanellakis, 1995). A time sequence *T* can be normalized as $t_i' = (t_i - mean(T)) / std(T)$. This normalization was done on each channel, to reduce scaling between channels (Kahveci, Singh, & Gurel, 2002). We also cut off four seconds from the beginning of each task from the assumption that it does not contain brain activation information as it takes 4-5 seconds for the blood activation in the brain to be picked up by the fNIRS device (Bunce, et al., 2006).

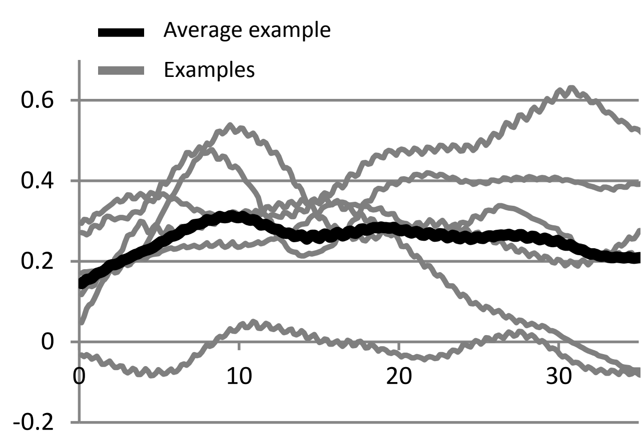

Figure 4-3. Example of fNIRS data for condition WL4.

The black, ticker line indicates the mean of all six trials.

4.2.2 Machine Learning Analysis

We used the sliding windows classification method to automatically produce task predictions. We selected this algorithm in part because it could be transformed for real time classification, a long term goal. The *Sliding Windows* method transforms the data into a time independent dataset, permitting the use of traditional machine learning algorithms (Dietterich, 2002). For each time point, we look at a window of size *w* surrounding that point, including several data points before and several data points after the time point (Figure 4-4). For a time point t_i, a window of size 5 will contain $\{t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}\}$. Windows are given the label of t_i, even if the beginning or trailing points have another class label.

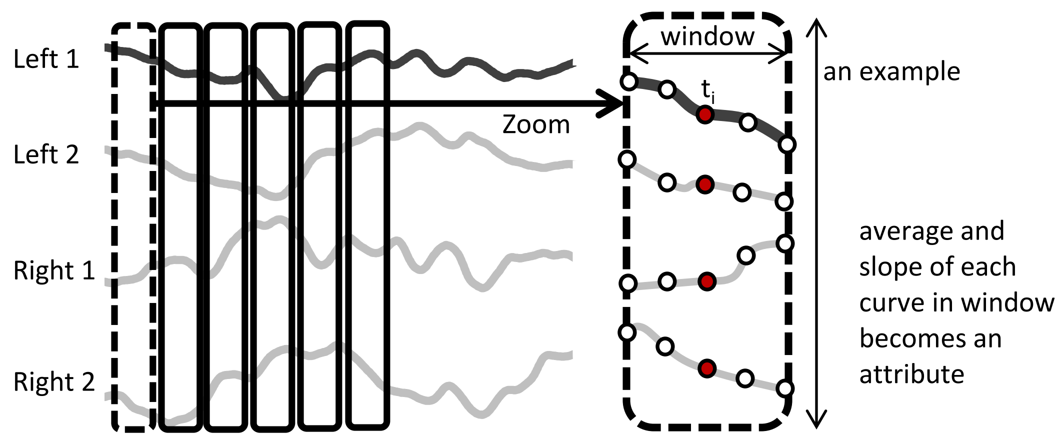

Figure 4-4. The Sliding Windows approach.

Each curve (one collected brain measure) is sliced into task-sized chunks, with each time point as a classification feature.

We generated features for the average and slope of each window. Averaging over each window for each channel smoothed out some of the artifacts in the data from breathing and heartbeat. We find the slope over each window to incorporate the increasing and decreasing nature over that window. We calculate the slope using the averaged values of the time points at the extremities of the window. The process is repeated for every time point, shifting by one time point, creating overlapping windows. Therefore, this resulted in 32 features for each instance (16 channels x 2 features per window).

The Sliding Windows method produces classification examples for approximately every time point. This results in every condition having a large number of examples on which to learn and test.

We selected a window size of 41 (approximately 6 seconds of data). We used the Weka machine learning toolkit (Hall, et al., 2009) to run experiments, with the multilayer perceptron as classification algorithm. Multilayer perception is a neural net with backpropagation.

The sequential nature of brain sensing data is important: measurements occurring near each other in time are closely related, leading to non-independent readings. In our previous example, there is a correlation with a reading at time t_i and the readings at time t_{i-1} and t_{i+1} because the reading corresponds with oxygenation in the blood which changes somewhat gradually. In this case, random sampling during cross validation gives misleading, high classification results since the training and test sets are not independent. For instance, random classification could put t_i in the training set and t_{i+1} in the testing set, which would make it t_{i+1} to be correctly classified. Therefore, we implemented a blocked cross-validation scheme to assess our accuracy (Lee & Tan, 2006) based on our blocked experimental design. There were six sets (of 5 conditions) in the experiment. We created a fold for each set, and we ran cross validation on each possible combination of training on five folds and testing on the unseen sixth fold of data. We averaged the results of these tests together to determine each classifier's accuracy for the current subject.

We were interested in determining whether we could distinguish different workload levels from the fNIRS data alone using machine learning. First, we calculated the presence of brain activity by comparing WL0 (no activity) with each workload level individually. For example, using data from WL0 and WL2, we ran classifiers to determine if we could distinguish the two classes from each other given training and testing data for only those two classes. We then calculated the accuracy of distinguishing each combination of graphical workload levels (three combinations of two levels; four combinations of three levels and one combination of all four levels). For example, we compared WL0, WL3 and WL4. Finally, we tested the classification of all five workload levels from each other, as well as comparing graphical and physical WL3. We ran a total of 14 tests for each dataset.

4.3 NASA-TLX Results

Using the NASA-TLX, we computed the results of each subject's overall workload for each condition and averaged them together, displayed in Figure 4-5. Overall, we observe that an increased number of colors lead to a higher workload level. This supports the underlying premise of our study that workload increases as colors on the rotating cube increase. A one-way analysis of variance indicates statistical significance on the *Task* factor ($p=0.0018$). Post-hoc Tukey HSD tests, designed to determine which groups differ from each other, revealed that only the TLX results between WL2 and WL4 are statistically different ($p<0.05$). This indicates that only those two states are perceived to be different.

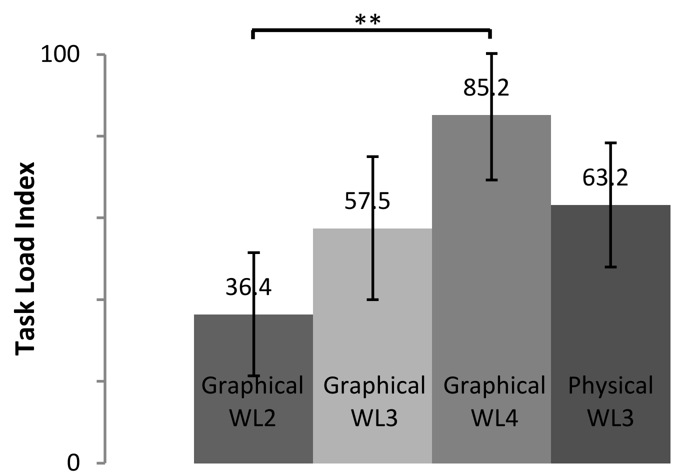

Figure 4-5. Total Workload calculated with NASA-TLX.

4.4 Classification Results

Table 4-2 displays the accuracy obtained when averaging over all subjects for different condition combinations. Appendix B details the classification results per subject.

Table 4-2. Average accuracy and standard deviation over all subjects, with multilayer perceptron.

Conditions Combinations	***Average Accuracy (stdev)***	**Chance level**
WL0 - WL3 physical	76.6% (21.6%)	50.0%
WL3 - WL3 physical	75.0% (18.6%)	50.0%
WL0 - WL2	56.1% (16.4%)	50.0%
WL0 - WL3	61.7% (16.5%)	50.0%
WL0 - WL4	71.2% (13.0%)	50.0%
WL2 - WL3	55.9% (8.0%)	50.0%
WL2 - WL4	63.4% (8.9%)	50.0%
WL3 - WL4	56.4% (12.3%)	50.0%
WL0 - WL2 - WL3	40.6% (7.9%)	33.3%
WL0 - WL2 - WL4	59.0% (12.5%)	33.3%
WL0 - WL3 - WL4	48.6% (9.3%)	33.3%
WL2 - WL3 - WL4	40.1% (7.6%)	33.3%
WL0 - WL2 - WL3 - WL4	34.8% (8.8%)	25.0%
All five conditions	34.4% (10.5%)	20.0%

4.4.1 Comparing Four and Five Conditions

When classifying between five workload levels, a random classifier would ‘guess’ with an accuracy of 20% across the five classes. Accuracy for the multilayer perceptron with all five workload conditions averages at 34.4%, ranging from 20.4% to 49.8% across subjects with all five workload conditions. The lowest classification accuracy was attained by a subject that produced many motion artifacts during the experiment, especially in the WL3 physical condition (Subject 2).

Similar accuracy results are obtained when comparing the four graphical conditions. An average accuracy of 34.8% yield similar conclusions (compared to a chance level of 25%). Individual results range from 22.5% to 45.8%. In this case, the subject with the lowest accuracy was the only left handed participant (Subject 5). It has been hypothesized that left and right handed participants have a different brain organization, which might be reflected in the data results (Toga & Thompson, 2003).

Overall, it is apparent that we can distinguish between four or five classes with accuracies better than random. However, results suggest that the granularity between a large number of workload classes was not good enough to differentiate each class in the presence of the other classes with high accuracy. Therefore, our further analysis focuses on subsets of workload conditions.

4.4.2 Analysis of Graphical Blocks

In this section, we make comparisons between workload levels viewed in the graphical interface. All combinations yield better results than average, but some perform better than others. We will analyze them in two subgroups by comparing them two by two, or three by three.

When observing the results from the classification of two classes at a time (Table 4-3), we observe an average accuracy of 60.8%. This accuracy is low (compared to a chance level of 50%), but we see potential in it.

Table 4-3. Accuracy from the comparisons of 2 workload levels

	WL2	WL3	WL4
WL0	56.1%	61.7%	71.2%
WL2		55.9%	63.4%
WL3			56.4%

Specifically, results from the comparison of two contiguous workload levels are the lowest of the group (approximately 56% for the comparisons of WL0 versus WL2; WL2 versus WL3; and WL3 versus WL4) while we obtain the largest accuracy when comparing WL0 and WL4.

Results from the 3 condition comparisons yield lower values, but the difference with chance level is approximately the same (11.3%). The results containing workload level three (WL3) all yield lower results, which indicates that this level might not be independent from the others (WL2 or WL4), so it is harder to classify. Given this observation, we are interested in looking in more details at two comparisons that do not include WL3.

Case study: comparing no, low and high workloads

Consider the results of workload level 0, 2, and 4, as displayed in Figure 4-6. Classification accuracies range from 41.15% to 69.7% depending on the subject. Given that a random classifier would have 33.3% accuracy, the results are promising. We observe a correlation between performance and accuracy results in subject five, which had the lowest classification accuracy: this subject also had incorrect responses to the number of each color seen for every WL4 task. Therefore, it is possible that the subject 'gave up' or became distracted part way through the WL4 tasks, which could result in skewed WL4 activations. However, we observed this subject do a high number of motion artifact, which is likely to be the cause of the results. Despite the lower classification accuracies for subject 5, it seems that we can predict, with some confidence, whether the subject was experiencing no workload (WL0), low workload (WL2), or high workload (WL4).

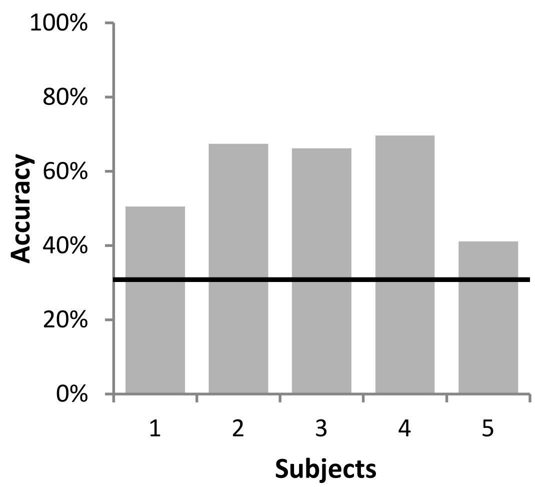

Figure 4-6. Accuracy with WL0, WL2, and WL4 considered.

The horizontal line represents chance level at 33%.

Case study: comparing low and high workloads

We observe a slight increase in accuracy when comparing low (WL2) and high (WL4) workload levels only by removing WL0 from the training and testing data although the chance level is now at 50%. In this case, average classification accuracies were 69%, 69%, 60%, 70% and 49% for subjects 1 to 5, respectively (Figure 4-7). Again, the fifth subject's results are much lower than the other subjects' results for the same reasons expressed before. While average classification accuracies were higher when we considered only WL2 and WL4, the ability to classify three classes of workload as opposed to two classes may be worth a slight decrease in accuracy.

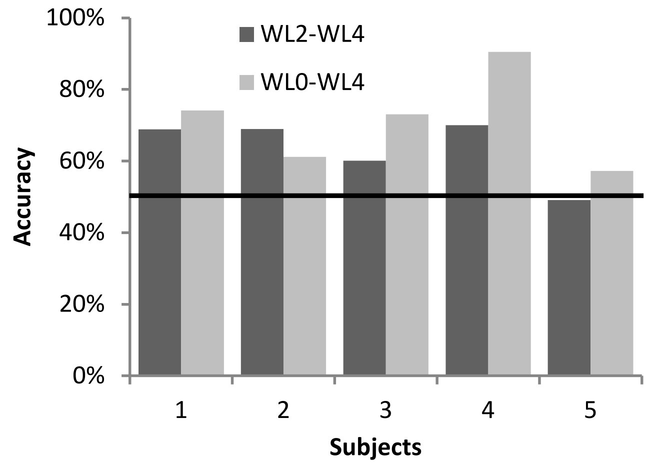

Figure 4-7. Accuracy with WL4 compared to WL2 or WL0.

The horizontal line represents chance level at 50%.

We see a similar situation when we remove WL2 from our previous case study data and only focus on differentiating between WL0 and WL4. In this case, classification accuracies range from 57% through 90% accuracy depending on the subject. Subject 5 had the lowest accuracies in all situations. This could be attributed to the subjects' response to WL4 tasks. These results indicate our ability to differentiate the presence of brain activity in the data.

4.4.3 Analysis of Graphical versus Physical Blocks

We now observe the differences between the graphical and physical user interfaces for the third workload level. The average accuracy was 75%, with a range from 44.6% to

90.6%, and accuracy greater than 73% for all but one subject (Figure 4-8). The subject with the lowest accuracy was left handed. The results show differences between the two types of displays, which indicate cognitive differences that may be due to the activation being located in different areas of the brain.

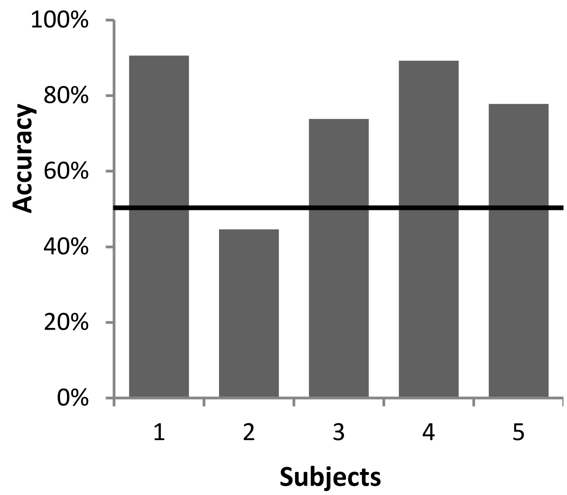

Figure 4-8. Accuracy with WL3 Graphical and WL3 Physical.

The horizontal line represents chance level at 50%.

4.5 Discussion

With the exception of the subject with motion artifacts, we observed positive classification results, which are useful from a HCI perspective. However, our current results show that we have moderate success at differentiating a large number of mental workload states. This can be attributed to both the algorithm chosen for analysis and the task granularity of the experimental protocol. Higher results were obtained by comparing noncontiguous levels of workload, mainly by eliminating the third condition

(WL3). This condition is likely to be too similar to workload levels two and four. This is corroborated by the NASA-TLX results obtained.

We also found distinguishable differences between the same workload levels when the cube was displayed in a graphical vs. physical user interface. Although we can accurately distinguish between the cognitive activities experienced in these two conditions, it is hard to identify the source of the difference, whether attributable to the workload of the interface, the workload of the task, or other variables affecting brain activity. Further studies would be necessary to establish that. However, these results encourage further exploration into cognitive workload associated with different interaction styles.

Examining our results across different subjects showed considerable individual differences. Our low participant number is partly to blame and we believe a more stable accuracy could be extracted from a larger participant pool. Given the results obtained with the left-handed subject with the physical condition, we also hypothesize cognitive difference due to handedness. We also observed that the subject that produced a large number of motion artifacts had consistently low accuracy.

Overall, we achieved our goal to test the ability of the fNIRS device to detect levels of workload in HCI, to develop classification techniques to interpret its data, and to demonstrate the use of fNIRS in HCI. Our experiment showed several workload comparisons with promising levels of classification accuracy. One of our long term goals is to use this technology as a real time input to a user interface in a realistic setting, which will be addressed in Chapter 6.

Chapter 5:

Distinguishing Difficulty Levels[3]

Maintaining the player's involvement is a key component of successful games. It can be achieved by adapting the game's content or difficulty in order to keep the user optimally challenged (Chanel, et al., 2008; Chen, 2007). As Chapter 4 demonstrated the feasibility of using fNIRS to evaluate the user's mental load, we are interested in evaluating fNIRS ability to do the same in a gaming context. The goal of this present study is to measure brain activity using fNIRS' during game play, and to distinguish the brain signal collected with fNIRS between different intensity levels of a computer game. The study is designed to ultimately lead to adaptive games and other interactive interfaces that respond to the user's brain activity in real time.

[3] The work in this chapter was originally described in Girouard, et al. "Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements" in the proceedings of Human-Computer Interaction - INTERACT (2009) pp. 440-452.

The present study applies fNIRS to the human forehead, measuring the anterior prefrontal cortex, a subset of the prefrontal cortex. Research shows a prefrontal cortex response to video game playing, which lead us to believe that the video game Pacman could produce similar activations. Note however that most of the fNIRS studies measure a larger brain region, with probes that are much different than ours, although our current probe format has the advantage of a simple and comfortable setup.

The arcade game of Pacman was chosen in this experiment because of its great potential for passive adaptability: it is easy to change the amount of enemies to maintain interest without overwhelming the user. This selection was based both on its customizable environment and on a literature review of game play (see Chapter 2). Pacman offers different difficulty levels that keep all other aspects identical, such as the scene and the characters' behavior. We believe the results obtained with Pacman will translate to other games of similar mental demand.

We developed and implemented a computer version of the game of Pacman, originally released by Namco (Japan). Figure 5-1 displays a snapshot of our version of Pacman. The user directs Pacman through a maze by pressing arrow keys, with the goal of eating as many fruits and enemies as possible, without being killed.

As Chapter 4 concluded only moderate success at differentiating a large number of mental workload states, which suggest that a lower number of states might yield better results. We choose to test two activation levels—two game difficulty levels—in this experiment to simulate and improve the results obtained by comparing workload level two and four in the previous experiment.

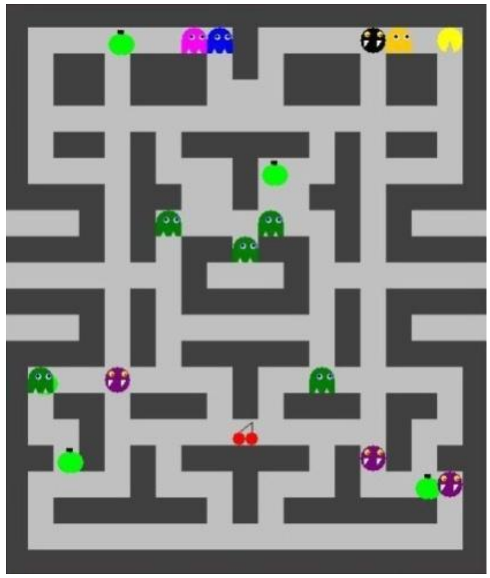

Figure 5-1. A snapshot of Pacman (the yellow character on the top right corner), enemies and fruits on the maze, as used in the experiment (hard level).

Two levels of difficulty, differentiated by pace and quantity of enemies, were selected through pilot testing. The enemies walk at a pace of one step per 1000ms for the easy level, and one step per 150ms for the hard level. There can be a maximum of 6 enemies at once on the board in the easy level, and 12 for the hard one. The maximum number of fruits on the board is identical for both levels of difficulty (7 fruits), with at most one cherry at any time. Each game started with a new, clean board. A new board contains four enemies and three fruits, dispersed on the board. The Pacman starts in one of the four corner positions, randomly selected.

Participants were hypothesized to be able to distinguish these difficulty levels, so it was also hypothesized that brain measurements would show distinguishable differences in addition to observed differences in performance.

Nine subjects (4 females) participated in this study (mean age of 24.2 years; std 4.15). All were right-handed, with normal or corrected vision and no history of major head injury. Informed consent was obtained, and participants were compensated for their time. All knew of the game, and all but one had previously played it. Participants practiced the game for about one minute to familiarize themselves with our version.

5.1 Design and Procedure

Participants completed ten sets of two trials (one in each difficulty level) over a twenty minute period. In each trial, participants played the game for a period of thirty seconds, and rested for thirty seconds to allow their brain to return to baseline. Conditions within each set were randomized for each subject. The experimental protocol of alternating 30s-long windows of activation and rest was designed to take into account the slow hemodynamic changes that occur in a time span of 6-8 sec (Bunce, et al., 2006) as well as a short game cycle that nonetheless allowed performance to level off. Figure 5-2 illustrates the experimental protocol.

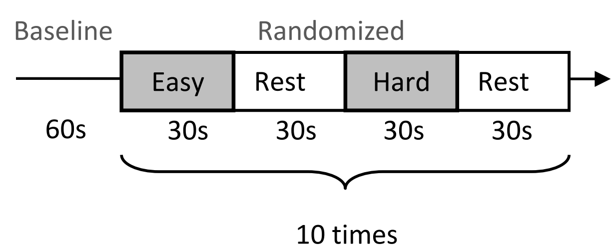

Figure 5-2. Experimental protocol: a minute of baseline, followed by 10 random sets of 30 seconds of playing time, then 30 seconds of resting time for each condition.

In addition to fNIRS data, we collected performance data—number of times Pacman is killed, as well as number of fruits and enemies eaten. At the end of the experiment, subjects were asked to rate the overall mental workload of each game level with the NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988), a widely used measure of subjective mental workload used here as a manipulation check. The NASA-TLX for each level was administered using a paper version (two in total).

5.1.1 fNIRS Equipment

We chose to use the data from the two last sources of each probe (with source-detector distances of 2.5 and 3cm), because they reach deeper into the cortex. The shallower source-detector axes are thought to pick up primarily systemic responses happening in or on the skin. Selecting deeper measures is hypothesized to improve our results.

5.2 Analysis Techniques and Results

5.2.1 Behavioral Results and Performance Data

We performed an analysis on the non-brain data collected, that is the NASA-TLX results and the game performance statistics. The NASA-TLX data was meant to confirm that users perceived the two difficulty levels as different. Results indicated an average mental workload index of 26 (std 12.9) for the easy level, and 69 (std 7.9) for the hard level, on a 100 point scale. This difference was significant according to a two sided t-test ($p<0.01$), and confirm our manipulation.

We also examined the performance data. Every data source collected showed a significant difference between the two difficulty levels ($p<0.05$). Figure 5-3 displays the average value of the data collected.

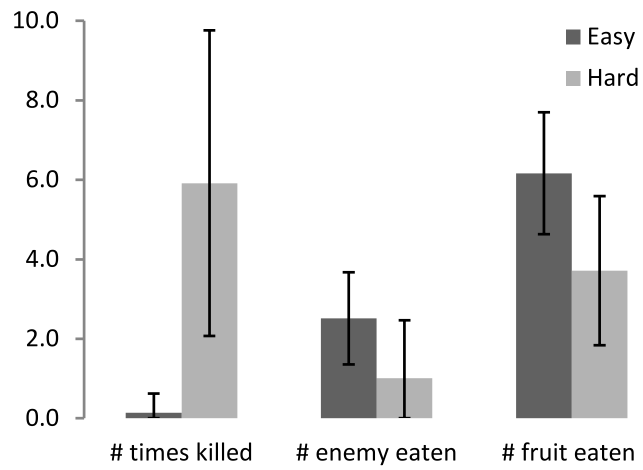

Figure 5-3. The difference between each level is significant for each data type. The graph shows data collected, with standard deviation, averaged over trials and subjects.

5.2.2 Brain Data Analyses

We performed two analyses of the brain data to confirm the presence of differences in hemoglobin concentrations for the different conditions: a classic statistical analysis to establish the differences between conditions, and a more novel task classification that will show the possibility of using this data in a real-time adaptive system.

5.2.3 Brain Data Preprocessing

Given the assumption that the brain returns to a baseline state during each rest period following the stimuli, even though it may not be the same baseline state in each rest period, we shift each trial so that the initial value is zero to control for differences in initial state. Finally, we separate each trial according to *Activeness*—whether the user was playing or resting. Figure 5-4 illustrates trials of data for a particular stimulus.

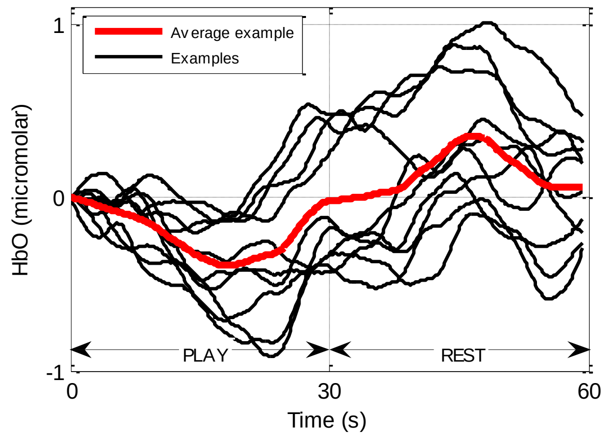

Figure 5-4. Example of fNIRS data, zeroed.

The red, ticker line indicates the mean of all trials. The left half of the data was taken when the user was playing the easy Pacman, and the right half was the rest period following.

5.2.4 Statistical Analysis of Brain Data

For the statistical analysis, we average each trial of each condition to get a mean value of oxygenated hemoglobin [HbO] and deoxygenated hemoglobin [Hb], for each difficulty level, activeness, hemisphere and channel. We then apply a factorial repeated measures analysis of variance (ANOVA) on *Difficulty level* (2) x *Activeness* (2) x *Hemoglobin Type* (2) x *Hemisphere* (2) x *Channel* (2). This factorial ANOVA will observe differences within each participant, and determine if they are significant across participants. This is the same analysis as performed in Chapter 3, apart from the two leading factor, specific to this study. The full statistical results can be found in Appendix A-2.

If the end result is to construct a system that can respond to different individuals with a minimum of training, we need to know how different we should expect individuals to be—hence including subjects as a factor in the analysis. Given the novelty of the fNIRS method, and the lack of well-established analysis methods in previous work in this area, the cortical distribution of the combination of channel and hemoglobin type effects cannot yet be predicted beforehand. In addition to the statistical significance, we report the effect size of the interaction (ω^2), which is the magnitude of the observed interaction, and indicates practical significance. An omega-squared measure of 0.1 indicates a small effect, 0.3 a medium effect and 0.5 a large effect (Field & Hole, 2003).

We found the main effect *Hemoglobin Type* to be significant, with a medium effect ($F(1, 8)=6.819$, $p<0.05$, $\omega^2=0.39$). This was expected, because [Hb] and [HbO] are present in different concentrations in the blood. The interaction of *Channel* x *Hemoglobin Type* is also significant, with a medium effect ($F(1, 8)=5.468$, $p<0.05$, $\omega^2= 0.33$), indicating that [Hb] and [HbO] are not the same at a given channel.

Game-playing compared to resting are significantly different as an interaction with channel with a large effect size (*Activeness* x *Channel*, $F(1, 8)=27.767$, $p<0.001$, $\omega^2= 0.75$), showing that there is a difference between playing Pacman and resting, and that this difference varies as a function of the cortical depth of the measurement (that is, the source-detector distance, or channel). We also observed that the interaction of *Activeness* x *Channel* x *Hemoglobin Type* is significant, with a medium effect ($F(1, 8)=5.412$, $p<0.05$, $\omega^2= 0.32$), as illustrated in Figure 5-5.

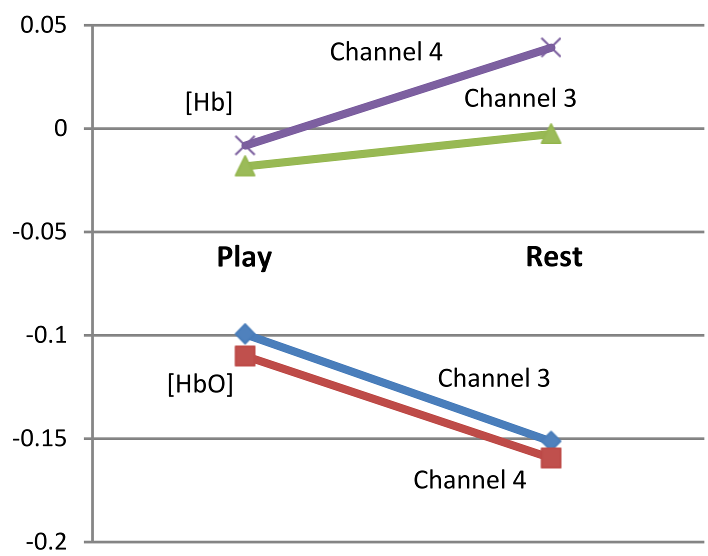

Figure 5-5. Mean plot of the interaction of *Activeness* x *Channel* x *Hemoglobin Type.*

Finally, we observed a significant interaction of *Difficulty Level* x *Activeness* x *Channel* x *Hemoglobin Type*, with a small effect size ($F(1, 8)= 7.645$, $p<0.05$, $\omega^2= 0.18$). This interaction shows that we can significantly distinguish between the activeness of the participant, and the degree of difficulty of the current game when data from all channels and hemoglobin type are used as features.

This confirms our initial hypothesis. The ANOVA results indicate significance between the play and rest conditions, and the two difficulty levels.

5.2.5 Machine Learning Classification of Brain Data

Statistical analysis confirmed our hypothesis that the brain signals in the different conditions were significantly different. We then wanted to determine whether this signal could be used in an adaptive user interface. To do this, we used machine learning to train a classifier.

We chose to explore a second type of classification technique, called sequence classification (Dietterich, 2002). While sliding windows demonstrated some potential in Chapter 4, careful observation of the fNIRS data revealed that the curves are exactly that, curves, not plateaus, as illustrated in Figure 5-4. Hence a technique that relies on the idea that small slices of the same condition will look alike is not as appropriate. As opposite of the sliding window, sequence classification considers the entire task as an example, instead of slicing it into a large number of examples. Specifically, sequence classification applies a label to an entire sequence of data, and uses each data point as a feature (Figure 5-6). In our case, a sequence is one trial, containing 180 points.

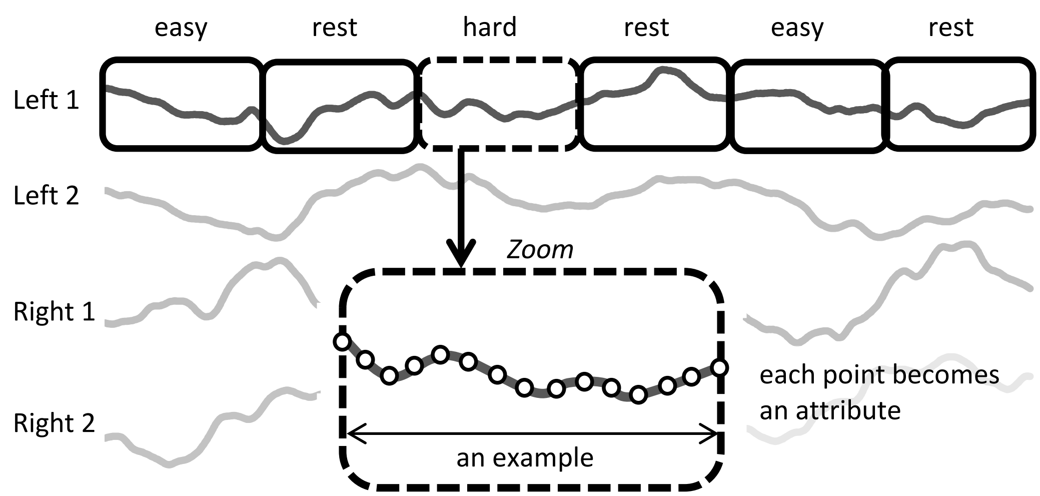

Figure 5-6. Schematic diagram of sequence classification.

Each curve (one collected brain measure) is sliced into task-sized chunks, with each time point as a classification feature.

Because of our multivariate data (8 recordings for each time point: 2 probes x 2 channels x 2 hemoglobin types), we classify each channel individually first. To combine the results of all these classifications, each classifier votes for the label of the example. We used a weighted voting technique that sums the probability distribution of each example by each classifier.

The classification algorithm used is k-nearest-neighbors (kNN), with k=3. kNN uses the label of the three most similar examples (the closest neighbors) to the example to classify, and assigns a label based on the weighted average of their labels. We used a random 10-fold cross-validation in all classifications. We trained the classifier on part of one subject's data, and then tested for this specific subject with the left out data. This procedure was repeated for each subject. The cross validation resulted in test sets of 2 or 4 examples of each class. This cross validation is similar to that of Chapter 4.

We used the same preprocessing as for the statistical analysis, but we explored the difference when zeroing the data. In the statistical analysis, we “zeroed” the data, meaning that we shifted the trial so that the first datapoint was zero, under the assumption that the brain had returned to baseline. In this analysis, we tested both zeroed data, and non-zeroed data (see Figure 5-7 for a visual example). We were however more interested in observing the results for non-zeroed data, because this data is more similar to the one we would have access to in a real time brain-computer interface.

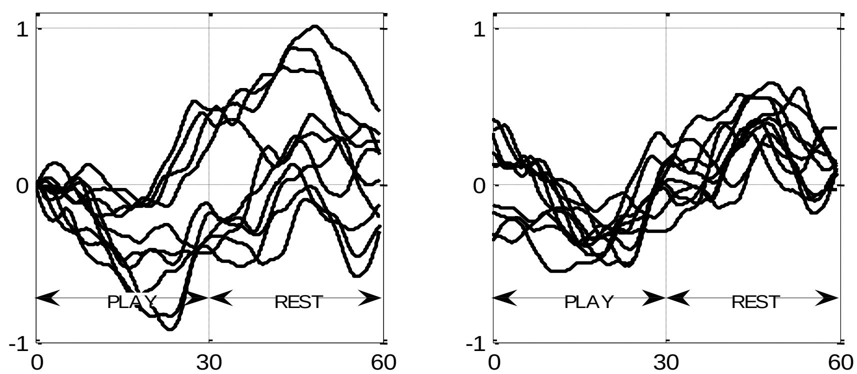

Figure 5-7. Example of zeroed (left graph) and non-zeroed data (right graph). Left graph identical to Figure 5-4.

We attempted three types of classification: (a) *Activeness* (play versus rest), (b) *Difficulty level* (easy versus hard), and (c) *Two difficulty levels and rest* (easy versus hard versus rest). To accomplish each classification, we selected and/or grouped the trials differently. For *Activeness*, we combined all playing trials into one class, and all resting trials into another to form two classes (20 examples of each class). For *Difficulty Level*, we compared the easy and hard levels using the play trials only (10 examples of each class). Finally, in *Two difficulty levels and rest*, we compared three conditions: the play period of the easy level, the play period of the hard level, and all rest periods.

Our initial implementation used individual classification (only trials of one subject classified together). We call this "Per subject classification". Most BCI work is done this way, per subject, where we train and test on one subject's brain activity. Figure 5-8 shows the average accuracy of each type of classification, for non-zeroed data, with classification done per subject (accuracy averaged over subjects). We began with non-zeroed data as it represents the most likely parameter for a real time system.

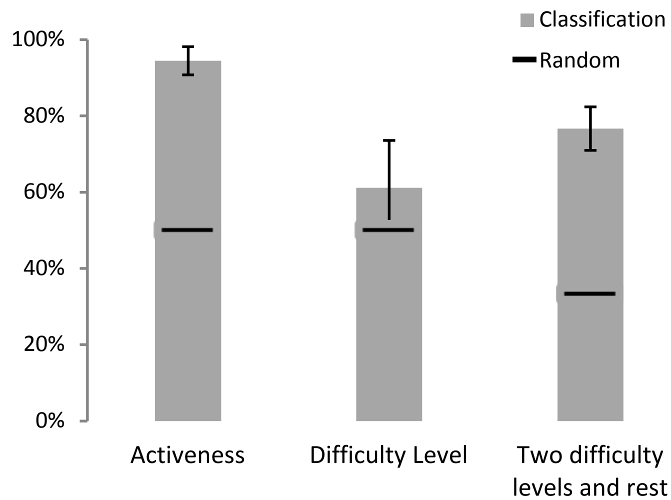

Figure 5-8. Average accuracy for different classifications for non-zeroed data, per subject classification, with standard variation and random classification accuracy.

We also explored the possibility of bypassing this step by pooling all the data together, which is to classify all subjects together. We label this method “Combined subjects classification”. Table 5-1 illustrates the three possible analyses produced: the table shows the average accuracy of each type of classification, for zeroed and non-zeroed data, per subject classification, and combined subjects classification.

Table 5-1. Average accuracy for different classification variations.

The gray cell indicates the highest result of the table. The standard deviation, when available, is indicated in parentheses.

Activeness	**Non-zeroed**	**Zeroed**
Per subject	94.4% (3.7%)	93.1% (6.1%)
Combined	83.3%	91.4%

Difficulty levels	**Non-zeroed**	**Zeroed**
Per subject	61.1% (12.4%)	55.6% (16.3%)
Combined	55.6%	54.4%

Two difficulty levels and rest	**Non-zeroed**	**Zeroed**
Per subject	76.7% (5.7%)	75.6% (8.4%)
Combined	67.8%	71.4%

There are two elements to observe in the tables of results: (1) a comparison between the results from averaging the results of the classification of each subject individually (per subject) or from running the data from all subjects together (combined subjects); and (2) an evaluation of using the non-zeroed data versus the zeroed data. The highest result (grayed out) in all tables happens to be the data per subject and non-zeroed, identical to Figure 5-8.

5.3 Discussion

While some might argue that performance data is sufficient to classify the difficulty level of a game and can be obtained without interference, the goal of this study is to investigate the use of the brain measurements with fNIRS as a new input device. In a more complex problem, performance and brain data coming from fNIRS might not be as related, e.g. if the user is working hard yet performing poorly at some point. In addition, distractions may also produce workload increases that would not obvious from monitoring game settings and performance, and thus may necessitate brain measurements. That is, a participant playing a simple game while answering difficult questions might also show brain activity relating to increased workload that would be incomprehensible based only on performance data (e.g. Nijholt, et al., 2008). In real, non-gaming situations, we might not have performance data like in the present case, as we don't always know what to measure— how hard is an air traffic controller working, or a person creating a budget on a spreadsheet? The use of the brain signal as an auxiliary input could provide better results in these situations.

Our analyses show that we can distinguish between subjects being active and passive in their mental state (*Activeness*), as well as between different levels of game complexity (*Difficulty Level*). The classic statistical analysis confirmed that these conditions produced different patterns in blood oxygenation level, and the machine-learning analysis confirms that these patterns can be distinguished by the classifiers used.

5.3.1 Brain Activation When Playing Pacman: Play versus Rest

Results indicate the presence of a distinct brain signal when playing Pacman, in comparison to the rest periods. The *Activeness* classification in Figure 5-8 yields an average accuracy of 94.4% (for non-zeroed data, classified per subject). It indicates a noticeable difference between the playing signal, and the resting signal. This corresponds to the results obtained with the statistical analysis, where *Activeness* was a significant factor in multiple interactions. This provides real time measurements that could be used in an adaptive interface. Our results corroborate those of previous studies that showed prefrontal cortex activity related to video games, measured with fNIRS.

5.3.2 Difficulty Levels: Easy versus Hard

The *Difficulty level* of the game was shown to be a significant factor in this experiment in both types of analyses. This is supported with the fact that users perceived the two levels as being significantly different according to the NASA-TLX. Hence, we can say that there was a significant cognitive difference between the two levels. Previous fNIRS game experiments (Matsuda & Hiraki, 2005, 2006; Nagamitsu, et al., 2006) only analyzed stimuli versus non-stimuli periods (which in this experiment we have called activeness), and not two levels of difficulty, making this result an advance over prior work.

However, the statistically significant interaction that included *Difficulty Level* had a small effect size, and classifying the difficulty of playing periods yields an average accuracy of 61.1% (for non-zeroed data, classified per subject). This relatively low accuracy indicates that it is difficult with this classifier to differentiate between the two levels, which relate to the small effect size found in the statistical analysis. We also observed significant inter-subject variability through a high standard deviation: only four participants scored between 65% and 85%. This indicates that the two difficulty levels might be significantly different with only part of the participants. As everyone's brain varies greatly, this is not a surprising result. Implications of this result for human computer interaction include the fact that a brain computer interface using such measures could only be accessible to a subset of the population. However, a study with a larger number of participants is necessary before making a clear statement to that effect.

A comparison of three types of conditions (*Two difficulty levels and rest*) indicates an encouraging average accuracy of 76.7% (for non-zeroed data, classified per subject), explained by the low differentiation between the difficulty levels, and the high separation between the activeness of the subjects. We must note that the difference in brain signal measure is not strong. One explanation may be that the difference in mental processes between each level manifests itself in other brain locations besides the anterior prefrontal cortex (location measured), such as in the dorsolateral prefrontal cortex. It could also be that the difference between the two difficulty levels was not big enough to cause strong changes in activation.

Results are consistent with prior work. Distinguishing work from rest was relatively easy, but discriminating different workload levels was harder, with significant inter-subject variability. Similar results have been found over decades of EEG work (e.g. Allison & Polich, 2008; Gevins & Smith, 2003), which may suggest fundamental limitations in making fine discriminations between two similar workload levels. Physiological signals produce similar results (Chanel, et al., 2008).

Current findings indicate the presence of brain activation in the anterior prefrontal cortex when playing Pacman. Because the activation of the different levels of difficulty is correlated with mental workload (measured with NASA-TLX), we can presume that the difficulty level in this experiment is also correlated with mental workload.

5.3.3 Exploring Different Classification Methods

During our machine learning classification, we explored two data processing and variation of in the analysis.

Per subject or combined subjects classification

At first, we examine the difference between individual and grouped datasets classification. If we observe the non-zeroed data, we observe an increase of 11.1% in accuracy of the Activeness classification by averaging the data individually. This trend can be observed for each type of classification: we note that the individual runs are on average 8.3% higher than the runs of all subjects when using non-zeroed data. However, the average increase when using zeroed data is only of 1.4%. We also observe a high standard deviation of the averaged individual runs with zeroed data, indicating that many per subject accuracy are below that of the classification of combined subjects.

Overall, this tells us that both types of accuracies are within similar range. We may get higher accuracies when classifying subjects individually when using non-zeroed data, and higher results when running all subjects combined when the data is zeroed. Those results are encouraging because it means we can use data from multiple subjects to train a classifier. However, because the cross-validation was run with random samples of the data (which are unlikely to be entirely from one subject), this does not indicate we can use a new subject without any training. Additional analyses training on the data of all but one subject, and testing on that left-out subject would be interesting.

Using zeroed or non-zeroed data

From the point of view of data zeroing, we observe that with per subject classification, the zeroing of data produces reduced accuracy (on average, by 3.5%) and increased individual variation (higher standard deviation), while it increases the accuracy when classifying all subjects at once (by 3.5% on average). For the individual results, we can attribute the decrease to the fact that the first few points of the data are very similar—the first point of every example is zero, and the following ones are very closely related (see Figure 5-7). Hence we are using a reduced amount of features to classify them. For the results of classifying combined subjects, we find that the zeroing performs a “normalization” of the data between subjects (by shifting data), which leads to better comparisons and classification. Other types of normalization could be possible, such as scaling the data, but they were not investigated in this analysis. More normalization could lead to better results, especially when classifying on multiple subjects at once.

Overall, we believe the machine learning results are noteworthy. They show the ability of fNIRS data to be classified easily and the potential they can have to be used in an adaptive interface. In the long run, our goal is to be able to classify data in real time. The data collected in this experiment suggest to run per subject classification when using non-zeroed data, and to use the classification of combined subjects when using zeroed data.

5.4 Conclusion

In this chapter, we have shown that functional near-infrared spectroscopy can distinguish between the brain at rest and the brain activated when playing a video game, both using statistical analysis and machine learning classification. We also demonstrated that we can differentiate two levels of difficulty with some success. The activation of the different levels of difficulty is correlated with mental workload, measured with NASA-TLX. Hence, we can presume that the difficulty level in this experiment is correlated with mental workload. However, our classification accuracy was low when distinguishing easy or hard levels.

Saito et al observed a larger activation cluster in the dorsolateral prefrontal cortex with the games of Othello and Tetris than with Space Invaders (Saito, et al., 2007). This was justified with the fact that Othello and Tetris require spatial logical thinking (planning and memory of prior moves). The game of Pacman relates more to Space Invaders than to Othello or Tetris, as both are arcade games, and not puzzles, suggesting the possibility of a stronger signal with a different game. In addition, previous work using fNIRS to study video games compare different types of games (e.g. shooter game versus puzzle game), which could be interesting to experiment with, such as contrasting different levels in other types of games. This could verify whether differentiating two levels of video games yield weak results in other game types, or that Pacman's main brain activation is located elsewhere. Finally, while their results were weak, Hattahara et al. (2008) implied that expertise is an important factor in the prefrontal cortex activity when playing games. It would be interesting extension to include a larger number of subjects with varying levels of experience with the game and compare their results.

In a larger research context, exploring the use of fNIRS in an adaptive interface would prove interesting for the HCI community. Results of the comparison of two different levels could be applied to other games of similar mental demand. The correlation between mental workload and difficulty levels in this experiment indicates we could also apply the current results to general applications that respond to such measurements.

Chapter 6:

Designing a Passive BCI using fNIRS Real Time Classification

While most work using fNIRS uses offline analyses to evaluate the data collected, the key component of brain computer interfaces is the ability to perform real time analyses. Many researchers argue that their work could be converted and done in real time (e.g. Sitaram, et al., 2007) including our previous chapters, yet we found few fNIRS systems in the literature that do (Coyle, et al., 2007; Luu & Chau, 2009). The existing systems use simple paradigms, making decisions with a threshold or by comparing the signal from two previous tasks. We also found many BCI systems that operate in real time, processing EEG data streams, and controlling interfaces. We learn from these tools and apply their principles to the design of an fNIRS real time system.

We have developed a software system that allows for real time fNIRS brain signal analysis and machine learning classification of affective and workload states called the Online fNIRS Analysis and Classification system (OFAC). This system receives and processes brain signals and event markers, automatically recognizes the current cognitive state using a database of previously recoded signals and machine learning techniques, and outputs this state to the interface, allowing for the creation of interfaces that adapt and change in real time according to traditional inputs as well as cognitive activity. OFAC offers the user an additional communication channel based on brain activity, providing multimodal interaction.

Our work aims at reproducing the procedures used offline in previous work, adapting them to be suitable for real time input to a user interface. This chapter presents the OFAC system, tests and proves the system's reliability and potential through two studies. Our first evaluation compares a previous offline analysis with our real time analysis. The second study demonstrates the online features of OFAC: its ability to record, process, classify and adapt simple interfaces in real time.

6.1 Online fNIRS Analysis and Classification System

We present a new system that uses machine learning to classify a large number of states in real time to obtain the user's cognitive state. OFAC works with fNIRS data in a real time pipeline to feed a user interface, which can in turn adapt to the information. In this research, we achieved the transformation of the offline characteristics of the ISS

Oxymeter system (Champlain, IL) into a real time system. While their system was never designed to run in online, we overcame technical issues to obtain and interpret the raw data in real time.

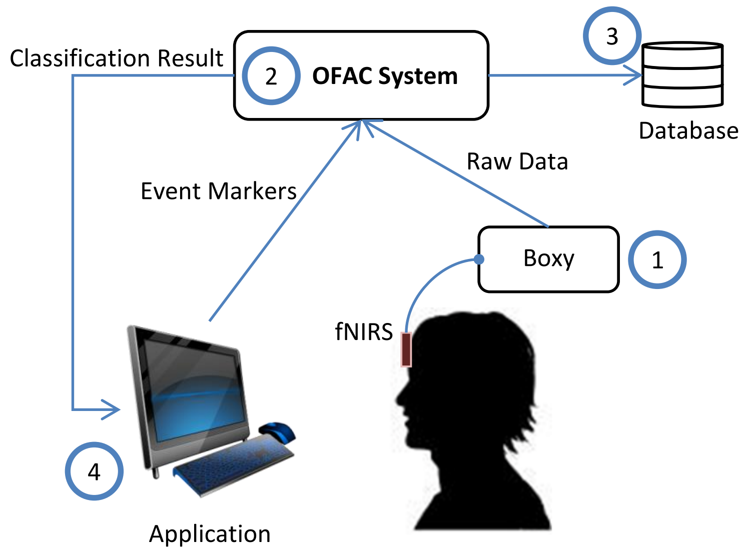

Figure 6-1. OFAC system's architecture

We created a flexible, modular architecture for the OFAC system using Matlab (The MathWorks, Natick, MA) (Figure 6-1). It allows for the substitution of single modules should another functionality be required, and accepts multiple input signals, such as the combination of fNIRS and EEG. The rest of the chapter will describe an fNIRS-only system used in the latter experiments.

OFAC contains three types of modules for data processing: modules to receive and record input data into a database (one for each type of input); to pre-process and to filter data; and to perform machine learning classification and output the brain signal classification to the interface. The current system takes two different types of input: raw brain data, and external markers from the application shown to the user. The raw data (from the fNIRS acquisition software Boxy, ISS Inc.) can include basic markers related to the start and stop of the real fNIRS data when the sensors are correctly in place and the experiment starts, as opposed to uncalibrated data. The external markers could contain behavioral data, for instance, to help with data classification.

OFAC System Loop

- fNIRS raw data acquisition
- Application data acquisition
- Data storage
- Signal processing
  - Transform to hemoglobin values
  - Filter
  - Real time visualization
- Feature generation and classification

Figure 6-2. OFAC high-level loop.

OFAC is a complex distributed system that can process each module on an individual computer. The system currently runs on two computers, one for the application with which the user interacts, and one for the fNIRS software and the real time processing program OFAC, illustrated in Figure 6-3. With this setup, the experimenter can monitor the user with the real time program, without interrupting the participant.

Should the processing program take a lot of CPU power and interfere with the fNIRS measurement software, every program can run on a different computer (Figure 6-4). We are required to have a serial (real or virtual) connection between Boxy and Matlab (a Boxy constraint), but there is no restriction on the type of communication protocol for the link between OFAC and the application.

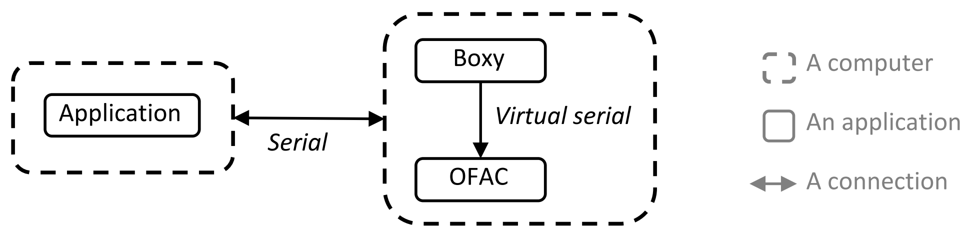

Figure 6-3. The real time system runs on two computers, communicating through a serial connection.

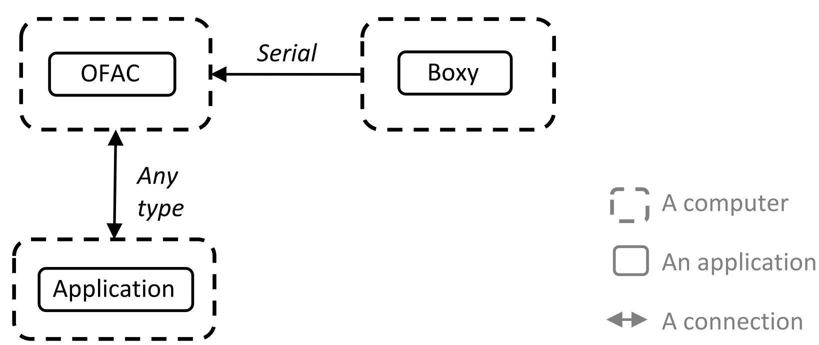

Figure 6-4. The real time system computer organization with one computer per program.

This architecture imposes minimal requirements for the application software, which can be written in any language on any platform. The only constraint is to have the ability to connect with OFAC and respect a defined communication protocol, currently done through a serial connection. The current protocol exchanges semi-colon separated data. The application sends event markers with the form:

```
trial number; task name (or code); application timestamp
```

In turn, OFAC sends classification results with the form:

```
predicted class; {class probability distribution}
```

OFAC provides both an *online* and *offline* mode. The latter provides a tool to explore previously recorded data with the OFAC system for research purposes, such as evaluating the impact of different filtering methods, or classification algorithms.

The following descriptions explain both the general components of OFAC and the specific implementation used in our studies.

6.1.1 Data Acquisition and Storage

The current system received two data streams: event markers from the application and raw brain signals. It stores the data as it comes in in a database, preventing data loss should the system have a major malfunction.

The synchronization of the system and the different data sources is done with a timestamp of the fNIRS data as soon as it comes in. The application event marker, read immediately after (if any), is time stamped with the raw data time.

6.1.2 Signal Processing

Our work aims at reproducing the procedures used offline in previous work (e.g. Chapter 5), adapting them for function suitable for real time input to a user interface. We first convert the raw values in light absorption changes. We apply a moving average, removing the high frequency noise (Figure 6-5). The filtered data is converted into oxygenated and deoxygenated hemoglobin concentrations using the modified Beer-Lambert law (Chance, et al., 1998).

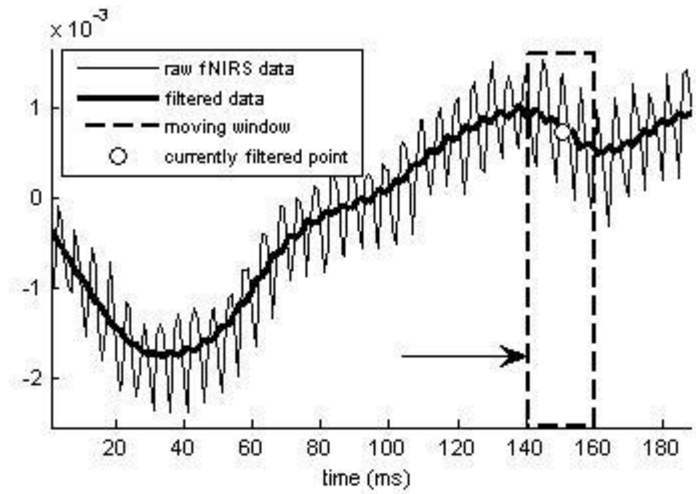

Figure 6-5. Moving Window of 19 points.

6.1.3 Feature Generation and Classification

To mimic the procedure used in the previous analysis, we implemented the machine learning technique called sequence classification (Dietterich, 2002). OFAC calls Weka (Hall, et al., 2009) to perform the training and testing of examples created using sequence classification. Using Weka gives OFAC access to a large library of classification algorithms. When enough training data has been accumulated, we first call the classification algorithm to obtain the classifier, and then this classifier is used to test the examples as they come in. To allow flexibility in the analysis, the system can test examples one at a time, or in groups of points.

6.1.4 Summary and Additional Features Implemented in OFAC

We have implemented a few functions which could be of use for future analyses, although they are not all employed in the following experiments. Table 6-1 summarizes the current data processing capabilities of OFAC.

Table 6-1. OFAC data processing capabilities.

Signal Processing	*Feature Extraction*	*Machine Learning Algorithms*
• Baseline filtering • Hemoglobin conversion • Band-pass filtering • Moving average filtering • Channel selection	• Data zeroing • Cutting a few seconds • Sequence classification • Sliding window (w=1) • Class merging • Class selection	• Access to Weka k-nearest-neighbor, support vector machines, etc. • Individual or batch testing • Baseline threshold

We have also integrated machine learning evaluation tools, such as calculating the accuracy of the subject's session, displaying the confusion matrix and providing a basic classification results visualization graph.

6.1.5 System Monitoring and Visualization

Real time monitoring of the data becomes a critical factor with online systems. We provide two ways to keep an eye on the process. First, we output status updates to the command line, giving a snapshot of the system. Those updates are mainly of five types: the general algorithm's phase (baseline, training, or testing phase, or not currently using the measurements); the measurement number, the trial number and the task in progress; event markers received and classification results sent; classification calls to Weka; and any system output or errors. Figure 6-6 displays an example of the displayed status updates.

Figure 6-6. Example of status messages while running a subject

```
[Point #0] NOT_MEASUREMENT period
Msg from reading the fNIRS buffer: "A timeout occurred before the
Terminator was reached."
[Point #0] was longer than 0.16s (took 2.0849s)
[Point #1] was longer than 0.16s (took 0.91737s)
[Point #92] fNIRS Marker: 1
[Point #97] Application Marker: 0;baseline;1000
[Point #97] BASELINE period
[Point #100] Status msg: baseline period; Current task: 0.5s elapsed
[Point #200] Status msg: baseline period; Current task: 16.5s elapsed
[Point #300] Status msg: baseline period; Current task: 32.5s elapsed
[Point #400] Status msg: baseline period; Current task: 48.5s elapsed
[Point #480] Application Marker: 1;video;62195
[Point #480] Baseline calculated
[Point #480] TRAINING period
[Point #480] was longer than 0.16s (took 0.27926s)
[Point #500] Status msg: training period; Current task: 3.2s elapsed
[Point #600] Status msg: training period; Current task: 19.2s elapsed
[Point #690] Application Marker: 1;video_rest;95864
[Point #700] Status msg: training period; Current task: 1.6s elapsed
[Point #715] Application Marker: 1;tetris;99864
...
[Point #6434] Application Marker: 14;tetris_rest;1014944
[Point #6459] Application Marker: 15;video;1018944
[Point #6459] TESTING period
[Point #6454] Classification: TRAINING data
Preparing data
Class <tetris> (0): 2631 points
Class <video> (1): 2647 points
Keeping 5278 of 5974 original data points -- Deleting classes not
used
Converting <fNIRS-train> from matlab to weka
Running weka with classifier: functions.SMO -C 1.0 -E 1.0 -L 0.0010 -
P 1.0E-12 -N 0 -V -1 -W 1
Saving training data and classifier
...
[Point #6641] Application Marker: 15;tetris;1052974
[Point #6700] Status msg: testing period; Current task: 9.4s elapsed
[Point #6800] Status msg: testing period; Current task: 25.4s elapsed
[Point #6828] Application Marker: 15;tetris_rest;1083004
Classification: TESTING
Sending to App: t;0.9;0.1
[Point #6853] Application Marker: 16;video;1087004
[Point #6900] Status msg: testing period; Current task: 7.5s elapsed
[Point #7000] Status msg: testing period; Current task: 23.5s elapsed
[Point #7041] Application Marker: 16;video_rest;1117019
Classification: TESTING
Sending to App: v;0.0;1.0
...
```

We also provide a visual display of the data (Figure 6-7). The interface has four panels: the top left plot contains the raw data (as it is received by the system); the bottom left plot contains the light absorption values and the filtered absorption values; the bottom right plot is hemoglobin values; and the top right plot shows the running times (to monitor lag time). While the interface can be updated with every new point as it comes in, this update slows down Matlab and it becomes unusable within a few minutes: Matlab is simply not designed to handle this type of plotting. Instead, we take a snapshot of the plots every few seconds and update the interface.

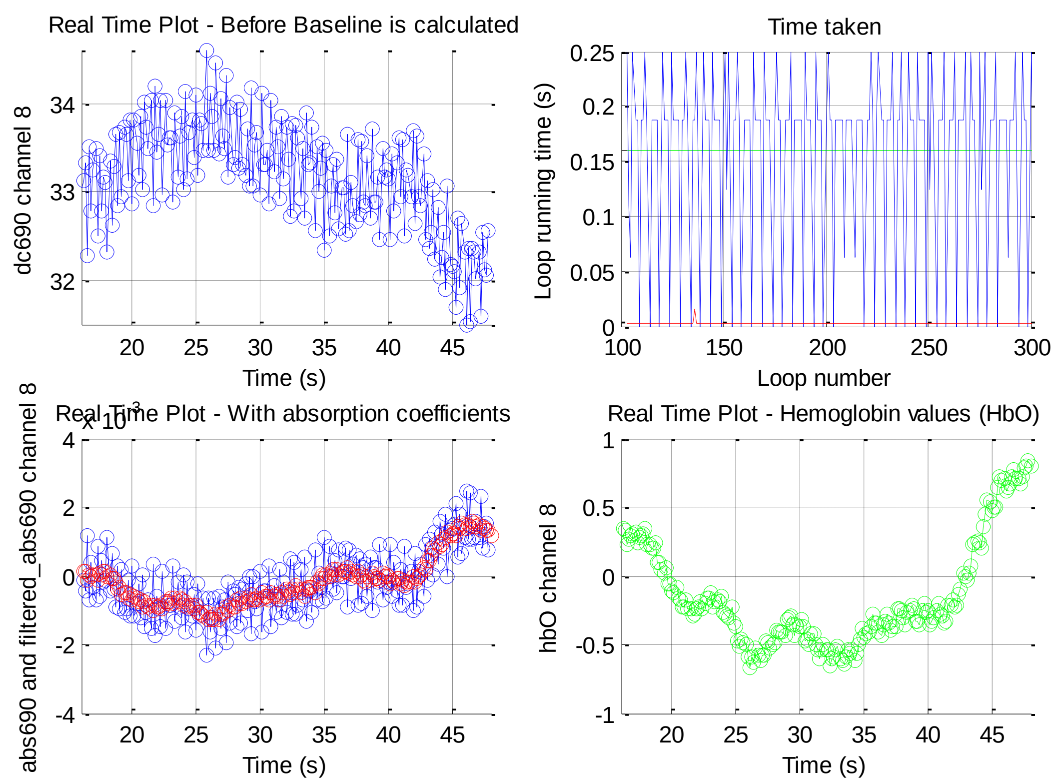

Figure 6-7. Real time visualization interface.

6.2 Generic Synchronous Experimental Protocol

We present a generic experimental protocol to use with the OFAC system (Figure 6-8). The system currently operates in a synchronous mode, meaning that the user is bound by a specific task schedule. However, since our signal is analyzed continuously, a different classification algorithm with the same architecture could allow for an asynchronous mode of operation.

In each protocol, we start by collecting baseline data (around a minute). This lets the user relax, and it allows us to measure their brain at rest. Following comes a training period, where the user repeats tasks for a certain number of times. This period is designed to collect examples of the user's brain signal for each task, in order to predict it later. The training period should include rest periods, used to allow the user's brain to go back to a rested state. This is a typical requirement for fNIRS experiments. The collected data is then used to train the chosen classifier. The remainder of the experiment uses the trained classifier to periodically test the brain signal received.

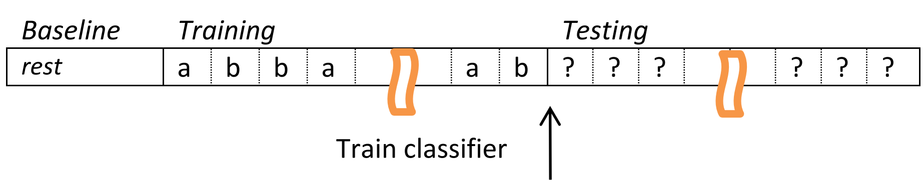

Figure 6-8. Generic experimental protocol.

6.2.1 Differences Between Online and Offline Analysis

The challenges when transforming offline analyses so they are performed in real time resides mostly with the fact that the full dataset is not equally available. In offline analyses, the whole dataset is available at once, while in an online system, we only have the amount of data collected so far, reducing the total information available. This element affects mostly two components of systems, the data filtering and the classification. Most of the differences in accuracy observed between offline and online analysis of the same data will be explained by this.

While the machine learning classification algorithm is not affected by the offline/online distinction, and we can use identical algorithms, the data available to train (and test) the algorithm does vary. For instance, offline algorithms can perform cross validation, and pick a random sample of the whole data set to be taken out for testing. In contrast, real time algorithms are bound to the order in which examples come in.

The second challenge to acknowledge is that online systems cannot afford to have time consuming procedures and algorithms. The chosen signal processing method, feature extraction and classification algorithms must require a low computational time and complexity as not to slow down the whole system. On the other hand, offline systems are not bound by such constraint, and can use computationally slow analysis tools. They can afford to use the very best analysis tools, even if very consuming.

Online algorithms also must not introduce a large delay in the data pipeline, though the delay’s magnitude is bound to the data stream’s frequency. Imagine a filtering algorithm that requires a window of 30 points on either side of the point to filter. If the data stream’s temporal resolution is 100Hz, this represents a delay of 180ms between the time this point comes in, and the time it is filtered, while the delay is of 5s for a stream frequency of 6Hz. As fNIRS has a slow response time, researchers must pay special attention to this challenge.

We evaluate the system's capabilities with two studies. Our first study compares online and offline analyses using a previous study. The second study is a proof-of-concept evaluation that uses the online features of OFAC to classify tasks in real time, and adapts an interface.

6.3 "Real Time" Analysis of a Previous Study

As a first evaluation of the OFAC system, we chose to run the real time analysis on previously recorded data to compare classification results with those obtained offline. To test data from a prior experiment, we used the offline mode in the OFAC system which loads previously recorded data and feeds the main loop one line at a time. This method tests whether the system can process data serially, and in a real time manner, meaning that we apply the same filtering algorithms as online, and the same machine learning classification design.

We selected the study presented in Chapter 5 differentiating levels of a computer game through brain activity. It showed promising classification results, in particular for the Activeness classification, comparing rest and play (average accuracy of 94%). We focus on their offline machine learning analysis, done using sequence classification with k-nearest-neighbor (kNN), k = 3, and 10-fold cross validation.

While the design of the offline analysis is typical and correct, it cannot be directly translated to an online analysis. For instance, the previous machine learning analyses used cross validation, which is not possible in real time as it requires the whole data set. Instead, we used the first sets of trials to be the training data, and tested on the rest, to reproduce a real time situation. The filtering algorithm suffers from the same problem, and we chose a method that only requires a partial dataset (moving window).

The rest of the pre-processing is identical: we re-implemented the sequence classification technique to work in real time, and used kNN (k=3) for the analysis. Sequence classification uses the entire sequence of data from the last task, and produces a class prediction. While in real time systems one might prefer algorithms that work on a time point basis, where classification is done on every point that comes in, we chose to continue to work on a task basis using sequence classification. This algorithm performed well in the previous chapter, and the goal of this chapter is to compare the analysis with the OFAC system to previous analyses, as well to prove that it performs well in real time. We leave algorithm optimization to the next round of improvements.

We tested the OFAC system's ability to classify between the participants playing and resting. The original data contained 10 examples of each difficulty level: we combined examples of both levels to obtain 20 examples of play and 20 of rest.

6.3.1 Analysis and Results

We varied the amount of examples of each class in the training dataset to evaluate the effect on accuracy and possibly observe a minimum amount of training required to run real time experiments as users should invest only a minimum amount of time for the training algorithm (Krepki, et al., 2007). We compared accuracy using training sets containing the first 2, 4, 6, 8, 10, 12, 14, 16 and 18 examples of each class (play or rest). We also evaluated multiple filtering windows to select a balance between a stronger filter and a minimum delay in the analysis stream, using filtering windows of size 1, 9, 19, 29, 39 and 49.

Before comparing the average values to previous results, we applied a factorial 2-way repeated measures analysis of variance on *Size of Training Set* (9) and *Length of Filtering Window* (6) to see which factors were significant in the analysis. As can be seen in Figure 6-9, the main effect of Size of Training Set has a significant effect on the accuracy of the classification ($F(8,64)=9.500$, $p<0.001$). The group containing higher accuracies is constituted of training sets of 12 or more examples of each class (12, 14, 16 or 18 examples). The main effect of Length of Filtering Window was not significant, nor was the interaction of these two factors.

We conclude that a minimum training set of 12 examples of each class was necessary to obtain meaningful classification results with this dataset. We also observe that every filtering window tested yielded similar results, and determine that the data is well classified with even a small filtering window which requires a smaller delay in processing. Note that using a training set of 18 examples did not yield the previous offline accuracy of 94% (even though we tested with two examples, similar to the 10-fold cross-validation). This result differs because the same filtering methods cannot be applied, and the entire dataset is not available before hand, making cross-validation impossible.

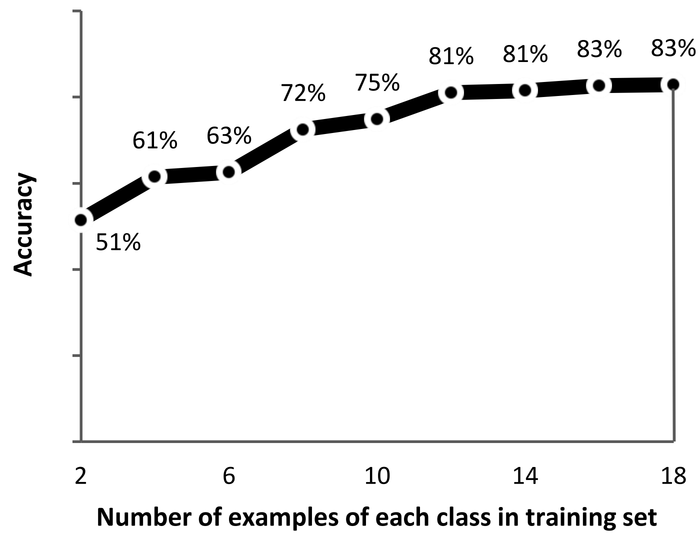

Figure 6-9. The first 12 examples (or more) in the training set produces a stable average accuracy of approximately 82%.

To compare the current results with those obtained with the offline analysis, we averaged the accuracy results obtained with training set of 12 examples and more, and all filtering windows. Over all participants, the classification accuracy obtained with the real time analysis is 82.0% (stdev 17.3%), while the one obtained with the offline analysis is 94.4% (stdev 3.7%), a decrease of 12.4%. Figure 6-10 shows the average accuracies obtained with both analyses, for each participant. We cluster the participants by their difference in accuracies: the first group contains participants with a real time accuracy equal or higher than offline, the second group shows lower accuracy (difference within 15%), and the third group with participants displaying a very low real time accuracy.

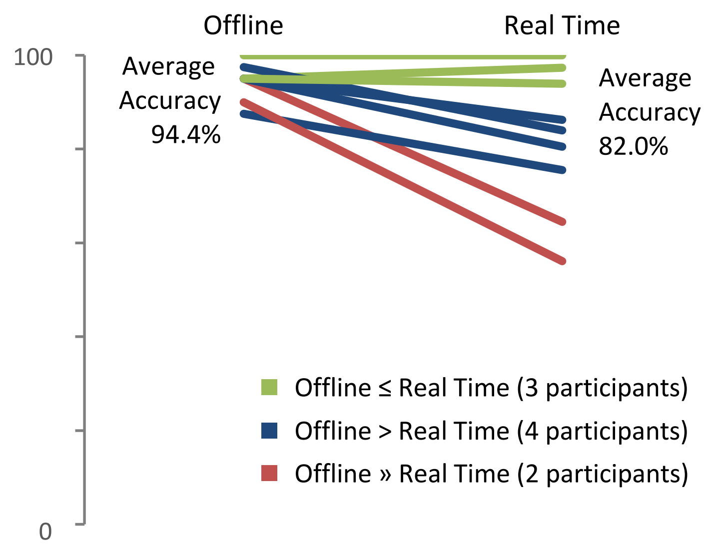

Figure 6-10. Comparing the real time and offline classification accuracy for each participant.

The results are promising: with the exception of two participants (s01 and s06), all results show accuracies over 80%, and differences between the two analyses of less than 15%. The real time results from three participants even equaled or surpassed offline accuracy.

The new analysis is, as predicted, performing worse than the offline analysis. However, we consider that this decrease in performance is outshined by the main advantages of the analysis: classifying in real time, and the ability to reuse this information to adapt the interface. Therefore, we find the real time analysis of a previously recorded dataset a success.

6.4 Real Time Task Classification and Adaptation

Having shown the classification capabilities of the OFAC system, we designed a simple proof-of-concept experiment to test the true online features of the Online fNIRS Analysis and Classification system. Our experiment has two goals: (1) to test if we can indeed process and classify in real time; and (2) to demonstrate a simple interface adaptation. To achieve these, we will distinguish between two tasks, and change the background music of the interface according to the predicted task.

Given our relative success at classifying activated periods versus rest periods, we selected tasks that would yield these two different types of signal. The first task consists of showing videos to the participant, while the second task has them play a short game of Tetris. Based on previous work, we expect the game task to activate the anterior prefrontal cortex (Chapter 5; Saito, et al., 2007), and the video task to deactivate it, if the videos are neutral or pleasant, and calm (Leon-Carrion, et al., 2007).

With this experimental design, we also hope to replace the typical rest condition with an engaging task that will not activate the aPFC. The rest task requires typically the user to watch a gray (or black) screen for 15 to 30 seconds, depending on the experiment. However, this task is not desirable in realistic graphical user interfaces: it is not engaging to the user; and it is not reasonable to interrupt the user to ask them to think of nothing for a short period in the middle of their work. But since fNIRS experimental designs are bound to include periods to allow the brain to go back to a rested state, we hypothesized that using an engaging task that does not activate the aPFC would accomplish the same goal. Specifically, we believe that using a video task instead of rest provides a better, more realistic HCI task than watching a grey screen.

We base our task selection primarily on four studies. León-Carrión et al. (2007) showed that non-arousing videos of neutral or positive valence show little activation in the prefrontal cortex (PFC) in a region encompassing both the aPFC and the dorsolateral PFC, both during and after the stimuli. In addition, Phan et al (2003) studied the effect of emotional arousal with pictures on the medial PFC with the use of fMRI. The activation found was located in areas too deep for the fNIRS probes to sense, and we do not expect this activation to be measured. Furthermore, pilot studies reproducing the latter work with fNIRS confirmed this statement. Based on those studies, we predict that a video task may be a suitable replacement for the typical rest task.

On the contrary, Chapter 5 showed activation of the aPFC during a game of Pacman, and suggested the possibility of a higher activation with a more intellectually demanding game such as Tetris, extrapolating their results with those of Siato et al. (2007).

Finally, we chose to change the background music according to the predicted task, to achieve our goal of demonstrating a simple interface adaptation. Music is often present when using computers, either related to the current task (such as when playing games), or in the background (Day, et al., 2009). Background music can play two different roles in regards to the user's attention: a distracter or an arousal inducer. Day et al. (2009) found that participants were more successful at decision making when the tempo of background music was faster, showing that faster tempo is more beneficial for harder tasks, acting as an inducer. In addition, Wakatsuki et al. (2009) showed using fNIRS no prefrontal cortex effect for music at low volume, while participants concentrating on music at high volume deactivated their PFC. We deduce from both studies that low volume background music at faster tempo might help the performance of a gaming task, and that a slower tempo should not impact the video task. While our main goal is to demonstrate an adaptation, it would be helpful to improve the performance or the satisfaction of the user's experience.

6.4.1 Participants, Protocol and Analysis settings

This study included 10 healthy volunteers (5 females), between the ages of 18 and 32 (mean 25.8, standard deviation 5.8). All participants were right-handed, had normal or corrected-to-normal vision, with no history of brain injury. Informed consent was obtained for all participants. This experiment was approved by the university's internal review board.

The session contained a total of 30 sets of two tasks (video and Tetris). Ordering of sets was randomized for each participant. Each stimulus was presented after a 3s fixation point. A minute of baseline at the beginning of the session allowed the user's brain to get to a rested state.

At the end of the session, participants answered a questionnaire pertaining to their experience with the tasks. Participants also rated the scenes using the same protocol as the one to select the videos.

To achieve our two goals, we divided the experiment into three parts: a training phase, and two testing phases - one without music, and one with adapting background music (Figure 6-11). To participants, part one and two are identical (no music).

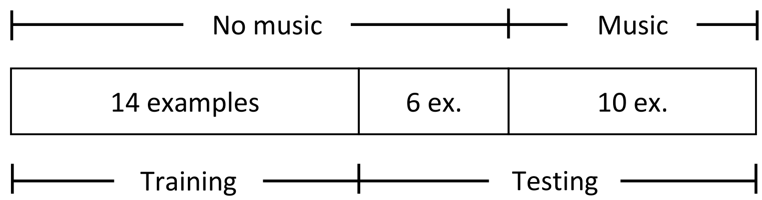

Figure 6-11. Experimental protocol and classification periods.

While we hypothesized that background music or its adaptation had no impact on the brain data, we tested some of the examples without music. Through pilot participants with this protocol, we identified that fourteen examples of each class was required to obtain stable classification accuracies, both to classify the no-music and the music examples. This correlates with results obtained with the previous study. Of the remaining sixteen examples, ten were assigned to the music condition (six to the no-music condition), because we are more interested showing the adapting interface.

We applied the signal processing method described earlier, with a moving average window of 9. We are using sequence classification with the support vector machine classification algorithm. The support vector machines algorithm finds the optimal class separation hyperplane, with the largest margin between examples of both classes.

As sequence classification produces a prediction at the end of a condition, the adaptation procedure was designed to use that prediction to change the music in the following block. The tasks' order is predetermined and alternating, meaning that the algorithm know that the following task is different than the current task. The algorithm works as follows: at the end of a task, the classifier produces a prediction. This prediction is then “inversed”, and the music associated with this inversed prediction starts to play. For example, if the current task is predicted to be a video, the algorithm “inverses” it to be Tetris, and the music associated with Tetris plays for the task that has just started.

6.4.2 Stimuli

For the video stimuli, we chose 30 clips that would fit the emotional criterion. The selected clips were 30 seconds in length, without sound, and containing mostly nature scenes (beach, forest, trees, streams, clouds). We preselected the scenes by asking an independent group of pilot participants (total of 12 participants) representative of our targeted experimental participants to rate a larger selection of scenes using the Self-Assessment Manikin (SAM) (Bradley & Lang, 1994). SAM measures the dimensions of pleasure (i.e. positive or negative) and arousal (i.e. calming or exciting) using a visual scale from 1 to 9. From those results, we chose a coherent group of scenes that scored below 5 on the arousing scale, and above 3 on the valence scale, using the same scale limits as León-Carrión et al. (2007).

For the game task, we used the Tetris Bean implementation of the classic game (Clee, 2002), with a board of 15 columns by 20 rows, illustrated in Figure 6-12. We limited the game play to 30 seconds, to mimic the video’s length. We increased the speed of the falling blocks to move every 250ms and difficulty of the game to ensure strong brain activation, and provided a preview of the block to come, to add planning to the mental task. Participants practiced a few games to familiarize themselves with this homemade Tetris version before the real experiment.

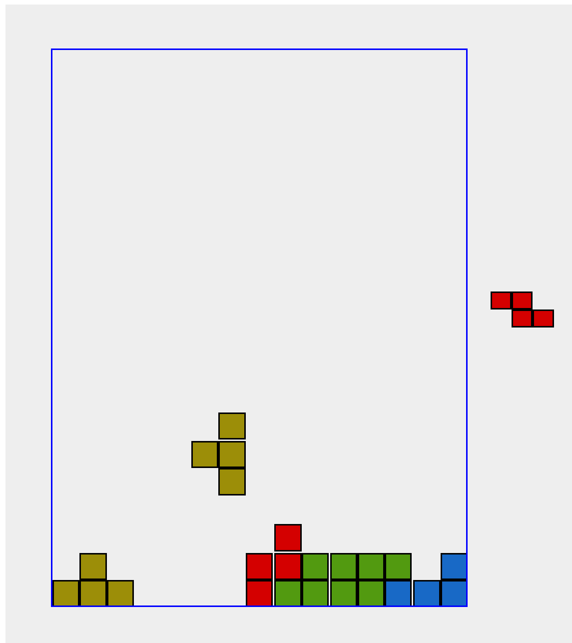

Figure 6-12. Screenshot of a Tetris game.

During the last part of the experiment, the application played quiet, continuous background music that changed according to the predicted task. If the predicted task indicated that the user was watching a relaxing video, the system would play slower tempo music, and it would play faster tempo music for the game task. The transition between the two was gradual. The slow music kept playing between tasks as to not interrupt the flow of the user, and the faster music was overlaid when Tetris was the predicted task. Both pieces of music were classical piano, intended not to cause a strong reaction among participants.

6.4.3 Results

We analyzed the results of this proof-of-concept experiment with two questions in mind. Are the classification results obtained meaningful? How did the adaptation affect the user's performance?

Behavioral Results

Clips used obtained an average valence rating of 5.7 (stdev of 1.6), and an arousal rating of 3.4 (stdev of 2.3). This confirms that the selected videos were calm and positive, as required.

We performed two t-tests on the Tetris games performance: first to determine a presence of a learning effect, and second to compare the scores with respect to music adaptation (Figure 6-13). We assigned each trial (game) to one of 3 groups: early games (trials 1 to 10), mid-session games (trials 11 to 20), and music games (trials 21 to 30).

To observe if a learning effect is present, we compared the first group (trials 1 to 10) to the second group (trials 11 to 20), both without music. We found a statistical significance between the two groups ($p<0.001$), showing a strong learning effect, with higher performance for the second set of games. Participants completed an average of 0.85 lines per game for the first trials, compared to 1.22 lines for the second set of trials.

We also compared the second group of trials with no music (trials 11 to 20) with the trials with music (trials 21 to 30) to determine the influence of the music on game performance. We observe no statistical difference between the performances with music playing, which indicates a neutral impact of the background music. This result is not contradictory of our initial goal, which was to simply demonstrate an interface adaptation.

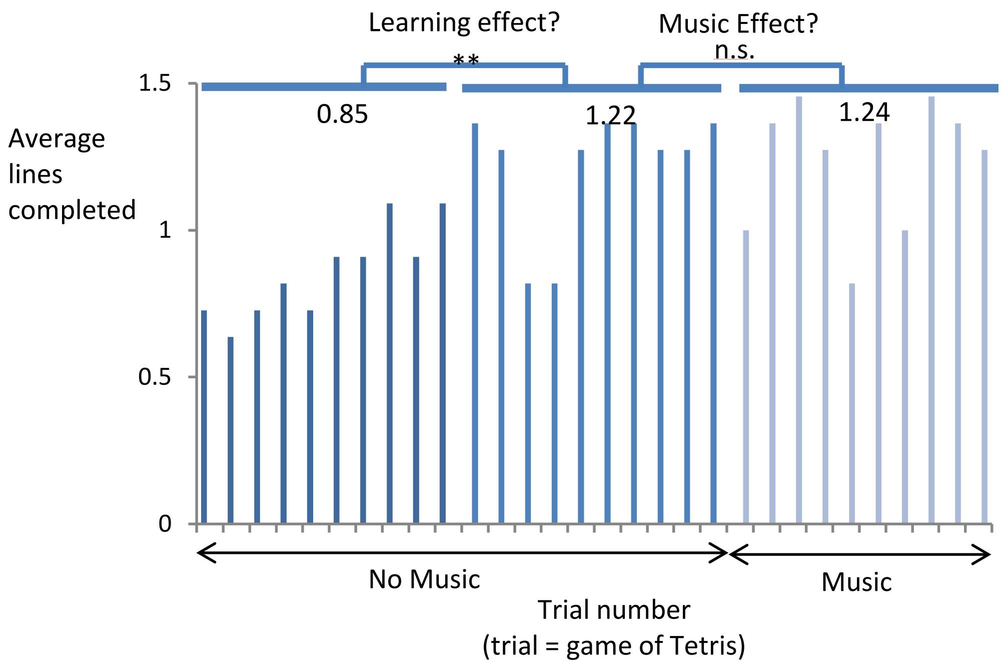

Figure 6-13. Tetris Game Performance.

Classification Results

The real time classification shows a high accuracy when distinguishing between our two types of tasks: watching a video or playing a game of Tetris. We present the results by averaging them into three groups of examples: (a) examples with no background music, (b) examples with background music, and (c) all examples (a combination of (a) and (b)). Figure 6-14 illustrates the results. When averaging the examples containing no music, the accuracy is 89.4% (stdev 8.8%), while the results with background music average to 82.5% (stdev 8.1%). A t-test showed no significance in the difference between the two groups. The overall real time accuracy is 84.9% (stdev 6.9%).

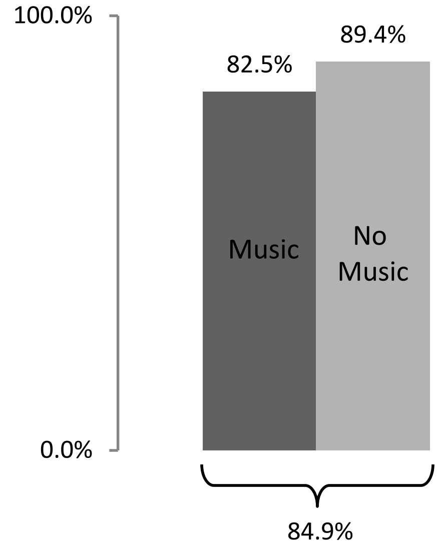

Figure 6-14. Accuracy results for real time classification of two tasks.

Individually, the classification accuracy for all examples varied between 71% and 97%. Three participants rate equal or above 90%, meaning that out of 32 classified examples, the algorithm got less than 3 wrong.

Subjective Survey Results

Our exit survey (Appendix C) revealed that all participants noticed music playing when performing both tasks, but that they only paid attention to it occasionally. Some participants focused on it when the videos came on while others "*when the music became more fast-paced."*

Their interest in the music rated 2.9 on average, on a five point scale, indicated a neutral opinion, which confirms our music choices. They all noticed classical piano, and observed the difference in tempo.

They did find the music played on a regular pattern. The slow paced music made them feel relaxed, calm, and occasionally bored, while the faster pace music was entertaining and exciting. Two participants noticed the presence of two tracks that were overlaid, and found it distracting. A better integration of two tracks (actually change between them, instead of overlaying them) would have improved their perception of that music. We do not believe, however, that this had a significant impact on their experience.

Most participants noticed that the faster music was associated with the Tetris game. However, the participant with the lowest accuracy during the music examples (65%) noted that "*it did not seem like the changes from one to the other were on a regular pattern*", which seems reasonable considering the low accuracy achieved.

They had a varied perception of the effect of the music on their performance. Half of the participants indicated that the music had no effect, while two said that it hindered, and three mentioned it helped their performance. They indicated that the positive influence was because they "*were more relaxed*," and that it gave them "*some rhythm*," while negative influence was because "*it grabbed [their] attention,"* or made them *"stressed*."

6.4.4 Discussion

Our results demonstrate the validity and reliability of the OFAC system. Not only did the system operate without fault when processing and classifying the data, but the results obtained are very encouraging. Using a simple classification algorithm, we have achieved moderate to high accuracies, of up to 97% with some participants. This compares well to Abdelnour and Huppert (2009) who achieved 79% when distinguishing between two motor tasks (left or right finger tapping) using a more complex analysis system, mainly an adaptive version of the general linear model that works in real time. With the OFAC system, we have created the first working real time passive BCI with fNIRS. We encourage other researchers to use our system with their experiments or interfaces with this demonstration.

Our task selection was successful: Tetris does activate the aPFC, and video watching does not. Furthermore, the video task did replace the ubiquitous rest task. This conclusion has many design implications for fNIRS' measured tasks, as experiments can move closer to real world scenarios.

The differences in accuracy between the music and non-music conditions could be related to brain processes. For instance, the training data was not obtained with music playing, which might not have caught brain processes differences should they have been present. However, given previous work and the analysis, we attribute the core of the differences to technical issues. Brain signals vary in time, even for the same task, due in part to the presence of a trend, meaning that the further from the training period each classified example is, the less accurate the algorithm is likely to be. Though the current protocol does well to answer our research goals, further studies containing adaptation should contain the adapted element in the training data and be counterbalanced.

Finally, we reached our adaptation goal as we demonstrate a simple adaptation through the modification of background music. However, as performance was not improved by the adaptation, and that it had mostly a neutral or positive impact on the user satisfaction, we cannot say that the adaptation improved the user's experience. Those results are in line with Kallinen et al. (2004), who found that music listening prompted higher overall user satisfaction and immersion, less boredom and more pleasure. However, low accuracies lead to higher inconsistencies in the adaptation, which can have negative impact on the performance of the participants.

6.5 OFAC Discussion and Future Work

We have learned a few lessons from our experience designing and working with a real time brain computer interface system. Delorme (2010) identifies and discusses problems related with the design and implementation of Matlab-based BCI systems, as we have observed some of these issues ourselves.

We recommend to initially test and to keep improving the overall analysis through offline analysis of the data, especially feature extraction and classification, a suggestion also pointed out by Wolpaw et al (2002). Such technique was used by Kerpki et al. (2007), who performed cross-validation on the data allocated to the classifier training, in order to evaluate the performance of the algorithm. In the case of multiple sessions with the same user and task, Millan et al. (2007) indicates optimizing the classifier with the new dataset from the current session. The integration of cross-validation methods in OFAC would also help estimating classifier performance in pilot studies.

More specifically, we recommend starting with a statistical analysis of the brain signal, before proceeding with a machine learning classification, if the first step was successful (Figure 6-15). The statistical analysis is meant to identify significant patterns in the brain activity signal, while the goal of the machine learning analysis is to assert the accuracy of classification, in order to perform a real time analysis. If offline classification results are successful, the online classification can be performed. Note that the term "successful" needs to be defined, as it may refer to a personal threshold, or a general pattern of results, for example.

Moving forward if previous step successful

1. Statistical analysis
2. Machine learning offline classification
3. Machine learning online classification

Figure 6-15. Conducting analysis in brain computer interfaces.

OFAC could benefit from more sophisticated signal processing and classification techniques. Additional data analysis could further resolve the temporal dynamics of classification efficacy, such as detecting workload changes within the first 2, 5, or 10 seconds instead of 30. Automatic determination of the amount of training data required might optimize the training period for every participant and improve accuracy. Online adaptation of the classifier would also have a positive effect of the classification accuracy. Finally, we hope to accumulate a large collection of OFAC components so the system can benefit a large research community.

On the application side, an algorithm could be designed to better integrate the predicted classes. Depending on the test paradigm, such as testing frequency, the application might need to acquire class results in a temporal queue as to filter this data to prevent quick changes in the interface (Krepki, et al., 2007). The application could also make use of the probability distribution to help the decision process. A predicted class with high probability should be adhered to with more certainty than a randomly selected class (when the probability distribution gives equal values to each class).

6.6 Adapting the Interface Passively with fNIRS

Once we are able to accurately measure mental workload, it becomes possible to create systems that adapt to the users current state of mind. As we have shown, fNIRS is well suited for such a task because it can be portable and produces real time results. The use of mental workload to create adaptive interfaces is an interesting research topic.

The design challenge is to use this information in a subtle and judicious way, as an additional, lightweight input that could make a mouse or keyboard-driven interface more intuitive or efficient. Specifically, we are exploring situations and interfaces that can be adapted slowly, in a manner that is subtle and unobtrusive to the user, which could increase productivity and decrease frustration. As a general rule for implicit interfaces, any visual modifications to the interface should be done carefully enough that the user hardly notices the change until he or she needs to (Fairclough, 2009)

There are many issues to consider when designing BCIs. Speed and accuracy have a large impact in the design of the integration of such new input into interfaces (Coyle, et al., 2007; Schlögl, et al., 2007b). We believe that user satisfaction is at least as important, especially for indirect, or passive, brain computer interfaces, as the main goal is not always to increase productivity.

Designers should be aware of the constraints and limitations of the measurement tool, and create adaptive systems that take advantage of them (or use them in a non-obstructive way). For instance, there are limitations to using fNIRS in real time, such as the metabolic response measured by fNIRS occurring over a few seconds*,* and the difficulty of filtering out motion artifacts in real time. Using fNIRS as a passive supplemental input will avoid some of these issues since the interface would not be dependent on this signal for interaction. fNIRS can be specialized in measuring short or medium length states, and not instant states like EEG. Interfaces can be adapted in a subtle matter, when a high degree of certainty in the user's cognitive state is achieved. In the case of an adaptive Pacman, changing the difficulty level should not be clearly noticeable to the user.

It has not been fully determined how to use mental workload to adapt an interface. Hancock et al. (1988) suggested that interfaces should adapt only in underload or overload situations. In situations of extreme workload, the performance is greatly reduced, and the need for regulated mental workload through interface adaptation becomes important. Others might argue the benefit of adapting the interface until mental workload is fully optimized. However, since mental workload varies with subtasks (Iqbal, et al., 2005), it becomes extremely hard to determine the exact, optimal workload for a particular user doing a particular task.

We view two main categories of adaptations that would build on the strength of fNIRS for passive BCIs: either adaptation through continuous changes, or through changes to a future input. Continuous changes adjust subtle, lightweight elements of the user's experience. Designers could modify windows properties. For instance, if the user is concentrated, we could minimize distractions by making the primary window the only one in focus, and fading the others (Girouard, et al., 2010a). The level of details of the content available to the user can vary according to the measured workload, presenting less information when the person is overloaded, to minimize their cognitive processing of that information (Figure 6-16). Changing the background music is another example of this type of adaptation.

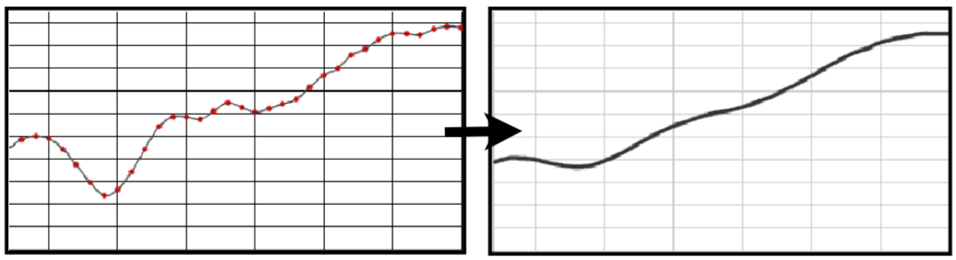

Figure 6-16. An example of high detailed graph (left), and one of low detail (right).

On the other hand, designers can also decide to make slightly stronger changes the next time the user does a particular action, or the system gets a new input. For instance, choices in menus could change according to the brain signal, such as unlocking magical functions in games if the user is in a relaxed state while in a tense situation (Ekman, Poikola, & Mäkäräien, 2008). It could also mean dynamically adjusting streams of information (emails, RSS, Twitter, etc.) to the user's mental workload so the high workload user is only interrupted with very important information, leaving them to browse the remainder when it will not disrupt as much their current state. Similarly, a video communication system like Pêle-Mêle (Gueddana & Roussel, 2009) which supports different levels of engagement (away, available, engaged) might become more active when the user is underloaded, and halt completely during periods of high workload.

Finally, the brain activity could also only be used in sporadic moments. The omnipresent screen saver application is designed to go on after a prolong period of computer inactivity. However, this detected inactivity might not be a true one: the user might be pondering on a question, reading a long text, observing a piece of code or discussing a slide that is displayed on the screen. We could detect the user's brain activity after long periods of computer activity, and if their brain is activated, they system would prevent the screen saver from starting.

One potential pitfall of interface adaptation is the dichotomy between interface design rules calling for consistency, and the ability to optimize mental workload of users, hence their comfort. There also is the risk of creating "clumsy automation" through shifts of workload, communication, attention, and coordination demands, reducing the performance of the user through additional cognitive load (Wiener, 1989). Finally, adaptation can create instability, if a first situation adapts into a second, which adapts back into the first (Alty, 2003). Designers must also be aware of the Midas Touch Problem (Jacob, 1990), where every brain signal would lead to a change in the interface, and should take steps to avoid the problem. These problems have not been fully solved, but by remaining aware of them, taking steps to avoid them when possible and making systems individually stable, and making the adaptive behavior consistent, designers can minimize their impact.

6.7 Conclusion

This chapter describes the OFAC system, our new, real time functional near-infrared spectroscopy analysis and classification system, and demonstrates through two studies the validity, reliability and potential of the system. Our first evaluation compares our previous offline analysis with our real time analysis. Results show a decrease of 12% in classification accuracy (94% to 82%), and that a minimum of twelve examples of each class is required to obtain a stable accuracy. We consider this decrease in performance to be outweighed by the main advantage of the analysis, classifying in real time, and the ability to reuse this information to adapt the interface.

The second study demonstrates the online features of OFAC: its ability to record, process, classify cognitive state signals and adapt simple interfaces in real time. We selected two tasks that would activate and deactivate the prefrontal cortex, respectively: playing a game of Tetris and showing calm videos. In addition, background music varies according to the predicted task: slower music for relaxing videos, and faster tempo for the game task. We evaluated this system through classification accuracy (average accuracy of 85%), as well as using user satisfaction of the adaptation.

Finally, we discussed optimal types of applications for brain input, such as continuous adaptation, or adaptation to a future input. The main work to be done remains to build adaptive user interfaces using the system, and determining how to evaluate them.

Chapter 7:

Conclusion

As interest for brain computer interfaces increases in the human computer interaction community, we now find more and more research that shows the ability for tools to measure and classify accurately specific brain signals. This new flow of information provides real time cognitive monitoring that can lead to usability testing, or to the creation of adaptive user interfaces. This opens up a new research direction, to study how we can intelligently make use of such brain signals as an input to the interface.

Contrary to much BCI research focusing on using this signal as the only input to the interface, I wondered how useful such technology could be if applied to the general public as an additional input. This body of work presents fNIRS as a new input device to the HCI community. fNIRS shows potential by its ability to measure different brain signals and its ease of use, and quick setup time. I believe this work to be an important step towards to using fNIRS in an adaptive user interface.

Throughout this dissertation, I tried to prove the hypothesis that *fNIRS is a good input technology for HCI, especially applied to the general public as an additional input*, and answer three questions that bring us closer to the goal of using real time fNIRS measures of cognitive load for adaptive interfaces.

First, I asked *what kind of cognitive states we can measure using fNIRS that can be used in HCI contexts*. Chapter 4 and 5 focused on identifying signals that can be reliably measured with fNIRS that are realistic to expect from human computer interfaces, and we found that the general concept of mental workload can be successfully measured in the anterior prefrontal cortex. Specifically, we observed in Chapter 4 different levels of mental workload qualified mostly as working memory, and successfully assess three levels of them at once. We also generated and measured two game difficulty levels which involve many elements that compose mental workload in Chapter 5. These two chapters are a step forward in conducting fNIRS experiments, as previous work only studied the activeness of the user, not multiple levels of it. However, the success of distinguishing multiple levels is limited.

Along the way, Chapters 3 and 6 also provided insights into user state measures. For instance, Chapter 3 reinforced our mental workload findings as it measured a significant cognitive activity during a short term memory task. Chapter 6 showed another game that activated the aPFC, but also identified that emotionally calm and neutral video content does not activate it. We can use those (intended) negative results to replace the typically necessary rest task in fNIRS experiments.

The second question posed the challenge of *developing technologies to identify those user states in real time*. Chapter 6 was dedicated to this question and introduced the Online fNIRS Analysis and Classification system to perform real time processed and classification of signals. We proved that the OFAC system is reliable, produces similar classification results as an offline analysis and can communicate with an application.

Question three inquired *how we should use this information as input to an adaptive user interface*. While I discussed adaptive systems in Chapter 6, I did not answer this question to its fullest. Instead, I categorized such adaptation and I proposed a series of adaptations possible using our system. I leave the implementation and in depth study of brain adaptive user interfaces to future work.

The subgoal transcending all the questions was *to find an accurate method for classifying multivariate sequential data from fNIRS*. The thesis used two methods of classifying fNIRS signals: sliding windows and sequence classification. Sliding windows allows the observation of the average and slope of data in small, time independent windows. On the other hand, sequence classification looks at an entire sequence of data, and uses each data point as a feature. We studied both as offline techniques and implemented real time versions of both algorithms in OFAC.

Finally, we presented a set of practical guidelines by synthesizing our results with prior work, showing which factors (physical activity or context) introduce irreparable distortions in the signal, and which can be recovered from (and how). This work confirms a large part of the hypothesis as it showed that typical interaction techniques are acceptable.

In this dissertation, I also proposed methodological insights to conducting work in this research area. In terms of experimental design, I outlined a generic experimental protocol to conduct real time fNIRS experiments. I also found a more HCI-friendly alternative to the typical rest task necessary to allow the brain to go back to a baseline state. Concerning machine learning classification, I noted the difficulty of such techniques and considerations for real time analysis and improvements.

This work also fills a gap in the literature outlined by Scerbo et al. (2001). In their report, Scerbo et al. presented a summary of psychophysiological measures and current applicability to adaptive automation through workload. Their table, however, was missing data concerning the sensitivity—the capability of the device to differentiate baseline for workload—and diagnosticity—the capability of the measure to distinguish different levels of workload—of fNIRS. The sensitivity and diagnosticity are concepts proposed by O'Donnel and Eggemeier (1986), in this case redefined by Scerbo et al. My work fills out this information: fNIRS has high sensitivity and moderate diagnosticity.

While I have shown that fNIRS is a good tool to use for HCI environments, the tool and the experiments present some limitations, especially to using fNIRS in real time, most of which have been discussed in Chapter 6. First, the metabolic response measured by fNIRS occurs over a few seconds, which creates a delay between the start of the brain activation and the start of signal measured. Second, it is difficult to filter out motion artifacts in real time. This has implications for a real time user interface such as the lack of an immediate, perfect response from the system. Using fNIRS as a passive supplemental input will avoid some of these issues since the interface would not be dependent on this signal for its interaction.

Another limitation is that fNIRS' signal is strongly dependent on the probe and area measured. Additionally, our work also was only moderately successful at distinguishing between levels of workload, indicating a limited diagnosticity. However, using another probe with more sensors, placed differently, could lead to a stronger signal, as it would pick up changes in blood oxygenation in more locations.

7.1 Future Work

With those limits outlined, there is still much interesting work to be done with fNIRS that could benefit the HCI community. I identify four specific domains of interest: interface adaptation, multimodal interaction, interface evaluation and machine learning development. Some of domains were mentioned by Nijholt et al. (2009), and Coffey et al. (2010), confirming their importance.

Developing Adaptive Interfaces

The pioneering use of fNIRS as a real time additional input unlocked a large unexplored research area that is the development of adaptive brain computer interface. I tackled a few related issues in the thesis, but more work is to be done for the creation of interactive systems that adapt their behavior to current information measured from the brain through fNIRS as a real time input. Specifically, OFAC opens up the research field to create applications using fNIRS brain signals as real time input. In addition, we hope to apply the BCI interface design and knowledge from the EEG community to fNIRS, as they have created a large number of BCI applications with EEG.

Exploring Multimodal Interaction

Chapter 2 mentioned that there is no inherent problem to combining different technologies with fNIRS. However, few researchers investigate it, probably because it generates an even larger data set. Kallenbach (2010) states that the "use of single psychophysiological measure is insufficient to create intelligent systems that adapt to the changing need, emotion, and behavior of their users during single interaction." Multimodal input should lead to better adaptive interfaces.

The combination of performance and fNIRS data may tell a better, more realistic story of the interaction of the user and the computer. A more specific evaluation of the data collected with each sensor, including performance data, could allow us to find and identify situations where performance alone is sufficient, or where fNIRS alone is sufficient.

Investigating Interface Evaluation

The area of interface evaluation with brain signals could use future work, as a literature review showed almost no dedicated effort. Evaluation of interfaces can lead to self-examination after specific computer activities and usability evaluation. It can also be presented to others while someone is being monitored, such as Chen et al.'s physiological blog (2008), or to guide semi-structured problem solving, providing knowledge to experimenters on when intervene for provide guidance.

The literature, however, shows a large number of studies measuring signals in static experiments in preparation for use in brain computer interfaces. Many of those preliminary studies could be transformed to compare interfaces.

Designing Better Classifiers

In this thesis, we designed and tested a few classification algorithms, but there are many more that could be applied to this problem. We are also interested in investigating the possibility of creating a generically trained classifier (trained in advance over a range of subjects) that classifies workload information and can be used to classify workload on a new user without an individual classifier training run.

Challenges to the community

Finally, we pose challenges to the community. Observed throughout this work and outlined by a community of psychophysiological researchers during a workshop at ACM CHI 2010 (Girouard, et al., 2010b), these methodological refinements would improve the work in brain computer interfaces, especially for human computer interaction researchers.

It would prove useful to define quality in BCI, for instance to recognize what accuracies are considered successful. It would also be helpful to understand and standardize how to get validity in the data, and how (if possible) to get and correlate it with ground truth.

We also must create a consensus on terms to use: for instance, some researchers use the terms brain computer interfaces, while other use brain machine interfaces. We also found two definitions of active and passive BCIs (see Chapter 2). A unified front would help disseminate and collect information in this field.

Other challenges reside in the field of adaptation, as defining the possible methods of adaption, and determining what or when to choose them. Finally, there is a body of work to be done in signal processing and machine learning classification and clustering.

Appendix A:

Expanded Statistical Analysis Results

This appendix reports the statistical analysis tests and results performed in the third and fifth chapter. The analysis in Chapter 3 was designed to test if the presence of certain artifacts would interfere with our ability to collect brain signals. The tests were done to answer the six questions posed in Figure 3-2. Each factorial repeated measures analysis of variance was performed on [Hb], [HbO] and [HbT]. Chapter 5's tests were designed to evaluate the presence of differences in hemoglobin concentrations between rest and game play, and between levels of difficulty.

A.1. Chapter 3 - Using fNIRS in Realistic HCI Settings

We report the p-value for tests with HCI significance, i.e. the tests from the factorial analysis with either the factor *Cognitive Task* or *Artifact*. Each table presents the results from one comparison, and we use the general term *Artifact* in place of the specific motion produced.

Table A-1 lists the tests performed to answer Comparison 2.1, with the following factorial ANOVA: Cognitive *Task* (cognitive task or rest) x *Hemisphere* (left or right) x *Channel* (4) x *Time Course* (7). Additional tests were performed for the [HbT] condition by adding Hemoglobin Type as a factor, creating the following factorial ANOVA: *Cognitive Task* (cognitive task or rest) *x Hemoglobin Type* (oxygenated or deoxygenated) x *Hemisphere* (left or right) x *Channel* (4) x *Time Course* (7).

Table A-1. P-value obtained in the Comparison 2.1 (Exp. 0) in HCI relevant interactions.

Interaction	***[Hb]***	***[HbO]***	***[HbT]***
Cognitive Task	0.330	0.755	0.488
Cognitive Task * Channel	0.031	0.087	0.188
Cognitive Task * Channel * Time Course	0.308	0.103	0.099
Cognitive Task * Hemisphere	0.300	0.534	0.402
Cognitive Task * Hemisphere * Channel	0.543	0.747	0.753
Cognitive Task * Hemisphere * Channel * Time Course	0.538	0.831	0.807
Cognitive Task * Hemisphere * Time Course	0.331	0.621	0.507
Cognitive Task * Time Course	0.292	0.544	0.420
Cognitive Task * Hemoglobin Type			0.337
Cognitive Task * Hemoglobin Type * Channel			0.059
Cognitive Task * Hemoglobin Type * Channel * Time Course			0.116
Cognitive Task * Hemoglobin Type * Hemisphere			0.800
Cognitive Task * Hemoglobin Type * Hemisphere * Channel			0.676
Cognitive Task * Hemoglobin Type * Hemisphere * Channel * Time Course			0.751
Cognitive Task * Hemoglobin Type * Hemisphere * Time Course			0.712
Cognitive Task * Hemoglobin Type * Time Course			0.297

Table A-2 . P-value obtained in the Comparison 1 (Experiment 1 to 4) in HCI relevant interactions.

Interaction	Clicking			Frowning			Head Movements			Typing		
	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]
Artifact	0.856	0.609	0.618	0.135	0.321	0.268	0.235	0.875	0.489	0.418	0.321	0.121
Artifact * Channel	0.232	0.790	0.716	0.068	0.035	0.035	0.082	0.729	0.862	0.344	0.677	0.536
Artifact * Channel * Time Course	0.436	0.659	0.762	0.082	0.065	0.064	0.405	0.674	0.685	0.253	0.724	0.582
Artifact * Hemisphere	0.866	0.631	0.716	0.857	0.639	0.742	0.803	0.954	0.900	0.915	0.371	0.502
Artifact * Hemisphere * Channel	0.841	0.128	0.215	0.371	0.474	0.538	0.073	0.233	0.116	0.689	0.918	0.814
Artifact * Hemisphere * Channel * Time Course	0.550	0.347	0.373	0.394	0.430	0.522	0.297	0.501	0.362	0.409	0.523	0.793
Artifact * Hemisphere * Time Course	0.964	0.473	0.712	0.810	0.560	0.643	0.570	0.915	0.820	0.711	0.698	0.645
Artifact * Time Course	0.756	0.750	0.713	0.199	0.134	0.071	0.452	0.330	0.600	0.735	0.034	0.087
Artifact * Hemoglobin Type			0.777			0.408			0.299			0.920
Artifact * Hemoglobin Type * Channel			0.547			0.046			0.387			0.791
Artifact * Hemoglobin Type * Channel * Time Course			0.502			0.067			0.493			0.696
Artifact * Hemoglobin Type * Hemisphere			0.678			0.491			0.961			0.428
Artifact * Hemoglobin Type * Hemisphere * Channel			0.137			0.268			0.411			0.895
Artifact * Hemoglobin Type * Hemisphere * Channel * Time Course			0.476			0.205			0.647			0.056
Artifact * Hemoglobin Type * Hemisphere * Time Course			0.307			0.448			0.849			0.822
Artifact * Hemoglobin Type * Time Course			0.761			0.284			0.107			0.288

Table A-3. P-value obtained in the Comparison 1.1 (Experiment 1 to 4) in HCI relevant interactions.

Interaction	Clicking			Frowning			Head Movements			Typing		
	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]
Artifact	0.947	0.908	0.972	0.642	0.498	0.515	0.471	0.987	0.649	0.385	0.688	0.300
Artifact * Channel	0.044	0.631	0.745	0.095	0.037	0.037	0.094	0.514	0.883	0.922	0.401	0.383
Artifact * Channel * Time Course	0.205	0.208	0.336	0.102	0.076	0.076	0.380	0.591	0.581	0.333	0.324	0.194
Artifact * Hemisphere	0.899	0.314	0.506	0.978	0.492	0.596	0.851	0.584	0.634	0.429	0.946	0.807
Artifact * Hemisphere * Channel	0.553	0.088	0.146	0.403	0.558	0.609	0.286	0.318	0.282	0.821	0.701	0.864
Artifact * Hemisphere * Channel * Time Course	0.702	0.345	0.487	0.427	0.490	0.566	0.600	0.485	0.454	0.752	0.574	0.729
Artifact * Hemisphere * Time Course	0.912	0.685	0.782	0.914	0.465	0.583	0.866	0.916	0.945	0.541	0.983	0.906
Artifact * Time Course	0.353	0.927	0.927	0.455	0.155	0.155	0.556	0.602	0.602	0.217	0.242	0.060
Artifact * Hemoglobin Type			0.873			0.486			0.558			0.735
Artifact * Hemoglobin Type * Channel			0.273			0.055			0.204			0.438
Artifact * Hemoglobin Type * Channel * Time Course			0.144			0.080			0.518			0.471
Artifact * Hemoglobin Type * Hemisphere			0.126			0.357			0.585			0.408
Artifact * Hemoglobin Type * Hemisphere * Channel			0.135			0.317			0.385			0.591
Artifact * Hemoglobin Type * Hemisphere * Channel * Time Course			0.419			0.268			0.626			0.543
Artifact * Hemoglobin Type * Hemisphere * Time Course			0.583			0.338			0.779			0.597
Artifact * Hemoglobin Type * Time Course			0.317			0.268			0.632			0.803

Table A-4. P-value obtained in the Comparison 1.2 (Experiment 1 to 4) in HCI relevant interactions.

Interaction	**Clicking**			**Frowning**			**Head Movements**			**Typing**		
	[Hb]	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**
Artifact	0.655	0.338	0.224	0.096	0.209	0.124	0.134	0.727	0.433	0.787	0.068	0.166
Artifact * Channel	0.690	0.673	0.816	0.061	0.042	0.041	0.167	0.841	0.719	0.206	0.323	0.668
Artifact * Channel * Time Course	0.671	0.291	0.451	0.094	0.074	0.069	0.568	0.222	0.260	0.366	0.273	0.311
Artifact * Hemisphere	0.852	0.013	0.034	0.762	0.797	0.898	0.471	0.432	0.395	0.556	0.108	0.252
Artifact * Hemisphere * Channel	0.651	0.495	0.554	0.354	0.389	0.46	0.028	0.337	0.077	0.042	0.726	0.512
Artifact * Hemisphere * Channel * Time Course	0.343	0.373	0.282	0.397	0.400	0.502	0.205	0.609	0.349	0.248	0.554	0.827
Artifact * Hemisphere * Time Course	0.722	0.037	0.129	0.642	0.688	0.725	0.240	0.572	0.395	0.817	0.314	0.413
Artifact * Time Course	0.498	0.131	0.534	0.461	0.204	0.163	0.066	0.08	0.483	0.154	0.052	0.192
Artifact * Hemoglobin Type			0.756			0.367			0.157			0.341
Artifact * Hemoglobin Type * Channel			0.540			0.052			0.585			0.115
Artifact * Hemoglobin Type * Channel * Time Course			0.229			0.092			0.304			0.287
Artifact * Hemoglobin Type * Hemisphere			0.048			0.643			0.570			0.191
Artifact * Hemoglobin Type * Hemisphere * Channel			0.453			0.227			0.374			0.556
Artifact * Hemoglobin Type * Hemisphere * Channel * Time Course			0.538			0.182			0.640			0.149
Artifact * Hemoglobin Type * Hemisphere * Time Course			0.069			0.587			0.798			0.432
Artifact * Hemoglobin Type * Time Course			0.143			0.415			0.002			0.012

Table A-5. P-value obtained in the Comparison 2 (Experiment 1 to 4) in HCI relevant interactions.

Interaction	Clicking			Frowning			Head Movements			Typing		
	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]	[Hb]	[HbO]	[HbT]
Cognitive Task	0.159	0.970	0.053	0.240	0.539	0.195	0.297	0.903	0.287	0.342	0.965	0.367
Cognitive Task * Channel	0.144	0.097	0.102	0.102	0.138	0.456	0.068	0.110	0.178	0.052	0.110	0.228
Cognitive Task * Channel * Time Course	0.689	0.128	0.070	0.387	0.043	0.089	0.124	0.088	0.128	0.085	0.114	0.206
Cognitive Task * Hemisphere	0.210	0.731	0.847	0.594	0.882	0.951	0.457	0.982	0.748	0.046	0.729	0.227
Cognitive Task * Hemisphere * Channel	0.302	0.915	0.744	0.448	0.519	0.519	0.592	0.916	0.777	0.720	0.760	0.646
Cognitive Task * Hemisphere * Channel * Time Course	0.622	0.763	0.852	0.726	0.605	0.685	0.666	0.859	0.812	0.820	0.669	0.604
Cognitive Task * Hemisphere * Time Course	0.338	0.359	0.507	0.869	0.678	0.733	0.217	0.455	0.321	0.110	0.525	0.376
Cognitive Task * Time Course	0.285	0.891	0.282	0.120	0.719	0.306	0.502	0.697	0.675	0.487	0.751	0.749
Cognitive Task * Hemoglobin Type			0.353			0.631			0.456			0.420
Cognitive Task * Hemoglobin Type * Channel			0.098			0.074			0.088			0.079
Cognitive Task * Hemoglobin Type * Channel * Time Course			0.195			0.042			0.080			0.088
Cognitive Task * Hemoglobin Type * Hemisphere			0.383			0.624			0.774			0.465
Cognitive Task * Hemoglobin Type * Hemisphere * Channel			0.762			0.505			0.974			0.865
Cognitive Task * Hemoglobin Type * Hemisphere * Channel * Time Course			0.654			0.497			0.837			0.741
Cognitive Task * Hemoglobin Type * Hemisphere * Time Course			0.215			0.650			0.569			0.238
Cognitive Task * Hemoglobin Type * Time Course			0.507			0.510			0.496			0.446

Table A-6. P-value obtained in the Comparison 2.2 (Experiment 1 to 4) in HCI relevant interactions.

Interaction	**Clicking**			**Frowning**			**Head Movements**			**Typing**		
	[Hb]	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**	**[Hb]**	**[HbO]**	**[HbT]**
Cognitive Task	0.160	0.765	0.082	0.290	0.443	0.268	0.354	0.944	0.373	0.532	0.394	0.399
Cognitive Task * Channel	0.695	0.220	0.160	0.249	0.354	0.744	0.316	0.185	0.306	0.503	0.361	0.485
Cognitive Task * Channel * Time Course	0.497	0.200	0.153	0.533	0.158	0.348	0.081	0.094	0.344	0.157	0.307	0.285
Cognitive Task * Hemisphere	0.189	0.089	0.288	0.954	0.499	0.667	0.930	0.314	0.499	0.047	0.599	0.180
Cognitive Task * Hemisphere * Channel	0.218	0.460	0.191	0.611	0.286	0.326	0.357	0.713	0.826	0.402	0.625	0.387
Cognitive Task * Hemisphere * Channel * Time Course	0.612	0.356	0.354	0.689	0.543	0.625	0.705	0.730	0.784	0.731	0.461	0.478
Cognitive Task * Hemisphere * Time Course	0.667	0.060	0.189	0.819	0.639	0.732	0.372	0.318	0.238	0.154	0.506	0.589
Cognitive Task * Time Course	0.155	0.657	0.384	0.269	0.669	0.462	0.725	0.946	0.828	0.155	0.598	0.204
Cognitive Task * Hemoglobin Type			0.464			0.898			0.652			0.789
Cognitive Task * Hemoglobin Type * Channel			0.286			0.154			0.166			0.380
Cognitive Task * Hemoglobin Type * Channel * Time Course			0.281			0.058			0.048			0.236
Cognitive Task * Hemoglobin Type * Hemisphere			0.047			0.240			0.351			0.088
Cognitive Task * Hemoglobin Type * Hemisphere * Channel			0.838			0.288			0.353			0.740
Cognitive Task * Hemoglobin Type * Hemisphere * Channel * Time Course			0.433			0.443			0.676			0.568
Cognitive Task * Hemoglobin Type * Hemisphere * Time Course			0.036			0.515			0.517			0.127
Cognitive Task * Hemoglobin Type * Time Course			0.233			0.901			0.775			0.630

A.2. Chapter 5 - Distinguishing Difficulty Levels

The following tests were designed for the Pacman experiment.

Table A-7. P-value obtained in HCI relevant P-value performed in Chapter 5.

Interaction	***p-value***
Difficulty Level	0.088
Difficulty Level * Channel	0.062
Difficulty Level * Channel * Hemoglobin Type	0.633
Difficulty Level * Hemisphere	0.255
Difficulty Level * Hemisphere * Channel	0.913
Difficulty Level * Hemisphere * Channel * Hemoglobin Type	0.302
Difficulty Level * Hemisphere * Hemoglobin Type	0.766
Difficulty Level * Hemoglobin Type	0.375
Activeness	0.798
Activeness * Channel	0.001
Activeness * Channel * Hemoglobin Type	0.048
Activeness * Difficulty Level	0.764
Activeness * Difficulty Level * Channel	0.294
Activeness * Difficulty Level * Channel * Hemoglobin Type	0.024
Activeness * Difficulty Level * Hemisphere	0.417
Activeness * Difficulty Level * Hemisphere * Channel	0.491
Activeness * Difficulty Level * Hemisphere * Channel * Hemoglobin Type	0.836
Activeness * Difficulty Level * Hemisphere * Hemoglobin Type	0.366
Activeness * Difficulty Level * Hemoglobin Type	0.866
Activeness * Hemisphere	0.333
Activeness * Hemisphere * Channel	0.266
Activeness * Hemisphere * Channel * Hemoglobin Type	0.358
Activeness * Hemisphere * ex	0.385
Activeness * Hemisphere * Hemoglobin Type	0.106
Activeness * Hemoglobin Type	0.235

Appendix B:

Detailed Classification Results presented in Chapter 4

Table B-1 displays the accuracy obtained per subjects for different condition combinations.

Table B-1. Average accuracy per subjects.

Conditions Combinations	*S1*	*S2*	*S3*	*S4*	*S5*	*Average Accuracy*
WL0 - WL3p	75.4%	39.6%	87.2%	90.2%	90.7%	76.6%
WL3 - WL3p	90.5%	44.6%	73.9%	89.2%	77.8%	75.2%
WL0 - WL2	39.2%	51.1%	72.7%	74.1%	43.3%	56.1%
WL0 - WL3	51.7%	44.5%	82.0%	76.4%	53.6%	61.7%
WL0 - WL4	74.1%	61.1%	73.1%	90.4%	57.2%	71.2%
WL2 - WL3	69.5%	50.2%	56.1%	50.1%	53.8%	55.9%
WL2 - WL4	68.8%	68.9%	60.1%	70.0%	49.1%	63.4%
WL3 - WL4	60.0%	70.5%	62.1%	51.6%	37.9%	56.4%
WL0 - WL2 - WL3	36.4%	39.2%	45.2%	51.4%	31.0%	40.6%
WL0 - WL2 - WL4	50.5%	67.4%	66.2%	69.7%	41.2%	59.0%
WL0 - WL3 - WL4	52.1%	43.5%	58.6%	53.6%	35.1%	48.6%
WL2 - WL3 - WL4	52.7%	35.1%	40.7%	38.7%	33.1%	40.1%
WL0 - WL2 - WL3 - WL4	37.1%	38.6%	30.2%	45.8%	22.5%	34.8%
all	33.0%	20.4%	36.6%	49.8%	32.1%	34.4%

Appendix C:

Self-Report Survey

How would you rate the game of Tetris compared to the average level of games you usually play?

	1	2	3	4	5	6	7	8	9	10	
Easiest	○	○	○	○	○	○	○	○	○	○	Hardest

How many lines do you think you completed, on average in the condition WITHOUT music?

○ 0 ○ 0.5 ◉ 1 ○ 1.5 ○ 2 ○ 2.5 ○ 3 or more

How many lines do you think you completed, on average in the condition WITH music?

○ 0 ○ 0.5 ◉ 1 ○ 1.5 ○ 2 ○ 2.5 ○ 3 or more

How satisfied are you with your performance at Tetris?

	1	2	3	4	5	
Not satisfied	○	○	○	○	○	Very satisfied

What was your strategy when playing Tetris? [Long answer]

Did that strategy change when the music was playing? [Long answer]

Did you think about something specific while you were watching the videos? Please describe. [Long answer]

Did you have any emotional reaction to some of the videos? If so, please describe the reaction for each video. [Long answer]

Did you notice music playing when you were performing the tasks (playing Tetris or watching a video)? [Yes/no]

Once the word music appeared on the screen, how often did you pay attention to the music?

○ Never ○ Sometimes ○ Always

Was there something that made you pay attention to the music? Please describe. [Long answer]

What can you tell us about the music? Please be as detailed as possible. [Long answer]

Did you perceive changes in the music, or was the same song playing in a loop? Please describe. [Long answer]

How did the music make you feel? Please describe. [Long answer]

Did you find the music to be on a regular pattern? Please describe. [Long answer]

If you noticed changes, were they associated with a particular task? Please describe. [Long answer]

Did the music help, hinder, or had no effect on your performance?

○ Help ○ No effect ○ Hinder

Did it help, hinder, or had no effect on your performance? -- Please describe. [Long answer]

Appendix C: Self-Report Survey

How much did the music help your performance?

	1	2	3	4	5	
Didn't help at all	○	○	○	○	○	Helped a lot

How much did the music disrupt your performance?

	1	2	3	4	5	
Didn't disrupt at all	○	○	○	○	○	Disrupted a lot

Did you like the music played?

	1	2	3	4	5	
Disliked a lot	○	○	○	○	○	Liked a lot

For how long would you continue doing this last set of tasks WITHOUT music playing?

	1	2	3	4	5	
Not one more minute	○	○	○	○	○	A long time

For how long would you continue doing this last set of tasks WITH music playing?

	1	2	3	4	5	
Not one more minute	○	○	○	○	○	A long time

Bibliography

Abdelnour, A. F., & Huppert, T. (2009). Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. *NeuroImage, 46*(1), 133-143.

Adams, R., Bahr, G. S., & Moreno, B. (2008). *Brain Computer Interfaces: Psychology and Pragmatic Perspectives for the Future.* Paper presented at the AISB 2008 Convention, Aberdeen, Scotland.

Akgul, C. S., B. (2005). Spectral Analysis of Event-Related Hemodynamic Responses in Functional Near Infrared Spectroscopy. *Journal of Computational Neuroscience*, 67-83.

Allanson, J., & Fairclough, S. H. (2004). A research agenda for physiological computing. *Interacting with Computers, 16*, 857-878.

Allison, B. Z., & Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological Psychology, 77*(3), 277-283.

Alty, J. L. (2003). Cognitive Workload and Adaptive Systems. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Anderson, C., & Sijercic, Z. (1996). Classification of EEG Signals from Four Subjects During Five Mental Tasks. Solving Engineering Problems with Neural Networks: Proceedings of the Conference on Engineering Applications in Neural Networks, 407--414.

Ayaz, H., Shewokis, P., Bunce, S., Schultheis, M., & Onaral, B. (2009). Assessment of Cognitive Neural Correlates for a Functional Near Infrared-Based Brain Computer Interface System *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (pp. 699-708).

Berndt, D. J., & Clifford, J. (1996). Finding patterns in time series: a dynamic programming approach *Advances in knowledge discovery and data mining* (pp. 229-248): American Association for Artificial Intelligence.

Bertini, E., Lam, H., & Perer, A. (2010). *BELIV'10: beyond time and errors novel evaluation methods for information visualization*. Paper presented at the Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, Atlanta, Georgia, USA.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *J Behav Ther Exp Psychiatry, 25*(1), 49-59.

Bunce, S., Devaraj, A., Izzetoglu, M., Onaral, B., & Pourrezaei, K. (2005). Detecting deception in the brain: a functional near-infrared spectroscopy study of neural correlates of intentional deception *Proceedings of the SPIE International Society for Optical Engineering, 5769*, 24-32.

Bunce, S. C., Izzetoglu, M., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2006). Functional Near Infrared Spectroscopy: An Emerging Neuroimaging Modality. *IEEE Engineering in Medicine and Biology Magazine, Special issue on Clinical Neuroengineering, 25*(4), 54 - 62.

Burgess, P. W., Quayle, A., & Frith, C. D. (2001). Brain regions involved in prospective memory as determined by positron emission tomography. *Neuropsychologia, 39*(6), 545-555.

Butti, M., Caffini, M., Merzagora, A. C., Bianchi, A. M., Baselli, G., Onaral, B., Secchi, P., & Cerutti, S. (2007). *Non-invasive neuroimaging: Generalized Linear Models for interpreting functional Near Infrared Spectroscopy signals.* Paper presented at the Proceedings of CNE '07 Conference on Neural Engineering, 461-464.

Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y., Pidikiti, D., & Thomas, R. (1998). A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express, 2*(10), 411-423.

Chance, B., Zhuang, Z., Chu, U., Alter, C., & Lipton, L. (1993). Cognition activated low frequency modulation of light absorption in human brain (Vol. 90, pp. 2660-2774): Proc. Natl. Acad. Sci.

Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2008). *Boredom, engagement and anxiety as indicators for adaptation to difficulty in games*. Paper presented at the Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era, Tampere, Finland.

Chen, D., Hart, J., & Vertegaal, R. (2008). Towards a Physiological Model of User Interruptability *INTERACT'07* (pp. 439-451).

Chen, J. (2007). Flow in games (and everything else). *Commun. ACM, 50*(4), 31-34.

Chenier, F., & Sawan, M. (2007). *A New Brain Imaging Device Based on fNIRS.* Paper presented at the BIOCAS 2007, 1-4.

Clee, S. (2002). Tetris Bean Retrieved January 24, 2010, from http://www.ibm.com/developerworks/java/library/j-tetris/

Coffey, E. B. J., Brouwer, A.-M., Wilschut, E. S., & van Erp, J. B. F. (2010). Brain-machine interfaces in space: Using spontaneous rather than intentionally generated brain signals. *Acta Astronautica, 67*(1-2), 1-11.

Coyle, S., Ward, T., & Markham, C. (2004). Physiological Noise in Near-infrared Spectroscopy: Implications for Optical Brain Computer Interfacing. Paper presented at the Proc. EMBS, 4540 - 4543.

Coyle, S., Ward, T., Markham, C., & McDarby, G. (2003). *On the Suitability of Near-Infrared Systems for Next Generation Brain Computer Interfaces.* Paper presented at the World Congress on Medical Physics and Biomedical Engineering, Sydney, Australia.

Coyle, S. M., Ward, T. E., & Markham, C. M. (2007). Brain-computer interface using a simplified functional near-infrared spectroscopy system. *Journal of Neural Engineering*(3), 219.

Coyne, J., Baldwin, C., Cole, A., Sibley, C., & Roberts, D. (2009). Applying Real Time Physiological Measures of Cognitive Load to Improve Training *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (pp. 469-478).

Cutrell, E., & Tan, D. S. (2008). *BCI for passive input in HCI.* Paper presented at the ACM CHI'08 Conference on Human Factors in Computing Systems Workshop on Brain-Computer Interfaces for HCI and Games.

Daly, I., Nasuto, S. J., & Warwick, K. (2008). *Towards natural human computer interaction in BCI.* Paper presented at the AISB 2008 Convention, Aberdeen, Scotland.

Day, R.-F., Lin, C.-H., Huang, W.-H., & Chuang, S.-H. (2009). Effects of music tempo and task difficulty on multi-attribute decision-making: An eye-tracking approach. *Computers in Human Behavior, 25*(1), 130-143.

Delorme, A., Kothe, C., Vankov, A., Bigdely-Shamlo, N., Oostenveld, R., Zander, T., & Makeig, S. (2010). MATLAB-Based Tools for BCI Research. In D. S. Tan & A. Nijholt (Eds.), *Brain-Computer Interfaces* (pp. 241-259): Springer.

Devaraj, A., Izzetoglu, M., Izzetoglu, K., & Onaral, B. (2004). Motion Artifact Removal for fNIR Spectroscopy for Real World Application Areas. *Proc. SPIE, 5588*, 224-229.

Dietterich, T. G. (2002). Machine Learning for Sequential Data: A Review *Structural, Syntactic, and Statistical Pattern Recognition* (Vol. 2396, pp. 15-30): Springer-Verlag.

Ehlis, A.-C., Bähne, C. G., Jacob, C. P., Herrmann, M. J., & Fallgatter, A. J. (2008). Reduced lateral prefrontal activation in adult patients with attention-deficit/hyperactivity disorder (ADHD) during a working memory task: A functional near-infrared spectroscopy (fNIRS) study. *Journal of Psychiatric Research, 42*(13), 1060-1067.

Ekman, I. M., Poikola, A. W., & Mäkäräien, M. K. (2008). *Invisible eni: using gaze and pupil size to control a game*. Paper presented at the CHI '08 extended abstracts on Human factors in computing systems, Florence, Italy.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers, 21*(1-2), 133-145.

Field, A. P., & Hole, G. (2003). *How to design and report experiments*. London; Thousand Oaks, Calif.: Sage publications Ltd., 384p.

Fink, A., Benedek, M., Grabner, R. H., Staudt, B., & Neubauer, A. C. (2007). Creativity meets neuroscience: Experimental tasks for the neuroscientific study of creative thinking. *Methods, 42*(1), 68-76.

Folley, B. S., & Park, S. (2005). Verbal creativity and schizotypal personality in relation to prefrontal hemispheric laterality: A behavioral and near-infrared optical imaging study. *Schizophrenia Research, 80*(2-3), 271-282.

Fuller, D., Sullivan, J., Essif, E., Personius, K., & Fregosi, R. (1995). Measurement of the EMG-force relationship in a human upper airway muscle. *Journal of Applied Physiology, 79*(1), 270-278.

Gevins, A., & Smith, M. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4*, 113-131.

Girouard, A., Solovey, E., Hirshfield, L., Chauncey, K., Sassaroli, A., Fantini, S., & Jacob, R. (2009). Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements *Human-Computer Interaction – INTERACT 2009* (pp. 440-452).

Girouard, A., Solovey, E. T., Hirshfield, L. M., Peck, E. M., Chauncey, K., Sassaroli, A., Fantini, S., & Jacob, R. J. K. (2010a). From Brain Signals to Adaptive Interfaces: Using fNIRS in HCI. In D. S. Tan & A. Nijholt (Eds.), *Brain-Computer Interfaces* (pp. 221-237): Springer.

Girouard, A., Solovey, E. T., Mandryk, R., Tan, D., Nacke, L., & Jacob, R. J. K. (2010b). *Brain, Body and Bytes: Psychophysiological User Interaction.* Paper presented at the Proc. ACM CHI 2010 Extended Abstracts, 4433-4436.

Goldin, D. Q., & Kanellakis, P. C. (1995). *On Similarity Queries for Time-Series Data: Constraint Specification and Implementation.* Paper presented at the Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming (CP'95).

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. N. (2008). *Feasibility and pragmatics of classifying working memory load with an electroencephalograph.* Paper presented at the Proc. CHI'08, 835-844.

Gueddana, S., & Roussel, N. (2009). *Effect of Peripheral Communication Pace on Attention Allocation in a Dual-Task Situation*. Paper presented at the Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II, Uppsala, Sweden.

Guhe, M., Liao, W., Zhu, Z., Ji, Q., Gray, W. D., & Schoelles, M. J. (2005). *Non-intrusive measurement of workload in real-time.* Paper presented at the 49th Annual Conference of the Human Factors and Ergonomics Society, 1157-1161.

Hacker, W. (2006). Mental Workload *Encyclopedia of Occupational Health and Safety* (4th ed.).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *SIGKDD Explorations, 11*(1), 10-18.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man and Cybernetics, 18*(4), 647 - 658.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theorical research. In P. Hancock, Meshkati, N. (Ed.), *Human Mental Workload* (pp. 139-183). Amsterdam.

Hattahara, S., Fujii, N., Nagae, S., Kazai, K., & Katayose, H. (2008). *Brain activity during playing video game correlates with player level*. Paper presented at the Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology, Yokohama, Japan.

Herrmann, M. J., Huter, T., Plichta, M. M., Ehlis, A.-C., Alpers, G. W., Mühlberger, A., & Fallgatter, A. J. (2008a). Enhancement of activity of the primary visual cortex during processing of emotional stimuli as measured with event-related functional near-infrared spectroscopy and event-related potentials. *Human Brain Mapping, 29*(1), 28-35.

Herrmann, M. J., Woidich, E., Schreppel, T., Pauli, P., & Fallgatter, A. J. (2008b). Brain activation for alertness measured with functional near infrared spectroscopy (fNIRS). *Psychophysiology, 45*(3), 480-486.

Hirshfield, L. M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E. T., Jacob, R. J., Sassaroli, A., & Fantini, S. (2009a). *Combining Electroencephalograph and Functional Near Infrared Spectroscopy to Explore Users' Mental Workload.* Paper presented at the Proceedings of the 5th International Conference on Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: Held as Part of HCI International 2009, San Diego, CA, 239-247.

Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J. K., Sassaroli, A., Tong, Y., & Fantini, S. (2007). *Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy.* Paper presented at the Proceedings of the ACM UIST'07 Symposium on User Interface Software and Technology.

Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J. K., Sassaroli, A., & Fantini, S. (2009b). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems, Boston, MA, USA, 2185-2194.

Hockey, G. R. J., Healey, A., Crawshaw, C. M., Wastell, D. G., & Sauer, J. (2003). Cognitive demands of collision avoidance in simulated ship control. *Human Factors, 45*, 252-265.

Horn, N. R., Dolan, M., Elliott, R., Deakin, J. F. W., & Woodruff, P. W. R. (2003). Response inhibition and impulsivity: an fMRI study. *Neuropsychologia, 41*(14), 1959-1966.

Hoshi, Y. (2009). Near-Infrared Spectroscopy for Studying Higher Cognition *Neural Correlates of Thinking* (pp. 83-93).

Huppert, T., & Boas, D. A. (2005). Hemodynamic Evoked Response NIRS data analysis GUI Program User's Guide, 2009, from http://www.nmr.mgh.harvard.edu/PMI/resources/homer/HomER_UsersGuide_05-07-28.pdf

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). *Towards an index of opportunity: understanding changes in mental workload during task execution*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Portland, Oregon, USA.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). *Task-evoked pupillary response to mental workload in human-computer interaction.* Paper presented at the CHI '04 extended abstracts on Human factors in computing systems, Vienna, Austria, 1477 - 1480.

Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., & Pourrezaei, K. (2003). *fNIR Spectroscopy As a Measure of Cognitive Task Load.* Paper presented at the Proc. IEEE EMBS.

Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., & Pourrezaei, K. (2004a). *Functional Near-Infrared Neuroimaging.* Paper presented at the Proc. IEEE EMBS.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., & Chance, B. (2004b). Functional Optical Brain Imaging Using Near-Infrared During Cognitive Tasks. *International Journal of Human-Computer Interaction, 17*(2), 211-231.

Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2007). Functional brain imaging using near-infrared technology: assessing cognitive activity in real-life situations. *Engineering in Medicine and Biology Magazine, IEEE, 26*(4), 38-46.

Izzetoglu, M., Devaraj, A., Bunce, S., & Onaral, B. (2005a). Motion Artifact Cancellation in NIR spectroscopy Using Wiener Filtering. *IEEE Transactions on Biomedical Engineering, 52*(5), 934-938.

Izzetoglu, M., Izzetoglu, K., Bunce, S., Ayaz, H., Devaraj, A., Onaral, B., & Pourrezaei, K. (2005b). Functional Near-Infrared Neuroimaging. *IEEE Trans Neural Syst Rehabil Eng, 13*(2), 153-159.

Jacob, R. J. K. (1990). *What You Look At is What You Get: Eye Movement-Based Interaction Techniques.* Paper presented at the Proc. ACM CHI'90 Human Factors in Computing Systems Conference, 11-18.

Jacob, R. J. K. (1993). Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. In H. R. H. a. D. Hix (Ed.), *Advances in Human-Computer Interaction, Vol. 4* (pp. 151-190). Norwood, N.J.: Ablex Publishing Co.

John, M. S., Kobus, D., Morrison, J., & Schmorrow, D. (2004). Overview of the DARPA Augmented Cognition Technical Integration Experiment. *International Journal of Human-Computer Interaction, 17*(2), 131-149.

Ju, W., & Leifer, L. (2008). The Design of Implicit Interactions: Making Interactive Systems Less Obnoxious. *Design Issues, 24*(3), 72-84.

Kahveci, T., Singh, A., & Gurel, A. (2002). *Similarity Searching for Multi-attribute Sequences* Paper presented at the SSDBM'02.

Kallenbach, J. (2010). Measuring Interaction Experiences: Integration of Multiple Psychophysiological Methods *Proc. ACM CHI 2010 Workshop on Brain, Body and Bytes: Psychophysiological User Interaction*.

Kallinen, K. (2004). The effects of background music on using a pocket computer in a cafeteria: immersion, emotional responses, and social richness of medium. Paper presented at the CHI '04 extended abstracts on Human factors in computing systems, Vienna, Austria.

Keirn, Z. A., & Aunon, J. I. (1990). A new mode of communication between man and his surroundings. *Biomedical Engineering, IEEE Transactions on, 37*(12), 1209.

Kennan, R. P., Horowitz, S. G., Maki, A., Yamashita, Y., Koizumi, H., & Gore, J. C. (2002). Simultaneous recording of event-related auditory oddball response using transcranial Near Infrared Optical Topography and surface EEG. *Neuroimage, 16*, 587-592.

Koechlin, E., Corrado, G., Pietrini, P., & Grafman, J. (2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences*, 130177397.

Kok, A. (1997). Event-Related-Potential (ERP) Reflections of Mental Resources: A Review and Synthesis. . *Biological Psychology, 45*, 19-56.

Kondo, T., Dan, I., & Shimada, S. (2006). Functional near-infrared spectroscopy (fNIRS) neuroimaging for infants, people with disabilities, and healthy adults: advantages and problems. Paper presented at the CODATA06 (Abstract).

Kono, T., Matsuo, K., Tsunashima, K., Kasai, K., Takizawa, R., Rogers, M. A., Yamasue, H., Yano, T., Taketani, Y., & Kato, N. (2007). Multiple-time replicability of near-infrared spectroscopy recording during prefrontal activation task in healthy men. *Neuroscience Research, 57*(4), 504.

Krepki, R., Curio, G., Blankertz, B., & Muller, K.-R. (2007). Berlin Brain-Computer Interface-The HCI communication channel for discovery. *Int. J. Hum.-Comput. Stud., 65*(5), 460-477.

Kubota, Y., Toichi, M., Shimizu, M., Mason, R. A., Findling, R. L., Yamamoto, K., & Calabrese, J. R. (2006). Prefrontal hemodynamic activity predicts false memory--A near-infrared spectroscopy study. *NeuroImage, 31*(4), 1783-1789.

Kuikkaniemi, K., Turpeinen, M., Laitinen, T., Kosunen, I., Saari, T., & Lievonen, P. (2010). *Designing Biofeedback for Games and Playful Applications* Paper presented at the ACM CHI 2010 Conference on Human Factors in Computing Systems Workshop on Brain, Body and Bytes: Psychophysiological User Interaction.

Lee, J. C., & Tan, D. S. (2006). *Using a Low-Cost Electroencephalograph for Task Classification in HCI Research.* Paper presented at the Proc. ACM Symposium on User Interface Software and Technology 2006, 81 - 90.

Leon-Carrion, J., Damas, J., Izzetoglu, K., Pourrezai, K., Martin-Rodriguez, J. F., Martin, J. M. B. y., & Dominguez-Morales, M. R. (2006). Differential time course and intensity of PFC activation for men and women in response to emotional stimuli: A functional near-infrared spectroscopy (fNIRS) study. *Neuroscience Letters, 403*(1-2), 90.

Leon-Carrion, J., Martin-Rodriguez, J. F., Damas-Lopez, J., Pourrezai, K., Izzetoglu, K., Barroso y Martin, J. M., & Dominguez-Morales, M. R. (2007). A lasting post-stimulus activation on dorsolateral prefrontal cortex is produced when processing valence and arousal in visual affective stimuli. *Neuroscience Letters, 422*(3), 147.

Luu, S., & Chau, T. (2009). Decoding subjective preference from single-trial near-infrared spectroscopy signals. *Journal of Neural Engineering*(1), 016003.

Maki, A., Yamashita, Y., Ito, Y., Watanabe, E., Mayanagi, Y., & Koizumi, H. (1995). Spatial and temporal analysis of human motor activity using noninvasive NIR topography. *Med. Phys. , 22*, 1997-2005.

Mappus, R. L., Venkatesh, G. R., Shastry, C., Israeli, A., & Jackson, M. M. (2009). *An fNIR Based BMI for Letter Construction Using Continuous Control.* Paper presented at the Proc. CHI 2009 Extended Abstracts, 3571-3576.

Marshall, S., Pleydell-Pearce, C., & Dickson, B. (2003). Integrating psychophysiological measures of cognitive workload and eye movements to detect strategy shifts. *IEEE. Proceedings of the 36th Annual Hawaii International Conference on System Sciences*.

Matsuda, G., & Hiraki, K. (2005). Prefrontal Cortex Deactivation During Video Game Play. In R. Shiratori, K. Arai & F. Kato (Eds.), *Gaming, Simulations, and Society: Research Scope and Perspective* (pp. 101-109).

Matsuda, G., & Hiraki, K. (2006). Sustained decrease in oxygenated hemoglobin during video games in the dorsal prefrontal cortex: A NIRS study of children. *NeuroImage, 29*(3), 706-711.

Matsui, M., Tanaka, K., Yonezawa, M., & Kurachi, M. (2007). Activation of the prefrontal cortex during memory learning: Near-infrared spectroscopy study. *Psychiatry and Clinical Neurosciences, 61*, 31-38.

Matthews, F., Pearlmutter, B. A., Ward, T. E., Soraghan, C., & Markham, C. (2008). Hemodynamics for Brain-Computer Interfaces. *Signal Processing Magazine, IEEE, 25*(1), 87-94.

Meek, J., Elwell, C., Khan, M., Romaya, J., Wyatt, J., Delpy, D., & Zeki, S. (1995). Regional Changes in Cerebral Haemodynamics as a Result of a Visual Stimulus Measured by Near Infrared Spectroscopy. *Proc. Roy. Soc. London, 261*, 351-356.

Merzagora, A., Izzetoglu, M., Polikar, R., Weisser, V., Onaral, B., & Schultheis, M. (2009). Functional Near-Infrared Spectroscopy and Electroencephalography: A Multimodal Imaging Approach *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (pp. 417-426).

Millan, J. (2003). Adaptive brain interfaces. *Commun. ACM, 46*(3), 74-80.

Millan, J. d. R., Ferrez, P. W., & Buttfield, A. (2007). The IDIAP Brain-Computer Interface: An Asynchronous Multiclass Approach. In G. Dornhege, J. d. R. Millan, T. Hinterberger, D. J. McFarland & M. Klaus-Robert (Eds.), *Towards Brain-Computer Interfacing* (pp. 103-110): MIT Press.

Millán, J. d. R., Renkens, F., Mouriño, J., & Gerstner, W. (2004). Brain-actuated interaction. *Artificial Intelligence, 159*(1-2), 241-259.

Minnery, B. S., & Fine, M. S. (2009). Neuroscience and the future of human-computer interaction. *interactions, 16*(2), 70-75.

Moore, M. M. (2003). Real-world applications for brain-computer interface technology. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*(2), 162.

Morinaga, K., Akiyoshi, J., Matsushita, H., Ichioka, S., Tanaka, Y., Tsuru, J., & Hanada, H. (2007). Anticipatory anxiety-induced changes in human lateral prefrontal cortex activity. *Biological Psychology, 74*(1), 34.

Morioka, S., Yamada, M., & Komori, T. (2008). Frontal Lobe Activity during the Performance of Spatial Tasks: fNIRS Study. *J Phys Ther Sci, 20*, 135-139.

Nagamitsu, S., Nagano, M., Tamashita, Y., Takashima, S., & Matsuishi, T. (2006). Prefrontal cerebral blood volume patterns while playing video games-A near-infrared spectroscopy study. *Brain & Development, 28*, 315-321.

Nijholt, A., Bos, D. P.-O., & Reuderink, B. (2009). Turning shortcomings into challenges: Brain-computer interfaces for games. *Entertainment Computing, 1*(2), 85-94.

Nijholt, A., Tan, D., Allison, B., Milan, J. d. R., & Graimann, B. (2008). *Brain-computer interfaces for hci and games.* Paper presented at the CHI '08 extended abstracts on Human factors in computing systems, Florence, Italy, 3925-3928.

Nishimura, E., Stautzenberger, J. P., Robinson, W., Downs, T. H., & Downs, J. H. (2007). A new approach to functional near-infrared technology. *Engineering in Medicine and Biology Magazine, IEEE, 26*(4), 25-29.

Nishimura, E. M., Rapoport, E. D., Wubbels, P. M., Downs, T. H., & Downs, J. H. (2010). Functional Near-Infrared Sensing (fNIR) and Environmental Control Applications. In D. S. Tan & A. Nijholt (Eds.), *Brain-Computer Interfaces* (pp. 121-132): Springer.

Noel, J. B., Bauer, K. W., & Lanning, J. W. (2005). Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. *Computers & Operations Research, 32*, 2713-2730.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload Assessment Methodology. In K. Boff, L. Kaufman & J. Thomas (Eds.), *Handbook of Perception and Human Performance Vol. II: Cognitive Processes and Performance* (pp. pp. 42/41-42/49). New York: Wiley Interscience.

OCZ Peripherals. (2010). Neural Inpulse Actuator Retrieved April 18, 2010, from http://www.ocztechnology.com/products/ocz_peripherals/nia-neural_impulse_actuator

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors, v38*(n4), p665(625).

Pfurtscheller, G., Müller-Putz, G. R., Schlögl, A., Graimann, B., Scherer, R., Leeb, R., Brunner, C., Keinrath, C., Townsend, G., Vidaurre, C., Naeem, M., Lee, F. Y., Wriessnegger, S., Zimmermann, D., Höfler, E., & Neuper, C. (2007). Graz-Brain-Computer Interface: State of Research. In G. Dornhege, J. d. R. Millan, T. Hinterberger, D. J. McFarland & M. Klaus-Robert (Eds.), *Towards Brain-Computer Interfacing* (pp. 65-84): MIT Press.

Phan, K. L., Taylor, S. F., Welsh, R. C., Decker, L. R., Noll, D. C., Nichols, T. E., Britton, J. C., & Liberzon, I. (2003). Activation of the Medial Prefrontal Cortex and Extended Amygdala by Individual Ratings of Emotional Arousal: A fMRI Study. *Biological Psychiatry, 53*, 211-215.

Pickup, L., Wilson, J., Norris, B., Mitchell, L., & Morrisroe, G. (2005). The Integrated Workload Scale (IWS): a new self-report tool to assess railway signaller workload. *Applied Ergonomics, 36*(6), 681-693.

Quaresima, V., Ferrari, M., Torricelli, A., Spinelli, L., Pifferi, A., & Cubeddu, R. (2005). Bilateral prefrontal cortex oxygenation responses to a verbal fluency task: a multichannel time-resolved near-infrared topography study. *Journal of biomedical optics, 10*(1).

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nat Rev Neurosci, 5*(3), 184-194.

Raz, A., Lieber, B., Soliman, F., Buhle, J., Posner, J., Peterson, B. S., & Posner, M. I. (2005). Ecological nuances in functional magnetic resonance imaging (fMRI): psychological stressors, posture, and hydrostatics. *NeuroImage, 25*(1), 1-7.

Riche, N. H. (2010). *Beyond system logging: human logging for evaluating information visualization.* Paper presented at the ACM CHI 2010 Conference on Human Factors in Computing Systems Workshop on BELIV'10: beyond time and errors novel evaluation methods for information visualization.

Robertson, C., Douglas, S., & Meintjes, M. (2010). Motion Artefact Removal for Functional Near Infrared Spectroscopy: a Comparison of Methods. *IEEE Transactions on Biomedical Engineering, PP*(99), 11p.

Rolfe, P. (2000). In Vivo Near-Infrared Spectroscopy. *Annu. Rev. Biomed. Eng, 02*, 715-754.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology: an International Review 53*(1), 61-86.

Saito, K., Mukawa, N., & Saito, M. (2007). *Brain Activity Comparison of Different-Genre Video Game Players.* Paper presented at the Proceedings of ICICIC '07 International Conference on Innovative Computing Information and Control 402-406.

Sakatani, K., Yamashita, D., Yamanaka, T., Oda, M., Yamashita, Y., Hoshino, T., Fujiwara, N., Murata, Y., & Katayama, Y. (2006). Changes of cerebral blood oxygenation and optical pathlength during activation and deactivation in the prefrontal cortex measured by time-resolved near infrared spectroscopy. *Life Sciences, 78*(23), 2734.

Sakurai, Y., Yoshikawa, M., & Faloutsos, C. (2005). *FTW: fast similarity search under the time warping distance*. Paper presented at the Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Baltimore, Maryland.

Sassaroli, A., Frederick, B. d., Tong, Y., Renshaw, P. F., & Fantini, S. (2006). Spatially weighted BOLD signal for comparison of functional magnetic resonance imaging and near-infrared imaging of the brain. *NeuroImage, 33*(2), 505-514.

Saulnier, P., Sharlin, E., & Greenberg, S. (2009). *Using bio-electrical signals to influence the social behaviours of domesticated robots*. Paper presented at the Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, La Jolla, California, USA.

Scerbo, M. W., Freeman, F. G., Mikulka, P. J., Parasuraman, R., Di Nocero, F., & Prinzel III, L. J. (2001). The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology: NASA Langley Technical Report Server.

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., & Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *Biomedical Engineering, IEEE Transactions on, 51*(6), 1034.

Schlögl, A., Brunner, C., Scherer, R., & Glatz, A. (2007a). BioSig: An Open-Source Software Library for BCI Research. In G. Dornhege, J. d. R. Millan, T. Hinterberger, D. J. McFarland & M. Klaus-Robert (Eds.), *Towards Brain-Computer Interfacing* (pp. 347-358): MIT Press.

Schlögl, A., Kronegg, J., Huggins, J. E., & Mason, S. G. (2007b). Evaluation Criteria for BCI Research. In G. Dornhege, J. d. R. Millan, T. Hinterberger, D. J. McFarland & M. Klaus-Robert (Eds.), *Towards Brain-Computer Interfacing* (pp. 327-342): MIT Press.

Schroeter, M. L., Kupka, T., Mildner, T., Uludag, K., & von Cramon, D. Y. (2006). Investigating the post-stimulus undershoot of the BOLD signal-A simultaneous fMRI and fNIRS study. *NeuroImage, 30*, 349-358.

Schroeter, M. L., Zysset, S., & von Cramon, D. Y. (2004). Shortening intertrial intervals in event-related cognitive studies with near-infrared spectroscopy. *NeuroImage, 22*(1), 341-346.

Shneiderman, B., & Plaisant, C. (2005). Designing the User Interface: Strategies for Effective Human-Computer Interaction, Fourth Edition. Reading, Mass.: Addison-Wesley.

Simons, J. S., Owen, A. M., Fletcher, P. C., & Burgess, P. W. (2005). Anterior prefrontal cortex and the recollection of contextual information. *Neuropsychologia, 43*(12), 1774-1783.

Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., Shimizu, K., & Birbaumer, N. (2007). Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface. *NEUROIMAGE, 34*(4), 1416-1427.

Sjölie, D., Bodin, K., Elgh, E., Eriksson, J., Janlert, L.-E., & Nyberg, L. (2010). *Effects of interactivity and 3D-motion on mental rotation brain activity in an immersive virtual environment*. Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems, Atlanta, Georgia, USA.

Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., & Jacob, R. J. K. (2009). *Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines.* Paper presented at the ACM UIST'09 Symposium on User Interface Software and Technology, 157-166

Son, I.-Y., Guhe, M., Gray, W., Yazici, B., & Schoelles, M. (2005). Human performance assessment using fNIR. *Proceedings of SPIE The International Society for Optical Engineering, 5797*, 158–169.

Sutter, E. E. (1992). The brain response interface: communication through visually-induced electrical brain responses. *Journal of Microcomputer Applications, 15*(1), 31-45.

Tanida, M., Katsuyama, M., & Sakatani, K. (2007). Relation between mental stress-induced prefrontal cortex activity and skin conditions: A near-infrared spectroscopy study. *Brain Research, 1184*, 210.

Tao, L., Masaki, O., Wanhua, H., & Atsumi, I. (2005). *Do physiological data relate to traditional usability indexes?* Paper presented at the Proceedings of the 19th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: citizens online: considerations for today and the future, Canberra, Australia.

Tian, F., Sharma, V., Kozel, F. A., & Liu, H. (2009). Functional near-infrared spectroscopy to investigate hemodynamic responses to deception in the prefrontal cortex. *Brain Research, 1303*, 120-130.

Toga, A. W., & Thompson, P. M. (2003). Mapping brain asymmetry. *Nat Rev Neurosci, 4*(1), 37-48.

Tsang, P., & Velazquez, V. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics, 39*(3), 358-381.

Ullmer, B., Ishii, H., & Jacob, R. J. K. (2005). Token+Constraint Systems for Tangible Interaction with Digital Information *ACM Transactions on Computer-Human Interaction* (Vol. 12, pp. 81-118).

Vilimek, R., & Zander, T. (2009). BC(eye): Combining Eye-Gaze Input with Brain-Computer Interaction *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments* (pp. 593-602).

Villringer, A., & Chance, B. (1997). Non-Invasive Optical Spectroscopy and Imaging of Human Brain Function. *Trends in Neuroscience, 20*, 435-442.

Villringer, A., Planck, J., Hock, C., Schleinkofer, L., & Dirnagl, U. (1993). Near Infrared Spectroscopy (NIRS): A New Tool to Study Hemodynamic Changes During Activation of Brain Function in Human Adults. *Neurosci. Lett., 154*, 101-104.

Wakatsuki, J., Akashi, K., & Uozumi, T. (2009). *Changes in Concentration of Oxygenated Hemoglobin in the Prefrontal Cortex while Absorbed in Listening to Music.* Paper presented at the Biometrics and Kansei Engineering, 2009. ICBAKE 2009. International Conference on, 195-200.

Wickens, C., & Hollands, J. (1999). Engineering Psychology and Human Performance (3rd Edition): Prentice Hall.

Wiener, E. L. (1989). Human Factors of Advanced Technology ("Glass Cockpit") Transport Aircraft (NASA Contractor Report 177528). Moffett Field, CA: NASA Ames Research Center.

Wilson, G. F., & Fisher, F. (1995). Cognitive task classification based upon topographic EEG data. *Biological Psychology, 40*, 239-250.

Wilson, G. F., & Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors, 45*(4), 635-643.

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors, 49*(6), 1005(1014).

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*(6), 767-791.

Yang, H., Zhou, Z., Liu, Y., Ruan, Z., Gong, H., Luo, Q., & Lu, Z. (2007). Gender difference in hemodynamic responses of prefrontal area to emotional stress by near-infrared spectroscopy. *Behavioural Brain Research, 178*(1), 172.

Yuksel, B. F., Donnerer, M., Tompkin, J., & Steed, A. (2010). *A novel brain-computer interface using a multi-touch surface*. Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems, Atlanta, Georgia, USA.

Zander, T., Kothe, C., Jatzev, S., & Gaertner, M. (2010). Enhancing Human-Computer Interaction with input from active and passive Brain-Computer Interfaces. In D. S. Tan & A. Nijholt (Eds.), *Brain-Computer Interfaces* (pp. 181-199): Springer.