EEG-BASED COGNITIVE LOAD ESTIMATION

By

Max Quinn

A THESIS

Presented to the Department of Biomedical Engineering
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

February 2013

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of
Max Quinn
has been approved

Mentor/Advisor

Member

Member

Abstract

Understanding and monitoring cognitive processes is a difficult task; made more so by the apparently complex underlying resource networks that make up cognition, as well as by confounds such as variability in the effort expended by subjects while performing cognitive tasks. In this paper, we describe an electroencephalogram (EEG) and machine learning-based approach to estimating the cognitive effort exerted by a subject while performing a task.

We describe the components of an EEG processing pipeline and a machine learning system, the challenges associated with employing these systems, and consider several implementation options. We contribute a novel method for separating ocular artifacts from cortical activity that represents a possible improvement upon existing techniques. In addition, we have investigated a number of alternative approaches to classification and found that these perform in a similar manner as those in prior research. Although these approaches did not produce desired improvements, they enable us to characterize some of the limitations of the current approach.

One important aspect of the present study is that it extends the domain of the EEG based approaches to a new population. Similar systems for estimating cognitive load have been successfully applied to normal subjects, or to subjects with mild cognitive impairments, while performing tasks that require substantial cognitive effort. We apply our system to subjects suffering from aphasia while performing a more naturalistic listening task with passages of varying complexity.

We find that our system is able to distinguish data recorded while listening to difficult passages from data recorded during easier passages for both controls and aphasia subjects, but we did not achieve classification results that were as high as those from earlier studies. We believe that this is a result of the subtlety of the manipulation used during data collection, and the inconsistent level of effort required during an instance of a listening task.

Contents

1 Introduction **1**
- 1.1 Prevalence of Neurocognitive Dysfunction . . . 2
- 1.2 Importance of Diagnosis and Characterization . . . 3
- 1.3 Limitations of Existing Neurodiagnostic Methods . . . 4
- 1.4 Cognitive Effort and Capacity Utilization . . . 4
- 1.5 Cortical Activity as an Indicator of Cognitive Efficacy . . . 5
- 1.6 EEG-Based Measure of Cognitive Efficacy . . . 7
- 1.7 Evidence of Effectiveness . . . 8
- 1.8 Further Developing the System . . . 9

2 Data Collection **12**
- 2.1 Naturalistic Listening Task and Subjects . . . 12
- 2.2 Data Gathered During Task Performance . . . 12
- 2.3 Appropriate Data Labeling for Individual Subjects . . . 13
- 2.4 Signal Processing and Classification Overview . . . 15

3 Overview of EEG Signals and Processing Methods **18**

4 Ocular Activity Separation **21**
- 4.1 Existing Methods and Limitations . . . 22
  - 4.1.1 Electrooculogram . . . 22
  - 4.1.2 Independent Component Analysis . . . 22
- 4.2 Template Based Approach . . . 25
  - 4.2.1 Identifying a Conserved Transient . . . 26
  - 4.2.2 Greedily Modeling the Signal . . . 27
  - 4.2.3 Updating a Transient Template . . . 27
  - 4.2.4 Iteration and Termination Conditions . . . 29
  - 4.2.5 Discussion . . . 29
- 4.3 Modeling Eye-Movement During a Reading Task . . . 29

5 Feature Extraction and Classification **31**
- 5.1 Feature Extraction . . . 32
- 5.2 Classifier . . . 36
- 5.3 Feature Selection . . . 37
- 5.4 Alternative Classifiers and Feature Selections . . . 39
  - 5.4.1 Hierarchical Logistic Regression . . . 39
  - 5.4.2 Artificial Neural Network . . . 39
  - 5.4.3 Principal Components Analysis . . . 42

6 Results and Discussion **43**
6.1 Cognitive Load Estimation using Nominal Difficulty Labels 43
6.2 Subject Responses 45

7 Conclusions **49**

References **51**

A Additional Results **55**

1 Introduction

Investigation into intelligent systems, both natural and artificial, has long held understanding the elements of human cognition as a key goal. For cognitive scientists, understanding isolated cognitive abilities is of interest on its own, but can also give insight into the internal representations of information used in complex reasoning. For clinicians, viewing cognition through component abilities provides a framework for understanding the various neuropsychological dysfunctions that result when these abilities are somehow perturbed.

Cognitive processes are not directly observable, leading clinicians and researchers to rely on behavioral testing, wherein batteries of tasks are used to ellicit observable behaviors. Data generated by these behaviors are then used to infer the function and limitations of cognitive resources such as memory, attention, and various sensory-motor processes. An individual subject exhibits significant variability in his or her performance of these tasks, leading researchers to rely heavily on population data gathered through tests constructed to isolate highly specific cognitive abilities. The result is that tools useful in mapping out an average resource network are not tailored to making inferences about an individual subject. Issues like learning, fatigue, the trade-off between speed and accuracy, and the impact of cognitive effort play an important and often unaccounted-for role in performance.

Traditional tests, often performed with pen and pencil under the supervision of a test administrator, can now be implemented electronically, making administration more effective and efficient, enabling more frequent assessment on a large scale, but not significantly addressing the limitations of the tests themselves. These tools have been effective in mapping out prototypical resource networks, but do not account for variables that impact task performance for individual subjects. In particular, the role of cognitive effort, the extent to which available cognitive resources are utilized for a particular task, presents a significant confound in the investigation of neurological and cognitive disorders. It may be the case that patients are able to maintain high task performance through the allocation of cognitive resources not normally associated with the task at hand. In a real-world setting, these additional resources may be occupied by concurrent activities, leading to an underestimation of the difficulty of the task for the patient. Rather than focusing entirely on performance, integrating a measure of cognitive effort can provide the information necessary to assess individual cognitive ability, without the opaqueness associated with existing tools [35]. However, research into systems that integrate a measure of effort have generally relied on a disruptive assessment, requiring subjects to switch between a primary task and a self evaluation.

Electroencephalography (EEG) methods show promise in using physiological data to estimate cognitive effort. Recent advancements in EEG hardware are quickly improving the ease-of-use of such systems, making EEG-based methods of estimating cognitive effort a viable addition to traditional cognitive testing, and creating the opportunity for the development of new methods for the evaluation of cognitive efficacy through ecologically valid activities. However, practices in EEG signal processing are often dependent on skilled

technicians to identify specific patterns in the resulting data. This reliance on a technician undercuts the potential application of EEG to long term and large scale cognitive monitoring, as the quantity of data collected quickly makes the expense associated with processing a problem, and the reliance on the judgement of a technician introduces variability that is difficult to account for, especially when data from a single subject might be tracked for years. On the other hand, automated methods for low level processing and higher level analysis do exist. Additional hardware can aid in the low-level processing, but interferes with the usability of the systems by involving a more elaborate setup procedure. Higher level analysis can be performed using systems like that developed by Advanced Brain Monitoring [1] but require comparison against a group norm, limiting the flexibility of systems, and their applicability to patients who have non-standard neural activity patterns due to their unique injuries and recoveries.

1.1 Prevalence of Neurocognitive Dysfunction

Acquired neurocognitive dysfunction can occur as the result of sudden events or cumulative processes, can impact patients at any time of life, and can leave sufferers with deficits that range in the specificity of impaired abilities.

Effects of brain injuries initiated by discrete events, such as traumatic injuries or stroke, often manifest clearly and soon after an injury, although the extent of and the details of resulting cognitive impairments may be difficult to characterize. Stroke impacts 540,000 Americans each year, with each instance potentially leading to a constellation of impairments that can interact to produce highly specific cognitive dysfunction. For example, aphasia, the loss of language function, can be a severely debilitating result of acquired brain injuries from stroke. Aphasia can arise from focal injuries to regions of the brain that play a critical role in language production and comprehension, such as Broca's and Wernicke's area, but it may also result from diffuse brain injuries that are more associated with impaired working memory.

The effects of stroke related brain injury often become less observable with time, but might persist for years. Even when the general nature of the impairment is clear, it can be difficult to monitor progress during recovery. It may seem appear that an impairment impacting understanding of spoken language has faded when a patient provides behavioral cues indicating that they understand and are comfortable with verbal instructions. Later it can be revealed that patients are expending great effort in performing simple listening tasks.

Traumatic brain injury, caused by events such as vehicular accidents that create strong forces on the brain and lead to diffuse microstructural axonal injury [33], impacts as many as 300,000 Americans each year, in addition to one sixth of veterans of combat operations in Iraq and Afghanistan [20, 21]. Most instances are classified as "mild", following a brief loss or alteration in consciousness [4], yet, even here, a significant minority of patients, 10-20%, exhibit cognitive impairments such as difficulty concentrating and memory problems

for months or years. As traumatic events such as stroke or vehicular accidents are made less fatal by improving medical technology, it has become clear that survival is frequently the first step in a long recovery process, with substantial challenges left to address.

Other cognitive dysfunction comes about as the result of a long term processes, either in the form of a cumulative injury, as in the case of chronic traumatic encephalopathy, or as the result of a neurodegenerative disorder, such as Alzheimer's or Parkinson's diseases. The research into the prevalence of chronic traumatic encephalopathy (CTE) is still early, but there is the possibility that repeated head impacts that characterize boxing and American football could be threatening the cognitive and brain health of many young athletes. CTE leads to cognitive impairments similar to those seen as a result of mild traumatic brain in jury, and long-term, debilitating symptoms similar to those found in advanced Parkinson's.

The case for the prevalence and future impact of Alzheimer's disease is far clearer. Primarily effecting the elderly population, Alzheimer's slowly dismantles the cognitive abilities necessary to function in everyday life, forcing patients into a life of dependency. A 2011 Congressional Research Service study projects that the proportion of the population over 65 will nearly double over the first half of this century, with this group expected to make up 20% of the population by 2050. With this shift, the prevalence of Alzheimer's is expected to follow suit. The same report suggests that we may already be approaching a tipping point, where the needs of the elderly will begin to dramatically outstrip the capacity of the medical infrastructure to provide necessary care.

1.2 Importance of Diagnosis and Characterization

For each patient suffering from a neurocognitive dysfunction, the specific effects on cognition can be unique, but in all cases, cognitive dysfunction can interrupt or permanently alter the trajectory of the sufferer's life. The key to making appropriate decisions about clinical response lies in understanding and characterizing the extent of an injury and the associated impairments.

For traumatic injuries and stroke, this means preventing further injury during the period of increased risk that follows the injury event. During the acute and subacute phase following TBI, physical activity presents this risk, and is believed to be inversely associated with an individuals stage of recovery from the original injury [8]. Further, the cognitive deficits commonly associated with acute/subacute mild TBI would be expected to result in reduced ability to perform common daily tasks. In some cases, this creates an added risk associated with in-home falls or vehicle accidents [29, 12, 30]. For these reasons, it is important to have practical and effective means to assess individuals soon after an injury, allowing for the identification of impairments and the modification of behavior to mitigate these resulting risks.

In the chronic phase following a trauma or stroke, individuals may require continued clinical care. Accurate characterization of neural injuries and cognitive functioning at this later stage is essential to the process of efficiently determining the medical and rehabilitative

treatments that may be necessary to reduce residual symptoms, returning the individual to the highest level of functioning possible.

In the case of neurodegenerative diseases, efforts fall on improving the efficacy and efficiency of care based on the changing needs of the patient. Modest progress in extending the autonomy of this patient population can significantly impact the quality of life and efficiency of care for patients, but this goal must be balanced with careful evaluation of the safety of independent living. Making progress in safely preserving autonomy requires that tools enable the identification of an individuals abilities through frequent cognitive assessment. These tools must be sensitive to subtle changes that build over time and must be compatible with ecologically valid tasks that indicate the ability of a patient to function independently.

1.3 Limitations of Existing Neurodiagnostic Methods

The limitations of existing diagnostics of brain injury and cognitive impairment have prevented the effective identification of need for assistance and delivery of rehabilitative treatment.

Although studies show the effects of mTBI, in the form of microstructural axonal injury, and of Alzheimer's in post-mortem studies, these effects are not visible in living sufferers of mTBI, and often only detectable in late stages of the disease in the case of Alzheimer's. Instead, researchers and clinicians rely on behavioral measures to identify acute symptoms of injury, residual symptoms long after injury, and symptoms of degeneration [16]. Individual tests can lack sensitivity to the effects of mild brain injuries, as performance on cognitive tests results from a variety of factors, some unrelated to the injury itself, and performance measures can often fall into normal ranges defined by population norms [10]. Given a large battery of tests, these cases can often be identified [38], but the high variance in resulting measures for individual subjects makes it difficult to track the recovery process with any granularity.

1.4 Cognitive Effort and Capacity Utilization

The concept of cognitive effort can provide an important additional dimension in the inference of cognitive health through behavioral studies.

This concept is understood through an information processing view of cognition. Through this perspective, cognition operates in a resource constrained environment, where higher level processes are constrained by the limitations of lower level cognitive resources. These resources include operations such as information retrieval and attentional allocation, and memory resources such as *working memory*; a capacity that stores information necessary for a task at hand [24, 39].

Within this framework, *capacity utilization* is a measure of the proportion of the total available resources that are being used over a given period of time. Determinants of resource

utilization include: the resources available to an individual subject, the demands of the task, and the additional considerations that accompany task performance, such as speed and accuracy concerns. As the demands of the task increase, more of the available resources are recruited. When the pool of available cognitive resources is exhausted, *cognitive efficacy* will begin to decline, leading to failures in the component processes, and a decrease in task performance. *Cognitive effort* is a measure of the perceptible experience of operating at high capacity utilization. In the case of tasks that have cognitive resource requirements significant enough to initiate errors, increases in the experience of cognitive effort can indicate the approaching failure before it occurs [40].

Disorders such as mTBI and Alzheimer's causes diffuse injury that affects distributed neural networks, compromising executive, storage, retrieval, and processing abilities; as well as reducing cognitive capacity. These changes in cognitive capacity can impact the ability for individuals to perform more complex tasks that rely on these component resources and require greater capacity.

1.5 Cortical Activity as an Indicator of Cognitive Efficacy

Neuroimaging studies have revealed neural correlates of effortful task performance. Research shows that the amplitude and spatial extent of cortical activity increases as performance becomes effortful, even as behavioral performance on tasks remains the same. For example, McAllister et al. demonstrated this effect using subjects having experienced a mild head injury resulting in a brief loss of consciousness, but no abnormalities observable through structural MRI. Patients in this study reported more symptoms from the head injury check list than controls, worse memory of recent events, and more difficulty concentrating. In their study, mTBI patients and healthy controls performed a well known memory task while cortical activity was recorded using functional magnetic resonance imaging (fMRI).

The cognitive test used in the study, the n-back, taxes working memory by requiring that a subject store and update an ordered queue, where the length of the necessary queue can be manipulated through a test parameter. Stimuli are drawn from a related set of objects such as digits, letters, words, images, or tones, and are presented in sequence. The subject is instructed to attend to the stimuli and indicate when two of the same stimuli have been presented in the sequence with a particular offset, indicated by the n in *n-back*. In a 1-back digit instance of the test, a subject might be presented with the sequence '1 1 2 3 2 4 3', and should indicate a detection on the second occurrence of '2', but not the second occurrences of '1' or '3', which would instead by correct detections in the 0-back and 2-back versions of the test, respectively. To perform the task, a subject must maintain a queue of the n+1 stimuli that preceded the current stimuli. If working memory is limited, we expect that some instance of the n-back will require storage of a sequence that exhausts the available storage, leading to errors. The relationship between the sequence length and the errors is used to infer properties of working memory.

Injured and healthy subjects show no difference in terms of task performance, despite significant differences on standardized, self-reported measures of short term memory and concentration ability. Both groups perform with high accuracy on low load instances of the test, with an offset of 0 or 1; with a slight decrease in performance with an offset of 2; and a significant decrease with an offset of 3. Accepting this similarity in performance as indicative of health may delay the detection of slowly developing deficits such as Alzheimer's and the detection of mild TBI using common test batteries such as the MMSE [14].

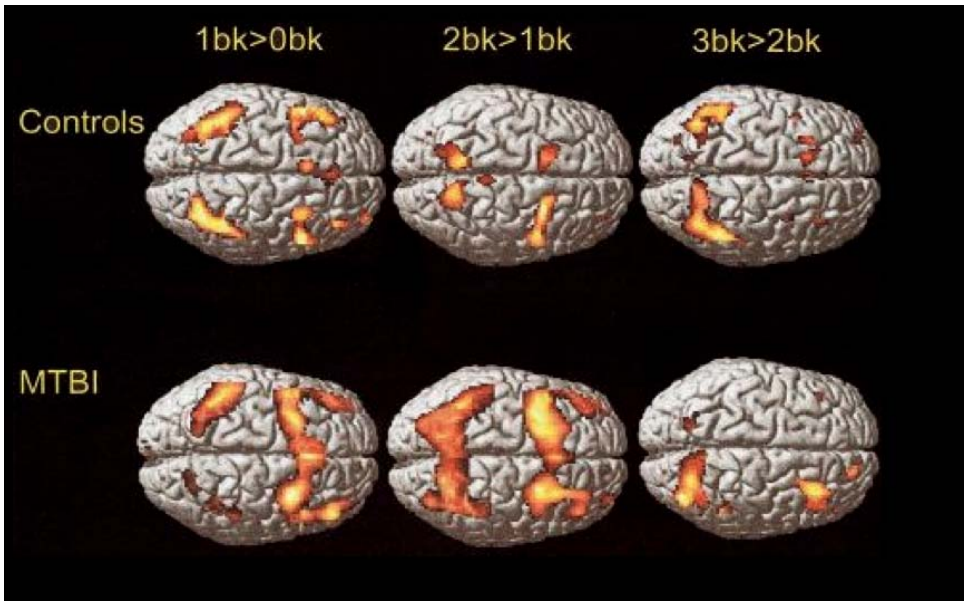

Figure 1: A figure from the McAllister et al. study shows fMRI data gathered from healthy subjects and subjects impaired by mTBI. Both groups performed the n-back task using n = 0,1,2,3. The change in activity associated with an increase in task complexity is more widely distributed and higher amplitude for mTBI sufferers, suggesting more resources are being recruited to perform the task. However, the change in performance and absolute performance for each group is unchanged [28].

However, fMRI data showed clear differences in the level and distribution of cortical activity under different task conditions between injured and control groups, despite equal performance. As can be seen in figure 1, controls showed limited increases in cortical activity in isolated regions with increased task difficulty. Injured subjects showed more broadly distributed changes in activity as the task complexity increased. The injured group presented this broad distribution of activity for task conditions with lower demands as well as during the highly demanding conditions, especially in parietal regions, which the researchers attributed to an increase in working memory utilization requirements to maintain performance [28]. It may be the case that a sufficiently granular test would reveal

differences through performance, but the physiological signal makes the differences clear.

Other studies show a similarly broad distribution of cortical activity in subjects reporting memory dysfunction for tasks expected to have low cognitive load requirements based on activity recorded from healthy subjects [13].

1.6 EEG-Based Measure of Cognitive Efficacy

These results suggest that cortical activity may be a reliable indicator of cognitive effort, even when performance based measures are not. However, to apply a method for estimating cognitive effort broadly for frequent evaluation, to use it in controlled rehabilitation, and to apply it in the development of new evaluative cognitive tasks, there are additional practical constraints that must be considered. It is important to consider the types of activities that can be performed while data is being gathered, and the cost associated with frequent evaluation.

Previous results have largely relied on data gathered through the use of fMRI, which is both expensive and impractical in many settings. fMRI requires that testing take place in a hospital setting and that subjects remain nearly still, requirements that rule out many existing cognitive tests that include a motor component, and preclude evaluation of natural activities that take place in familiar settings. Other hemodynamic measures such as near-infrared spectroscopy (fNIRS) and transcranial Doppler (TCD) have limited spacial resolution, which does not allow for the detection of the broad recruitment of regions that seems to be associated with increased cognitive effort, as well as being influenced by general changes in vascular activity not related to changing cortical activity. Electrophysiological data gathered from electrocardiogram (ECG) and galvanic skin response (GSR) have been used to evaluate cognitive state via polygraph, but neither measure cortical activity and are both influenced by non-cognitive factors that make their validity questionable in this context.

EEG sensors record data generated by neuronal firings that create electrical potentials that can read at the scalp. These recordings provide a rich source of information regarding cortical activity, which researchers have used to identify particular spatial and temporal patterns in cortical activity, evoked response potentials (ERPs), associated with particular cognitive events. Successfully characterizing these types of fast, perceptual events, which happen on the time-scale of tens of milliseconds, is made possible by the high temporal resolution of EEG. Frequency domain data has been used in the characterization of cortical activity associated with sustained cognitive activity such as vigilance and working memory. Gamma band activity over many regions is associated with increased cognitive load, as is alpha and theta band activity in the frontal midline cortex, but because there are many recording locations and potential frequency bands that make up the space of possible indicators of cognitive load, it has argued that an estimate of cognitive load should be derived from multiple regions and activity bands rather than any particular region or frequency [31].

1.7 Evidence of Effectiveness

Researchers at Honeywell's Human Centered Systems group have developed system for estimating cognitive load using an approach based on machine-learning and pattern recognition. Although initially motivated by the potential for brain state to be used in advanced user interface design and evaluation, the approach has been found to be effective in identifying scalp regions and cortical activity frequencies that are discriminative of cognitive load levels in a cognitively impaired population [27]. This finding motivates our further development of the system, and its application in cognitive load estimation for a more severely impaired patient population. What follows is a description of the Honeywell study, which served as a starting-point for our present research.

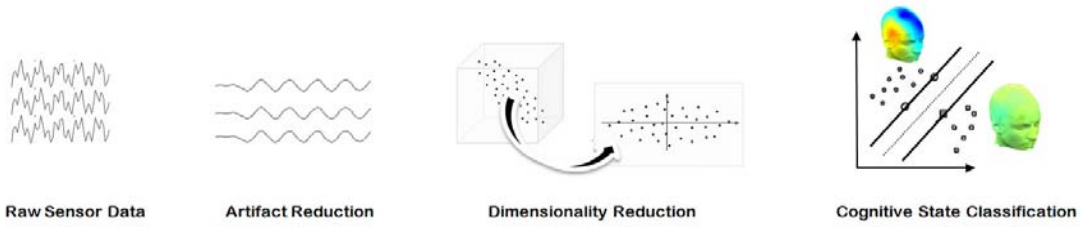

Figure 2: There are several processing steps involved in Honeywell's method for estimating cognitive load from EEG. Features based on spectral density estimates are extracted from short segments of EEG data. These features can be corrupted by artifacts that overlap with the frequency bands of interest, such as eye-blinks and muscle activity. In this initial study, these effects were not reduced through filtering, segment exclusion, or other methods. Spectral density features are extracted from each electrode location. A discriminative subset of electrodes and frequency bands is chosen using a feature selection algorithm such as sequential forward floating search (SFFS). A classifier (such as a support vector machine) is then built to discriminate between cognitive states using example data gathered from the intended subject.

Subjects, Task, and Data Ten patients reporting mild cognitive impairment following chemotherapy performed a reading task of high or low complexity. The passages were selected based on standard automated measures of text complexity, as well as subjective scoring provided by the researchers. These initial categorizations were supported by subjective scores provided by subjects following completion of the reading task. Subjects were instructed that they would read a passage, and then answer several questions regarding the content of the passage, although these responses were not considered in the analysis of the resulting data. While the passages were read, EEG data was recorded at 64 standard scalp locations with a sampling rate of 256Hz.

Feature Extraction and Selection Power spectral density features were extracted from

each scalp location using a five second window. Feature were extracted every 2.5 seconds, leading to an overlap of 2.5 seconds for adjacent windows of data. The power in four standard activity bands was integrated, creating a 256 element feature vector (64 locations $\times$ 4 bands) for each window of activity. These vectors were divided into training and testing sets. Because adjacent window of activity are defined with overlap, this division into training and testing sets uses sequential blocks of data, minimizing the signal shared between sets. A discriminative subset of electrodes was selected using sequential floating forward search (SFFS), a heuristic method that constructively builds a set of features and sometimes culls them based on several considerations [37]. The set is initialized with the single most discriminative feature. Additional features are added to the set on the basis of the increase in classification performance that they provide. Features already selected are removed conditional on the best classification previously achieved given a particular feature set cardinality. SFFS is described in more detail in section 5.3.

Classification and Analysis Features from windows of recorded data were predicted to have been recorded during a difficult reading task or easy reading task using logistic regression. Logistic regression returns an estimated probability that the data is associated with a particular label. When employing a classifier, the intended use of the classification has some bearing on what probability will be considered sufficient evidence when making a decision. However, without knowledge of how a system will eventually be employed, it is reasonable to consider the quality of the classifier using a metric that does not involve a thresholding step. Therefore, classification results were reported using the area under the receiver operating characteristic curve (AUROC), a unitless measure generated from unthresholded outputs of the regressor. If the model is providing uniformly random outputs, the AUROC would be near .5, where perfect classification results in an AUROC of 1.0 [11]. The Honeywell study reported a cross-validated AUROC of .84, averaged over the 10 subjects, a score associated with a clear effect, but one in which there is substantial noise.

1.8 Further Developing the System

These results are encouraging, suggesting that systems based on EEG recordings and machine learning-based processing are able to identify periods of high and low cognitive load with a good degree of accuracy. This paper documents our efforts toward improving this technique in terms of accuracy and reliability, improving usability, and evaluating applicability for patients with pronounced aphasia.

Before discussing our changes to the system, we will first discuss a few lingering issues that the initial study left unaddressed. The preliminary experiment made use of a reading task, the type of task enabled by the use an unobtrusively collected neurophysiological data, but one where factors usually controlled in neurocognitive imaging experiments were

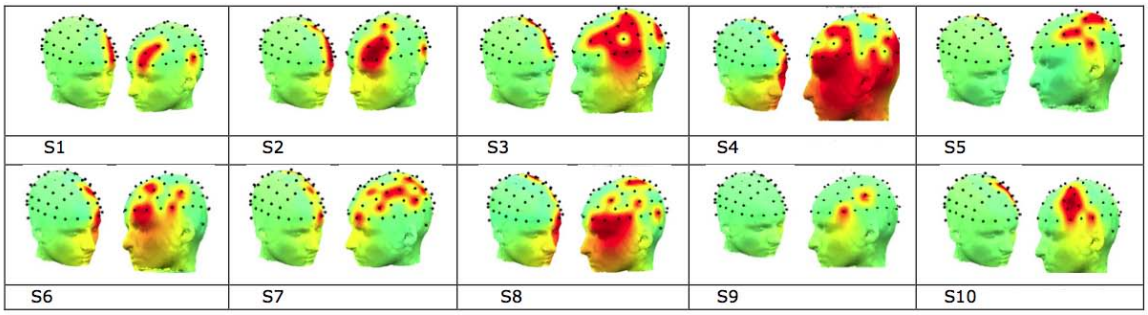

Figure 3: Ten subjects reporting mild cognitive impairment related to chemotherapy performed a reading task including passages of high or low complexity. A subset of scalp locations were selected for each subject that were found to be highly discriminative of the task condition. Scalp locations that were included in this subset are displayed in red and locations that were not included are displayed in green. Discriminative regions are primarily concentrated in left side frontal regions, consistent with expectations for a language intensive task.

not accounted for. This left the concern that signals not directly associated with cognition, such as the frequency of eye movements, could account for some of the system's classification performance. In investigating this potential alternative explanation, we considered signal processing techniques that might be useful in separating eye movements (as well as other signals not directly related to brain activity) from cortical activity that we assume to be cognitively relevant. We started with an approach based on independent component analysis (ICA), a technique for blind signal separation that has been shown to be effective in separating ocular activity from cortical activity in prior research [23]. However, this technique proved to be somewhat unreliable, displaying instability in results over repeated applications to the same data. As one of the goals of our project was improve usability of the method, we wanted to automate the processing of EEG signals as much as possible. Existing methods for automating ocular artifact removal rely on additional hardware [22], an added complexity that is counter to the goal of improved usability. We developed an alternative method for separating ocular components from cortical activity that is based on a modification of a general technique, matching pursuit. This method has been successfully applied in the EEG domain [8], but is not frequently used, we assume for runtime considerations. Our modifications significantly mitigate this issue and appears to separate ocular and cortical activity quite well, although further refinement and a thorough comparison of techniques remains future work.

The investigation into ocular activity supported the hypothesis that underlying cortical activity does account for the classification performance in the preliminary Honeywell experiment, as well as rejecting the hypothesis that eye activity alone can be used to distinguish high from low cognitive load during a reading task, in so far as one can be confident in our

methods for separating cortically generated signals from other signals. However, removing these ocular events did not effect a significant change in classification performance, as the EEG sites that were most significantly processed were rarely identified as useful sites for the purposes of cognitive load estimation.

Moving past the initial study and its associated data, we collected a new data set that could be used to answer questions regarding applicability and optimization of the cognitive load estimation technique for aphasia subjects engaging in listening tasks. We collected data from several patients that suffered from stroke related aphasia, as well as several control subjects, in conjunction with Honeywell and the Program in Occupational Therapy at Washington University in St. Louis. EEG data was recorded while subjects performed a spoken language understanding task in which they listened to short passages from sources of varying complexity. These passages were interleaved with multiple choice questions relating to the subjects understanding of the passages. The EEG recordings and the subject responses serves as the basis for the analyses that follow.

To evaluate our system for predicting cognitive load based on EEG activity, we need to apply labels to the EEG data both for building a model of the underlying activity, and for evaluating the predictions made using the model. Because cognitive load is not directly observable or reliably induced, we considered several methods using subjective responses to infer cognitive load associated with particular passages. We evaluated the inferred labels by comparing them to the nominal labelings provided by researchers. The motivating hypothesis for this was that the underlying cognitive state would at least correlate with the nominal labeling, but that the nominal labeling would be noisy for any particular subject. For example, a subject might not recognize the complexity of a passage, in which case the EEG data associated with this passage would not be wisely utilized in modeling activity associated with high load. By evaluating classification efficacy with alternative labelings, we hoped to find a better measure of the underlying cognitive state, which would provide a better target labeling for our signal based approach.

Finally, we considered alterations to the signal processing and classification pipeline. We investigated alternative feature extraction methods, dimensionality reduction methods, and classifier options. The combinatorics of an exhaustive search over a modest set of implementation options quickly becomes unfeasible, but we found was that the initial decisions made in the preliminary Honeywell study were quite reasonable with respect to all three pipeline components.

2 Data Collection

2.1 Naturalistic Listening Task and Subjects

Colleagues from the Washington University School of Medicine's Program in Occupational Therapy constructed a set of passages that would serve as the manipulation of cognitive processing. The challenge in constructing this set was in finding passages that both controls and subjects with aphasia would recognize as an easy or challenging passage. Passages were initially drawn from early reading tutorials meant for 1st or 2nd grade readers and from popular periodicals such as The Economist and The New Yorker, providing an initial pool of 40 passages between two and three minutes in duration. A pilot study was performed in which subjects performed a listening task with candidate passages and rated their difficulty. Nine control subjects were recruited from the Program in Occupational Therapy and seven subjects with aphasia were recruited from the Cognitive Rehabilitation Research Group Stroke Registry at Washington University. Subjects listened to passages and then provided ratings of difficulty and understanding on a five point scale, with a score of 1 represented a passage of low difficulty and a score of 5 representing a passage of high difficulty. Mean difficulty ratings for these passages were used to label passages as either easy or difficult, where passages with scores falling between 1 and 1.9 being assigned to the easy category, and passages with scores falling between 2 and 5 being assigned to the difficult category.

This norming study led to the selection of 16 difficult and 20 easy passages that subjects listened to over two sessions on sequential days. Subjects listen to easy or difficult passages and provide subjective responses relating to the perceived difficulty of the passages and to their understanding of the passages, and answered comprehension questions regarding the content of the passages. Difficulty was measured on a scale from 1 to 5, where 1 indicates a very easy passage and 5 indicates a very difficult passage. Understanding was reported on a 1 to 5 scale where 1 indicates poor understanding and 5 indicates very good understanding. Twelve subjects were recruited from the Cognitive Rehabilitation Research Group Stroke Registry. Aphasia subjects all suffered a left hemisphere stroke at least 6 months prior to the study, had chronic aphasia, and were English speakers. The mean age of these subjects was 64.5 years and were 58 months post stroke. Control subjects had a mean age of 25.5 and had no history of neurological disorders.

2.2 Data Gathered During Task Performance

While subjects performed trials of the listening task, EEG activity was gathered using a Biosemi ActiveTwo Mk2 system with 300 MOhm impedance and 31.25 nV resolution [3]. EEG signals were sampled at 2048Hz from 64 Ag-AgCl electrodes, held in place with an elastic cap, at scalp locations: Fp1, AF7, AF3, F1, F3, F5, F7, FT7, FC5, FC3, FC1, C1, C3, C5, T7, TP7, CP5, CP3, CP1, P1, P3, P5, P7, P9, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, Fpz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, CP6, CP4, CP2, P2, P4, P6, P8, P10, PO8, PO4, and O2. Signals were

recorded using a dedicated machine running ActiView, an application provided by Biosemi for sampling, filtering, and recording EEG data [4]. The ActiView software high-pass filtered the signals at 0.1Hz and down-sampled to 256Hz during recording.

A custom application was developed by researchers at Honeywell to manage trials. This software was developed using Presentation [32], an application and API used for the development of audio/visual psychological experiments that require precise timing, and run on a dedicated machine. In addition to playing audio files containing passages and displaying prompts for subjective responses and comprehension questions; it passed information regarding the state of the trial and subject responses to the recording machine via a low latency parallel port connection. Data pulses were sent to indicate the start of passage trials, the nominal difficulty of trials, the beginning of comprehension questions, the correctness of responses to questions, and the subjective difficulty ratings provided by the subjects. These data pulses were recorded by ActiView as an additional signal stream, in sync with the 64 signal streams from scalp electrodes. This allowed us to precisely identify the beginning and end of trials in the resulting EEG data.

2.3 Appropriate Data Labeling for Individual Subjects

The pilot study indicated that the passages used represented an effective manipulation for both controls and subjects with aphasia. However, one of the goals of the study was to avoid reliance on population norms when measuring cognitive effort. We wanted to account for the possibility that the nominal difficulty of a passage, recognized by researchers and the pilot cohort as being of a particular difficulty, may not be an effective manipulation for particular subjects. It is possible that the complexity of a passage is not recognized, failing to evoke an elevated cognitive load and the associated cortical activity; or that a nominally easy passage is difficult for a particular subject, evoking activity associated with heavy cognitive load. Using such events as examples of cognitive states associated with the nominal difficulty of the passage would contribute to a less effective measure of cognitive effort for the particular subjects.

To address this concern, we first considered using each subject's assessment of passage difficulty as the ground-truth labeling on associated EEG recordings. However, different subjects used the range of available responses differently. For example, some subjects would only rate the difficulty of passages 1 or 2 on the available 5 point scale, but might use more of the available range when indicating their understanding of the passage. Additionally, subjects seemed to express their experience of the passages using different questions that followed. Most patients seemed to provide consistent responses regarding the difficulty of the passage and their understanding, but not uniformly.

We addressed this additional subject to subject variability by constructing an individualized projection from the subject responses to a difficulty indicator variable. By using a simple, linear discriminant based on the Fischer criterion to relate the subject responses to the nominal labeling, we hoped to utilize the particular usage of the responses by the

	Nominal	Correctness	Confidence	Understanding	Difficulty	Projection
Nominal	1.00	-0.23	-0.49	-0.52	0.59	0.66
Correctness	-0.23	1.00	0.39	0.23	-0.24	-0.31
Confidence	-0.49	0.39	1.00	0.52	-0.52	-0.65
Understanding	-0.52	0.23	0.52	1.00	-0.64	-0.69
Difficulty	0.59	-0.24	-0.52	-0.64	1.00	0.90
Projection	0.66	-0.31	-0.65	-0.69	0.90	1.00

Table 1: The correlation matrices for passage properties, subjective responses, and the combined projection were calculated for each subject. For each pair of variables, the mean correlation coefficient over all subjects was calculated to produce the above table. Of note, the LDA-based projection is very closely related to the subjectively reported difficulty. Actual correctness of response is most closely related to reported confidence in the response, and is only weakly correlated with the nominal or perceived difficulty of the passage.

subject, but with a simple model that would not be strongly influenced by the occasional mismatch between nominal labeling and subject experience. The resulting projection produced a more granular measure of difficulty than any of the individual dimensions. These variables were then thresholded at the median of resulting values to assign new "difficult" and "easy" labels to the recorded EEG data.

The correlation matrix in table 1 outlines the relationship between the nominal labels applied to passages, the subjective responses provided by individual subjects, and the discriminative projections used to relabel the data. Notice that the correlation between the projection and the assessed difficulty is higher than the correlation between the projection and the nominal labeling that served as the target of the learned linear discriminant. This indicates that no combination of the subjective responses could very closely approximate the nominal labeling for individual subjects, supporting our concern that group norms may not provide the most accurate indication of brain state during the tasks. Instead, the resulting aggregate measure often weighted the subjective difficulty rating highly. Shifting from the nominal labeling to the projected labeling made us more confident that our labels reflected the state of the subject during the task.

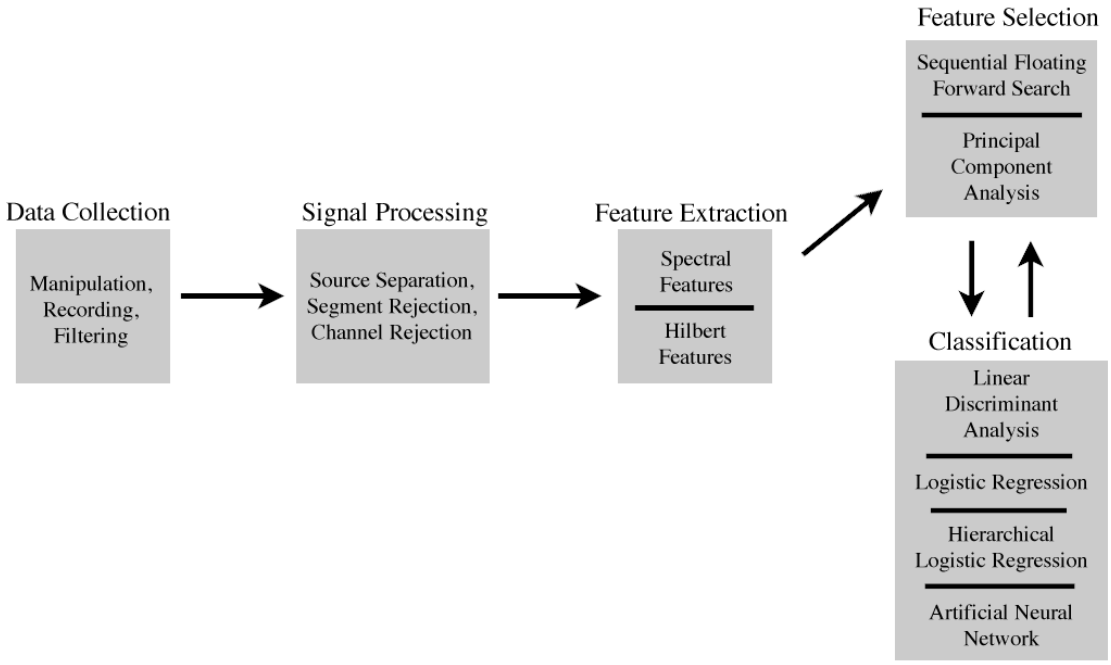

Figure 4: The system for cognitive load estimation consists of multiple modules, each of which may consist of several steps and may be implemented in a variety of ways. The above figure summarizes those modules, and several of the steps or options involved in each. Modules consisting of lists essentially express a related family of operations, all of which must be performed. Modules divided by horizontal separators denote available options that were evaluated in our analysis.

2.4 Signal Processing and Classification Overview

To further develop our cognitive load estimation system, we wanted to critically examine each of the basic components that make up the data processing and classification pipeline. The method consists of a sequence of steps that are common to many machine learning systems, each of which can be accomplished in numerous ways with varying techniques. Which techniques are used for a particular application depends on factors like the structure and type of data and the intended usage. In the first iteration of the system, these selections were made to demonstrate feasibility, relying on methods associated with prior successes in EEG-based brain state estimation [2], as well as those known for flexibility rather than specific domain success [25]. This left open the possibility that the particular implementation, although effective, was sub-optimal.

The first element of the pipeline consisted of a combination of traditional signal process-

ing steps, as well as techniques more specific to EEG processing, meant to non-destructively separate cognitively relevant elements of the recorded EEG data from unimportant or non-biologically derived signals. This included steps such as simple trend removal and more complex tasks such as the isolation and removal of ocular activity. These steps are performed in the absence of a target signal against which to compare techniques, so largely relies on prior knowledge to develop and evaluate.

Feature extraction is the process of transforming data in a manner that is meant to clear away information that is redundant or not useful for the indented use, while retaining important elements. For EEG data, which is characterized by high frequency temporal sampling and granular location sampling, this transformation is to a data space with a much lower dimensionality. After transformation, the data might be indexed by coarser temporal samples, broad frequency bands, and estimated sources that integrates multiple location samples. In some sense, the feature extraction process injects our expectations about what information in the EEG signal will be useful for estimating cognitive load.

Feature selection is a process by which the features that have been extracted are further pared down. When using classifiers, it is often the case that high-dimensional data will contain many statistical characteristics that will appear to be significant in predicting traits about the data. Statistical characteristics are often spurious, appearing to provide explanatory power for one dataset, but failing to generalize to new data. This issue is referred to as *over-fitting*, and is especially common when dealing with high-dimensional data such as EEG data. The frustrating association between high-dimensional data and poor generalization is commonly referred to as the *curse of dimensionality*. Feature selection can help to further summarize the data, which often mitigates this problem [7]. Some methods do this by explicitly evaluating subsets of features on the basis of generating models that generalize to new data. These methods, sometimes called *wrappers*, can be very robust in avoiding over-fitting and are represented in our study by sequential forward floating search (SFFS) [37]. Other methods allow for data to be further transformed based on statistical properties so as to de-emphasize lower amplitude noise and to make spurious associations less frequent. These are referred to as *filter methods* and are represented in our study by principal component analysis (PCA) [19]. Although effective in many contexts, we found over-fitting to remain a significant problem when applied to our data.

Classifiers make the final projection from a multi-dimensional feature space to a single dimensional feature space. The space of classifier options is diverse, as classifiers are more or less well suited to a particular application dependent on properties of the data. The amount of data available, the dimensionality of the data, the statistical relationships between available features, the reliability of the training data, the type of outputs being sought, and training and runtime considerations all play into the appropriateness of particular classifiers. Even when many of these characteristics are known, it still remains difficult to make the right selection of classifier. In practice, the selection is often an expression of prior knowledge and an exploration of several options.

In the case of binary classification, this single dimensional feature may be threshold so

that it consist of simply a binary inclusion or exclusion indicator. Classifiers generally fall into one of two classes, distinguished largely by how directly they can be understood within a statistical framework. *Generative models* construct a statistical model for observed data that includes a class variable. The model is built such that instances from the data can be interpreted as emissions of the model (hence generative), and the likelihood of a particular emission can be calculated explicitly. When new data is presented, the likelihood of an instance is calculated, conditional on each of the possible class labels (such as high cognitive load or low cognitive load). The label that produces the highest likelihood is interpreted as the classification.

Discriminative models may not construct a model that emits data, instead focusing the separation of training data using a family of transformation functions. These functions map values from a high dimensional representation of an input to a *decision variable* in the reals. This decision variable is thresholded to identify a class label. Discriminative classifiers include nearest neighbor approaches, support vector machines, Fischer criterion classifiers, logistic regressors, or naive Bayesian classifiers. Because much of our focus is on constructing a high performance system for estimating cognitive load, and because generalized linear models have performed well for classifying cognitive state in the past [36], we have remained largely restricted to discriminative models.

As in the Honeywell study, we evaluated the quality of a resulting classifier using the AUROC. The exclusion of a thresholding step prevents the incompatibility of comparison that arrises from accuracy and false positive reporting, or from sensitivity and specificity reporting, a property that has made it popular in a variety of classification and detection contexts. It can be intuitively interpreted through an equivalent probabilistic phrasing. Let a model assign a score to each instance from a binary classification task where each instance has a label of 0 or 1. Then the AUROC value for the model is equal to

$$P(score(x) > score(y)|label(x) = 1, label(y) = 0).$$

That is, if you were to draw one instance from each class, the AUROC for the classifier is the probability that the score for the instance from one particular class is higher than the score for the instance from the other [11]. The comparison of instances from separate classes means the system is being evaluated without respect to the distribution of data between the classes or to any particular threshold. Both of these properties may be more or less desirable when evaluating with a particular application in mind, but for data that is nearly balanced between classes, it allows for the easy comparison of systems.

3 Overview of EEG Signals and Processing Methods

The initial processing of EEG data is performed without reference to a target signal, leading to a reliance on prior knowledge regarding the generation and process of measuring the signals to select and evaluate reasonable processing methods. What follows is a brief discussion of common sources of electrical potentials recorded by the EEG, how they combine, and some of the processing options suggested by this view of EEG activity.

Recorded EEG signals are the composition of many signal sources, including biological sources such as brain activity, ocular activity, and muscle activity and non-biological sources of noise such as environmental electrical activity. Brain related signals are believed to be generated by the mass activation of large populations of neurons which create local net potentials. These local potentials then combine to create a net potential at each scalp location. The signal contributions combine additively and decrease in intensity as a function of the distance to the signal source. EEG uses an array of electrodes, placed on the scalp and held in positions with a specially designed cap, to sample these net potentials at a high frequency. At each of these locations, the spacing between the electrode and the signal sources creates a unique additive mixture of the source signals.

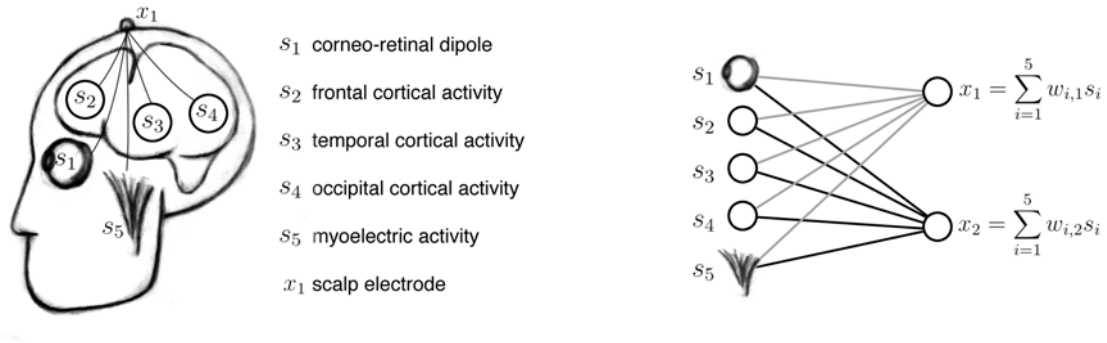

Figure 5: Electrical potentials are generated by the mass action of related populations of neurons or by other electrically active anatomical structures such as facial musculature or eyes. These potentials travel through the head with negligible delay. A mixture of these signals is recorded at the scalp with an electrode held in place with a specialized cap. The contribution of a signal source to a recording is a function of both its intensity and the distance between the source and the recording electrode. These source to recording site distances vary for each electrode, varying the influence of each source in the net potential. This situation is modeled by representing each electrode's record as a uniquely weighted sum of underlying source activities.

Some contributions to the mixed signal are expected to be related to cognitive activity, others less so, and others still are indicative only of a poor recording that should not be considered. For our purposes, the early processing steps are principally concerned with

decomposing the recorded signals into multiple signals associated with the contributing sources. These signals can then be used separately to inform cognitive load estimates.

Signal sources that contribute to the recorded EEG signal consist of biological factors such as,

- Skin impedance
- Respiratory activity
- Cardiac activity
- Brain related activity of a periodic nature
- Brain related activity of a transient nature
- Ocular activity, including blinks and eye movements
- Muscle activity

non-biological factors such as,

- Changes in electrolytic gel impedance
- Amplification with poor electrode to scalp contact
- Hardware problems
- Environmental electrical interference (ie, line noise)
- Perturbations of the electrode/scalp contact

Many of these contributing factors are discarded, as we do not consider them cognitively relevant, making it convenient to group unwanted signals by their signal properties, and how they might be removed, rather than by their source.

Slowly changing components can be removed or separated through high pass filtering. These components include,

- Skin impedance
- Respiratory activity
- Electrolytic gel impedance

Nearly stationary components can be sufficiently separated with notch filtering or adaptive filtering. These components include,

- Cardiac activity

- Respiratory activity
- Line noise

However, these signals are often low amplitude or fall into frequency ranges that do not overlap with the clinical frequency bands that we use in our analysis, they do not require specific attention to remove.

Outlier segments, where the distribution of frequency power and the signal amplitudes are dramatically different from what is expected to come from biological activity, should generally be discarded. During these segments, perturbations of the equipment or muscular activity create signals that are more powerful than cortical activity by orders of magnitude, making it unlikely that a reliable cortical signal could be recovered. These components include,

- Muscle activity
- Electrode/scalp perturbations

These segments can be hand marked for exclusion from model building. To further automate the early processing of EEG data, we identified such segments by their abnormal frequency power distributions, which are often characterized by unusually high power in high frequencies.

Frequent, transient signals present a more significant challenge, as they are non-stationary and generate broad-band activity in the frequency domain, making them difficult to filter. Although often recognizable, they are frequent enough that it would be high disruptive to simply remove signal segments that contain them, as can be done in the case of other muscle activity. These components include,

- Ocular activity, blinks
- Ocular activity, eye movements

There are several methods and active research in how best to deal with these sorts of transient events. We will discuss existing options and our own solution in the next section.

4 Ocular Activity Separation

Ocular activity can produce pronounced, high amplitude, impulse-like signals that create broadband activity when viewed in frequency space. An example of ocular activity can be seen in Figure 6. This activity is generated by the movement of a dipole created by retina and cornea. A potential is generated during any eye movement, such as horizontal eye movements made when reading or searching, or vertical eye movements made actively or made reflexively while blinking. It is preferable to limit eye movement during EEG recordings because this activity is not considered meaningful, with respect to cognitive activity, and can be disruptive when trying to analyzing resulting signals. Using EEG for cognitive monitoring may not allow for this limitation, as one of the appeals of the EEG approach is the potential for minimal modification of naturalistic activities such as reading, and failing to adequetly address such artifacts has been shown to degrade performance in similar EEG-based systems [17].

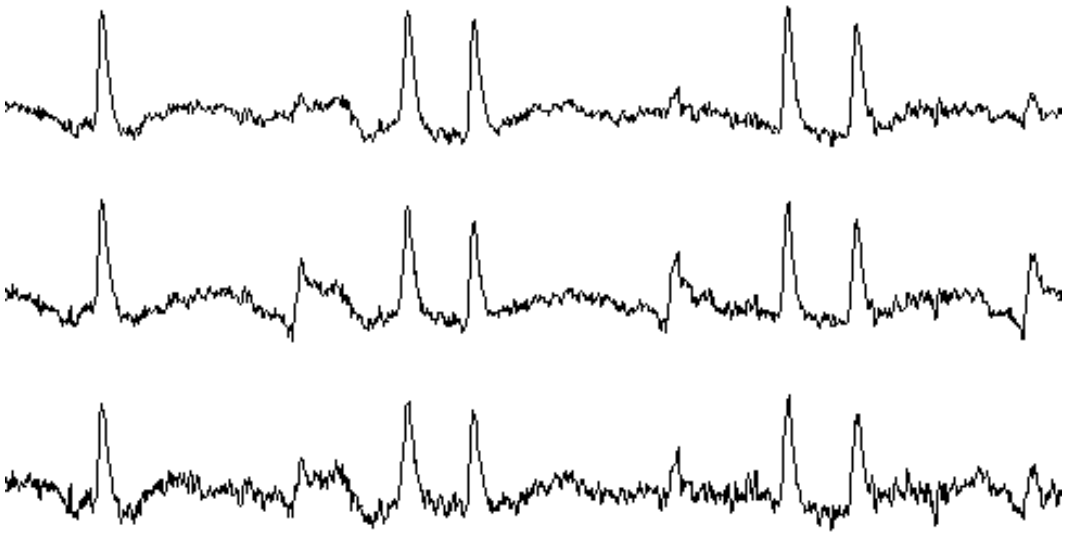

Figure 6: Above are EEG signals, recorded over several seconds, from three electrodes located near the front of the scalp. The peaked activations are ocular events. The similarity between these signals, with respect to the ocular events, forms the basis for the separation of the ocular contribution to the EEG recording.

Ocular activity left us with two concerns with respect to our investigation. The first was that the signals created by the frequent eye-movements during a reading task may be involved in the classification performance demonstrated in the preliminary Honeywell

study. The second was that these activations might be overwhelming useful information in the frontal channels that could otherwise be used by the cognitive load estimation system to discriminate between cognitive states. To investigate both of these concerns, we wanted to separate ocular and cortical signals into distinct components, allowing them to be evaluated individually.

4.1 Existing Methods and Limitations

The removal of ocular signals is a common element in research based EEG usage. Traditional and adaptive filter based approaches are generally insufficient, as eye related activations are high amplitude, broad-band, and non-stationary. The signals of interest overlap with the spectrum of the ocular signal, and are dramatically diminished or altered by applying a filter to the affected channels uniformly [17].

However, the conserved time-course of eye-related events makes them easily recognizable. A simple censoring approach is problematic, as eye movements are frequent, and removing affected segments of the signal would remove a significant portion of the data. Where a technician may be able to ignore evidence from particular channels when it is difficult to interpret (due to the presence of an ocular event), missing data is a challenging issue when automated classification systems are employed.

4.1.1 Electrooculogram

The inclusion of an electrooculogram (EOG) consists of the addition of one to many additional electrodes around the eyes. The expectation is that these electrodes will record eye activity, but not cortical activity. This signal will be used to infer what portion of scalp recorded signals can be attributed to eye activity. The EOG signal can then be subtracted from each electrode based on the estimated effect. Identifying weights for this process is non-trivial. It has been found that the weights relating the EOG to scalp electrodes depends on the type of eye activity being considered. That is, if $x_{brain} = x_{EEG} - \beta x_{EOG}$, then $\beta_{blink} \neq \beta_{saccade}$. Therefore, effective use of the EOG might also involve a method for recognizing types of eye movements, either during experimentation or during analysis. Furthermore, researchers have found that these methods benefit from the inclusion of multi-electrode recordings at each eye, which further reduces usability and increases potential variability due to poor placement [5].

4.1.2 Independent Component Analysis

Independent component analysis (ICA) is a statistical method that has been shown to be effective in separating ocular activity from cortical activity without recording additional data with an EOG [23]. ICA assumes that recorded signals X are generated through a linear combination of independent signal sources S with a mixing matrix M such that $X = MS$. These independent sources are recovered by estimating the inverse of this

mixing matrix through an optimization. There are several options for the measure to be optimized, but each is essentially a measure of non-Gaussianity, as the additive combination of independent random variables should be more Gaussian than the variables alone, by the Central Limit Theorem. One such measure of non-Gaussianity is kurtosis, in which case, the optimization is

$$\arg\max_{M^{-1}} \sum_i \mid kurt(S_i)| = \arg\max_{M^{-1}} \sum_i \left| \frac{\mathrm{E}((S_i - \mathrm{E}(S_i))^4)}{\mathrm{E}((S_i - \mathrm{E}(S_i))^2)^2} - 3 \right|$$

$$\text{where } S = M^{-1}X.$$

Tools such as gradient descent or Newton's method can be employed to perform this optimization. A visual representation of the effect of ICA can be seen in Figure 7.

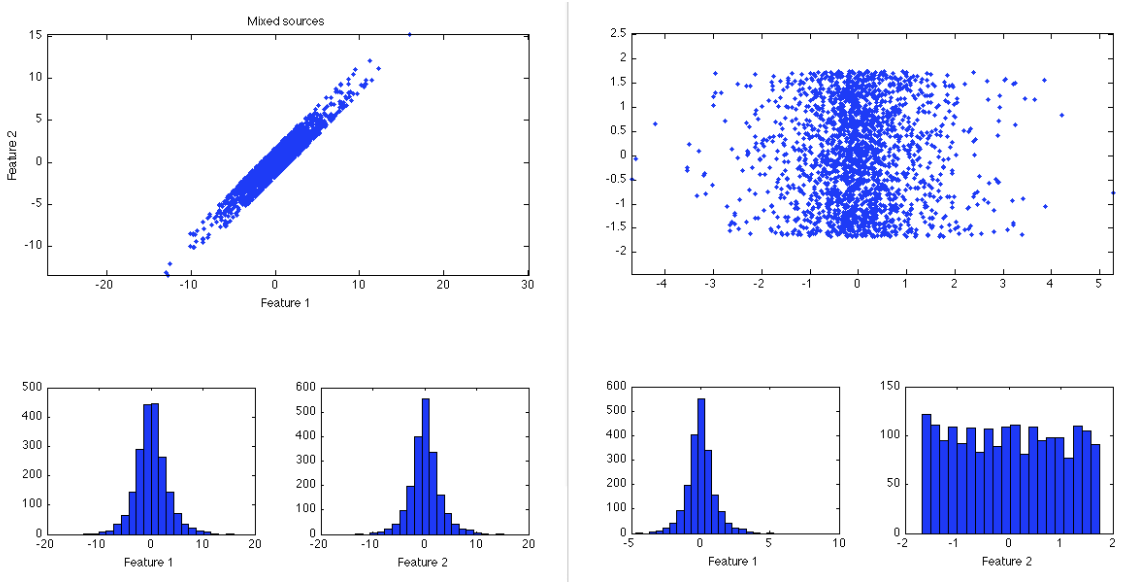

Figure 7: ICA is a technique for separating independent signals that have been additively combined. On the left, we see a two dimensional signal with a high degree of dependence between its dimensions. Below this are projections of this signal onto either of its dimension, both of which seem to have a Gaussian-like distribution. On the right, we see the same signal project onto a new set of axes derived using independent component analysis. Below this are projection onto the new dimensions. It becomes clear that the original signal was an additive mixture of independent signals.

The appropriateness of application of ICA to a set of signals is contingent on several conditions being met. Signals must be:

- additively mixed,

- non-normally distributed, and
- temporally independent.

These conditions are met to varying degrees by EEG activity in general. EEG potentials are known to mix additively based on the physics of the system. Most EEG signals are the result of population activation and may be normally distributed. Activation of spatially correlated brain regions often follow each other in sequence, such as during event related potentials (ERPs) that follow particular stimulation types, so are not expected to be temporally independent. Therefore, ICA is not always appropriately applied in the context of EEG.

However, for the special case of ocular activity, these conditions are largely satisfied. Amplitudes generated by ocular activity are far more peaked, appearing to have a Laplacian or exponential distribution, depending on the type of eye movement. To what degree eye activity is independent of underlying cortical activity is not entirely clear, but there is not an obvious dependence, and accepting the independence assumption produces empirically reasonable results.

Once the ICA operation has been performed, columns of the resulting un-mixing matrix M^{-1} can be viewed as spatially coherent signal generators when displayed as a topographic map. An example of such maps can be seen in Figure 8. Several heuristics regarding the distribution of activity in these maps, the activation time-course of the ICA components, and the distribution of their activations can be serve as a guide in identifying which components are likely related to ocular activity. An example of an activation time-course that is clearly eye-related can be seen in figure 9.

Once the ocular components are recognized by observing the spatial distribution of columns in M^{-1} and the activation time-course of $S = M^{-1}X$, they can be removed. This is done by reconstructing the signal with S and a modified mixing matrix M Let M^* be the matrix M where columns identified as eye-related have been replaced by $\mathbf{0}$. Then the cleaned EEG data $X_{clean} = M^*M^{-1}X$, and the isolated ocular information is $X_{ocular} = (M - M^*)M^{-1}X$.

Current practices for identifying these components rely on a technician to identify which independent components are associated with eye-related activity. The accepted heuristics are not always easily applied and it is often necessary to re-run the ICA decomposition multiple times, or with different subsets of data before a clear ocular component emerges. Furthermore, the instability in the convergence to an un-mixing matrix often leads to poor decompositions that may include brain-related activity in the ostensibly ocular component. Researchers have considered ICA in conjunction with other processing methods, such as adaptive filtering [18], but generally rely on the inclusion of a dedicated EOG recording for reliable removal of ocular components [6, 5]. For these reasons, we were interested in developing an alternative technique that leveraged both the recognizable spatial correlations used in the ICA technique, but also the distinctive time-course of activation that is employed in the component recognition heuristics.

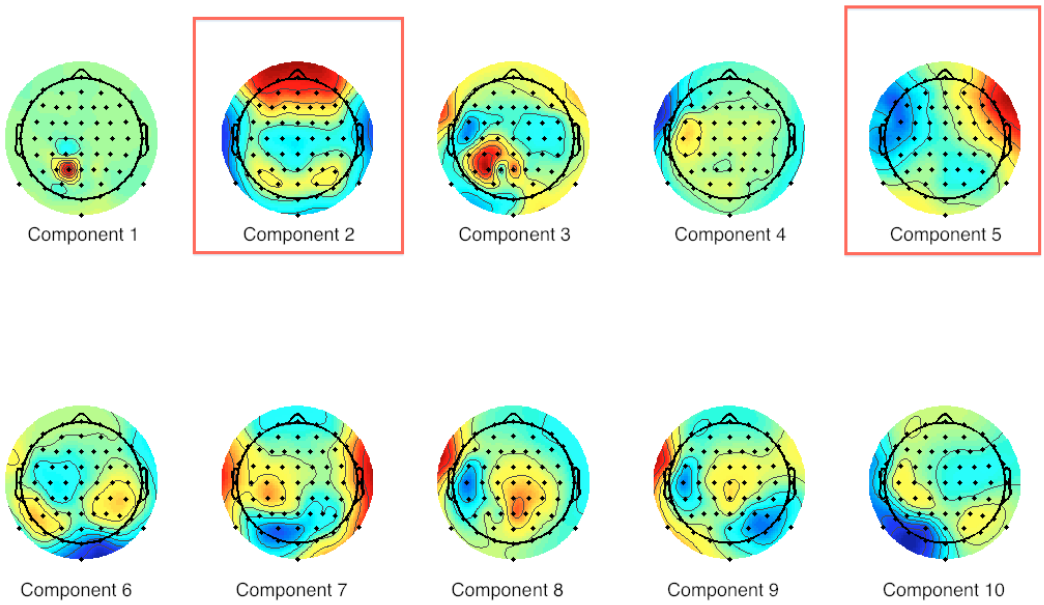

Figure 8: Above are 10 components that result from applying ICA to 64 channel EEG data. The spatial distribution of activation in highlighted components are characteristic of vertical and horizontal eye movements, respectively.

4.2 Template Based Approach

ICA provides an algorithmic approach to removing ocular components from EEG data, but includes manual steps and has unreliable convergence properties. Recognizing these facts, we developed an alternative method that takes advantage of spacial relationships in the EEG data, but is augmented by temporal features of ocular activity

Our new approach is based on *matching pursuit* [26], a method by which a signal is decomposed into a linear expansion through a greedy reconstruction process using an over-complete dictionary. The method has been applied to the analysis of EEG in the past [8, 9], but is not commonly referenced. This is attributable to, at least, the significant runtime required when a large dictionary is used, as is usually the case. We have made two modifications to the existing matching pursuit method. First, we make use of a small, adaptive dictionary that is updated iteratively. This allows us to model and extract the ocular activity; which is largely composed of repeated, conserved transients; without modeling the rest of the signal. Second, we discuss a modification to the method by which the signal is modeled using a lazy update that significantly reduces the number of comparisons between the signal and the dictionary.

Our method for modeling the signal consists of dictionary initialization, and itera-

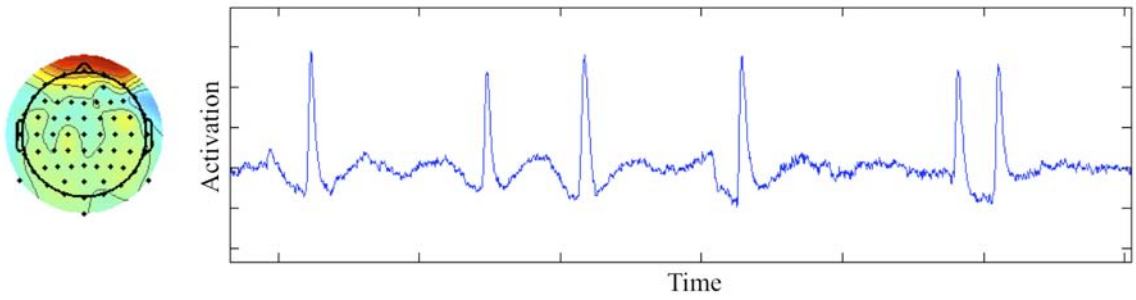

Figure 9: On the left is an example of a scalp map characteristic of a vertical eye-movement component. On the right is the activation time-course of the component. The peaked ocular events are clearly visible, but this information is only used during the manual identification of the component. If the component is discarded, additional information that may be attributed to the component is lost.

tion between signal modeling and dictionary updating. We will model the signal $\mathbf{x} = (x_1, x_2, ..., x_n)$ using a real vector $\mathbf{h} = (h_1, h_2, ..., h_m)$, which will serve as a single dictionary entry, and a real vector $\mathbf{c} = (c_1, c_2, ..., c_n)$, a scaling coefficient sequence.

The model $\mathbf{m}$ of the signal attributed to the dictionary entry can be constructed by $\mathbf{m} = \mathbf{h} * \mathbf{c}$, the convolution of the event template $\mathbf{h}$ with the coefficient sequence $\mathbf{c}$. From the signal $\mathbf{x}$ and this model $\mathbf{m}$, we can generate the residual $\mathbf{r} = \mathbf{x} - \mathbf{m}$, which represents the portion of the signal not attributed to the dictionary entry.

4.2.1 Identifying a Conserved Transient

We initially assume only that the pattern of interest is high-amplitude. For convenience, we will define an m unit segment of a signal x, centered at i, as $N_m(x_i) = (x_{i-\frac{m}{2}}, ..., x_{i+\frac{m}{2}})$, that is, the m unit neighborhood around sample i of x. We initialize our template by averaging signal segments that are centered on local amplitude extrema.

$$W = \{w_i : max(|N_m(x_i)|) = |x_i|\},$$
$$\text{where } w_i = \frac{N_m(x_i) - \mathrm{E}(N_m(x_i))}{var(N_m(x_i))}$$
$$\mathbf{h} = \mathrm{E}(W)$$

This initial template may be based on several high amplitude event types that should not be modeled together, or may contain multiple events captured within a single window, resulting in a poor estimate of the pattern of interest. However, we find that this initial estimate is sufficient to start the iterative updating process.

4.2.2 Greedily Modeling the Signal

To build a model of the signal $\mathbf{x}$, we use the matching pursuit algorithm, which operates on the input signal $\mathbf{x}$ and a dictionary to construct a coefficient sequence $\mathbf{c}$. Here, we are restricting our dictionary to a single template element $\mathbf{h}$. Initially, $c_i = 0$ for all i. We greedily search for a single coefficient using the following update procedure.

$$i^* = \arg\max_i cov(\mathbf{h}, N_m(r_i))$$

$$a^* = \arg\min_a \sum (a\mathbf{h} - N_m(r_{i^*}))^2$$

$$c_{i^*} = a^*$$

$$\text{Update } \mathbf{m} = \mathbf{h} * \mathbf{c} \text{ and } \mathbf{r} = \mathbf{x} - \mathbf{m}$$

Each iteration selects a location, fits a scaling coefficient, and updates the residual to account for the update. These steps are repeated until a termination condition is met. The termination condition may be a limit on the number of non-zero elements of $\mathbf{c}$ or a threshold on a^*.

The method, as described, differs from the common use of matching pursuit in a few small ways. When using a full dictionary, each iteration would find the arg max over all available dictionary elements and would construct a coefficient sequence for each dictionary element used. Additionally, care must be taken in setting the termination condition. The modeled segments of the signal are used to update the template used in the next iteration. If a poor termination condition is employed, and too many signal segments are used to update the template, the resulting template may be degenerate. In practice, a heuristic such as selecting the best matching 5% of detected covariance peaks works well, which led to quick convergence on a blink template after only a few iterations using a few minutes of EEG data.

An alternative termination condition can be used to explicitly avoid a degenerate template situation by evaluating the quality of the modeled signal after each coefficient is assigned. If the remaining variance of a residual $\mathbf{r}_j$, resulting from a new template learned with j modeled events, is less than the variance of $\mathbf{r}_{j+1}$ using $j+1$ modeled events, then the newly included event has essentially degraded the model, and it would be reasonable to stop modeling the signal.

4.2.3 Updating a Transient Template

Having selected coefficients $\mathbf{c}$ that allow us to generate a model $\mathbf{m}$ of our signal $\mathbf{x}$, we update the template $\mathbf{h}$. This update will serve two purposes. First, we would like to account for overlapping signals that are generated by events in close proximity to one another. In the case of eye-blinks, events often occur in rapid succession, creating an artifact in the blink template that corresponds to an immediately preceding and an immediately following

blink. This artifact, and the correction that comes from several iterations of the template updating procedure, can be seen in Figure 10.

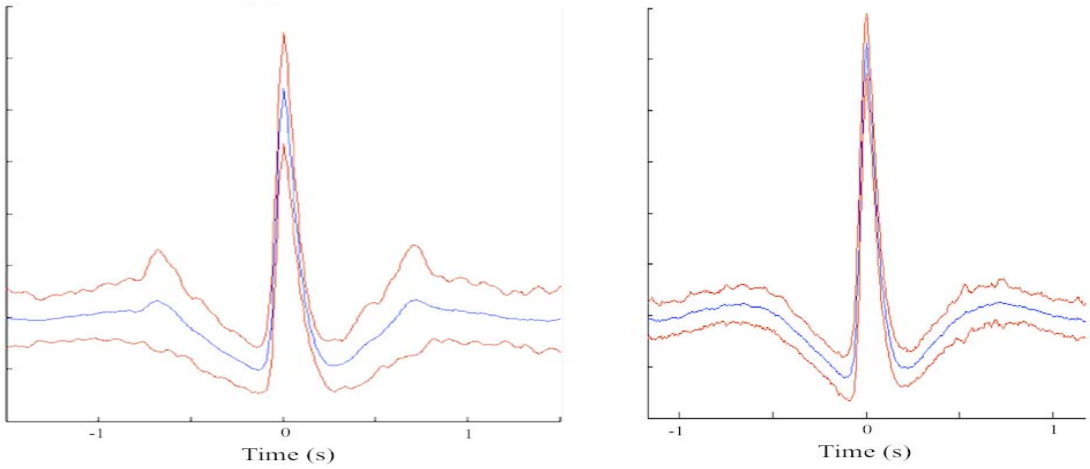

Figure 10: Templates may have artifacts as a result of overlapping transients with a consistent offset in the signal. On the left is an eye-blink template with leading and following artifacts (at approximately .75 seconds preceding or following) due to a consistent time between sequential blinks in the signal. The template itself, the result of taking a point-wise mean of training signal segments, is shown in blue. The red envelope shows one standard deviation from the training data about each point. The template on the right shows the result of applying the template update procedure several times, which removes the peaked artifacts. The remaining depressions before and after the blink event are the result of non-causal high-pass filtering, and are an accurate representation of the blink transient in the input signal provided to the template modeling system.

Second, we would like to account for the case where multiple transient types are present in the signal, producing high amplitude segments that have very different shapes, and should not be accounted for with a single template. Horizontal eye movements present such a situation, as left-to-right and right-to-left eye movements create distinct activation patterns.

We will update $\mathbf{h}$ as follows.

$$\text{Let } \mathbf{r}_i^\circ = \mathbf{x} - \mathbf{h} * (c_1, c_2, ..., c_{i-1}, 0, c_{i+1}, ..., c_n),$$
$$W = \{N_m(\mathbf{r}_i^\circ) : c_i \neq 0\},$$
$$\mathbf{h} = \mathrm{E}(W).$$

$\mathbf{r}_i^\circ$ is an object that shows us our current residual modified such that a single transient remains unmodeled. This allows us approximate the event at i in isolation, in so far as the

current model accounts for the other transients.

4.2.4 Iteration and Termination Conditions

Greedy modeling and template updating are repeated several times until an additional termination condition is met. A threshold on the change in the template from one iteration to the next is a good option.

Once this termination condition is met, we have a final template. A final signal estimate $\mathbf{m}_{final}$ may be built using the same matching pursuit process that is used during each of the update steps. In experiments, we used the simple 5% condition when learning the template, and the more complex condition when generating the final signal model and residual.

4.2.5 Discussion

A benefit of the greedy selection and residual method is that each included coefficient reduces the variance of the residual, and does so less than the previously added coefficient. If the order in which the coefficients are added is stored, then the first j coefficients account for the most signal variance possible using j or fewer coefficients. The method can be interpreted as a simple form of compression, emphasizing an accounting of as much variance as possible with each coefficient.

After several updates, the template becomes increasingly specialized to a particular transient type. If there are several distinct event types in the signal, the final residual $\mathbf{r}_{final} = \mathbf{x} - \mathbf{m}_{final}$ may be used to start the process over, effectively learning new dictionary entries in sequence.

In practice, we found that the entire method can be substantially sped up by replacing the matching pursuit procedure with a lazy procedure that approximates the process. Instead of modeling the segment of the input signal with the single highest correlation to the dictionary template, all local maxima in the local covariance measure are modeled, as long as their neighborhood does not overlap with that of another high-covariance peak. This reliance on local maxima makes it possible that some signal segments with a lower covariance with the template are modeled before signal segments with a higher covariance with the template, but allows the signal to be modeled thoroughly with only a few local covariance calculations, rather than after every template fitting step. This significantly speeds up the modeling process, which is repeated frequently during each iteration of the template updating procedure.

4.3 Modeling Eye-Movement During a Reading Task

To evaluate the degree to which ocular signals were involved in classification performance, we evaluated classification performance using the existing Honeywell processing pipeline, but with several treatments regarding the separation of ocular components. The treatments

included: the full EEG signal with no removal of ocular activity, the estimated ocular activity derived from our method, and the residual left when the ocular activity model is subtracted from the full EEG signal.

To model the ocular activity, we applied the ICA based technique to the EEG signal, identified ocular components based on their topographic maps and activation time-courses, and modeled the activation time-course using our modified matching pursuit method.

By applying the estimated mixing matrix to the modeled ocular signal, we shift from independent component space back to electrode space. This allows us to simply subtract the estimated ocular signal from the original EEG signal to get an estimate of the cortical signal without the ocular component. An example of a signal segment, the resulting modeled signal, and the resulting residual signal estimate can be see in Figure 11.

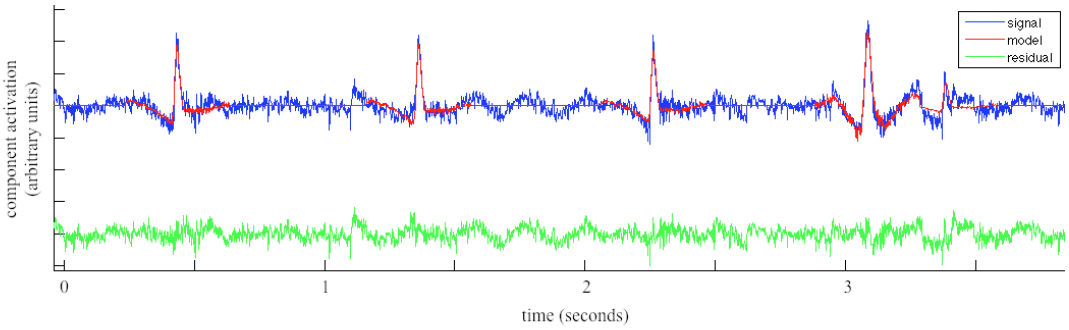

Figure 11: The original signal contains clear ocular activations throughout. The estimated ocular contribution appears to model this activation well, with little noise, but a clear, high-frequency peak. The resulting residual demonstrates that additional artifacts are not introduced, and transients that do not appear to be eye-related remain unmodified.

Each of these three treatments were then decomposed into frequency features and classified as described in Section 1.7. Figure 12 shows the classification AUROC values for each signal treatment and for each subject.

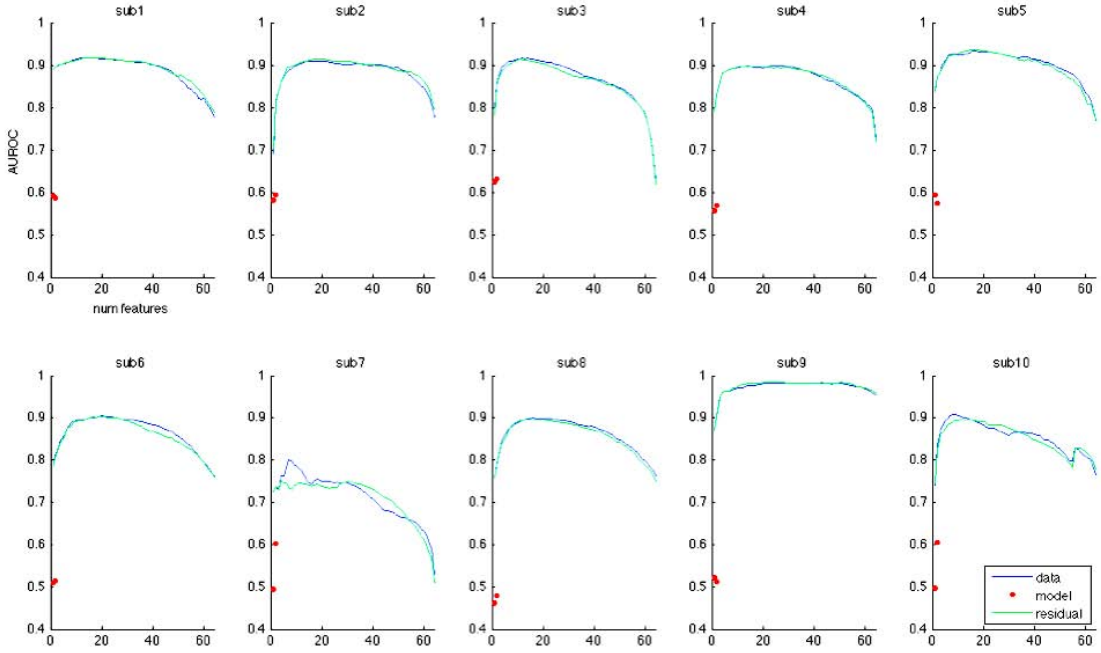

Figure 12: The classification for each subject and each preparation is shown. The x-axis shows the number of electrodes included to achieve the corresponding AUROC value on the y-axis. As there are only single channels for the ocular signals, they are placed on the x-axis arbitrarily. We see that they peak AUROC using the full EEG signal and the peak AUROC using the cleaned EEG signal are nearly equal for all subjects, supporting the claim that cortical activity explains the classification performance for both. Further, we see that the AUROC values associated with ocular components are low, usually below .6.

5 Feature Extraction and Classification

Since its development and its first application to humans, EEG signals have been interpreted in terms of their frequency composition, with alpha waves (in the 8-13 Hz range) being the first named frequency band by Hans Berger. Even these early observations related activity in particular frequency bands to brain state and cognitive activities such as mental arithmetic [34]. EEG signals are made up of both these frequency components and transients, and to what degree either of these are of interest impacts what decomposition of the signal is most appropriate. Frequency components can be extracted using traditional signal processing methods, such as Welch's method, as well as through wavelet transformations, which are more effective in modeling transient components. For our data, we don't expect transient signals to play a significant role in estimating cognitive load, as cognitive effort is associated with the sustained use of processing resources during the performance

of a task. For this reason, we've relied on frequency features without transients.

5.1 Feature Extraction

We considered two feature sets in our analysis of the EEG data gathered during the listening task. The first was based on the features from the Honeywell study described in section 1.7, which will be refered to as the spectrographic features. These features measure the average frequency power in clinical EEG bands, during a five second window, at a particular scalp location. To extract these features, we high-pass filter signals at 1 Hz to remove drift and slow changes in potential, reject unreliable locations and signal segments, and remove the estimated signal contribution related to ocular activity. We then extract spectrographic information using a short-time Fourier transform, sampling every 2.5 seconds and using a 5 second wide Hamming window. This brings the data from the time-domain to the frequency-domain, where average power in each band of interest can be calculated with a simple mean calculation. The clinical bands of interest included:

- Delta band (1-4Hz)
- Theta band (4-8Hz)
- Alpha band (8-13Hz)
- Beta band (13-30Hz) and
- Gamma band (30-50Hz).

The resulting feature space can then be indexed by time, location, and frequency.

It can be instructive to view the differences in activity based on task condition, although it may not be indicative of which features will be most discriminative on a single trial, user-specific basis. Figures 13 and 14 contain scalp maps that compare these power differences averaged over control subjects and aphasia subjects, respectively. These full scalp frequency maps, averaged over subject groups, are the type of features that could form the basis for cognitive load estimation that uses group norms. We see that the d' values associated with the differences in power distributions are quite low, but higher for control subjects, supporting the notion that subjects with aphasia will not be as effectively monitored on the basis of group norms. However, these types of features were shown to be effective in estimating cognitive load in previous studies on a subject-specific basis [27].

We were concerned that particular features of the signal, that may be indicative of load, were being smoothed over by using average power. In particular, it seemed that short bursts of activity in the alpha band may be related to the level of engagement of subjects while performing a cognitive task. This motivated consideration of the second feature set, one based on a more complex description of the activity within a particular 5 second window. To identify bursts of activity, we instead described the activity in a

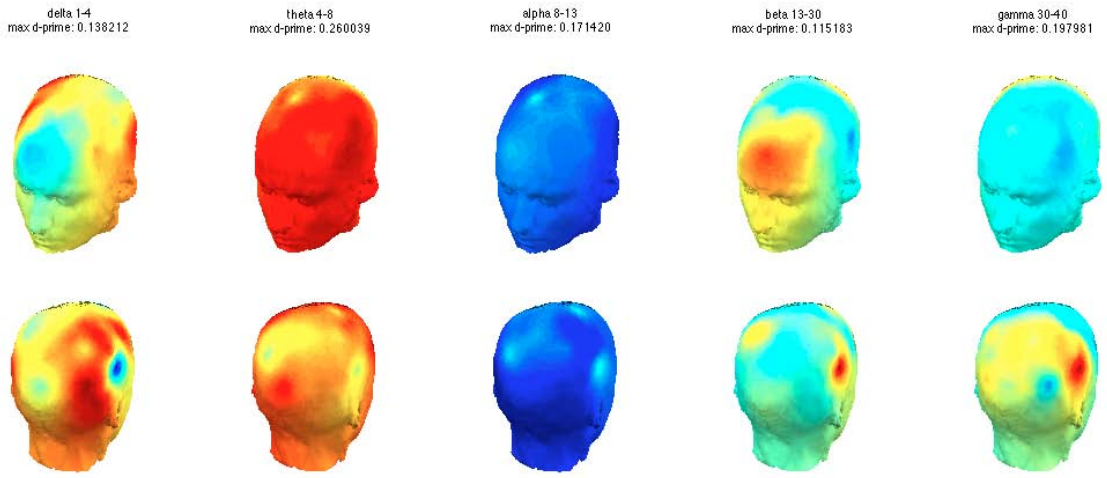

Figure 13: Above are scalp maps (viewed from two angles) that display the gross differences in activity between difficulty conditions, averaged over control subjects. Power was averaged over control subjects for each scalp location, each band, and each condition. The difference between the average activity associated with each condition is shown, scaled to fill the available color range. Each band is labeled with a d-prime value, indicating the extent of the difference between activity patterns. We see the most significant differences for control subjects appear in the theta and alpha bands.

window using a statistical summary of the activity. We initially chose to represent the activity using box-plot percentiles, storing the $2^{nd}, 25^{th}, 50^{th}, 75^{th}$, and 98^{th} percentiles of activation within the 5 second period. The intuition was that, although the 50^{th} percentile might not distinguish between a sampling window that contains short bursts of activity, or spindles, and one that does not, the 75^{th} or 98^{th} may show differences. We extracted these features using the Hilbert Transform [15], a transformation that interprets the recorded EEG signal as the real portion of an analytic signal. This transformation allows for the calculation of an amplitude envelope for the signal within a particular frequency band. This envelope can then be used to view frequency power with high temporal resolution. By filtering our EEG signals using a bandpass filter for a frequency band of interest, applying the Hilbert Transform, and by considering 5 second windows of activity, we were able to extract the summary statistics. These features will be referred to as Hilbert features. An example of the relationship between a signal and the Hilbert features can be seen in Figure 15.

For several different classifier types, we found that data represented using the Hilbert features could be nearly perfectly separated for a set of training data, but that the model that allowed for this near perfect separation would generalize very poorly (in fact, near

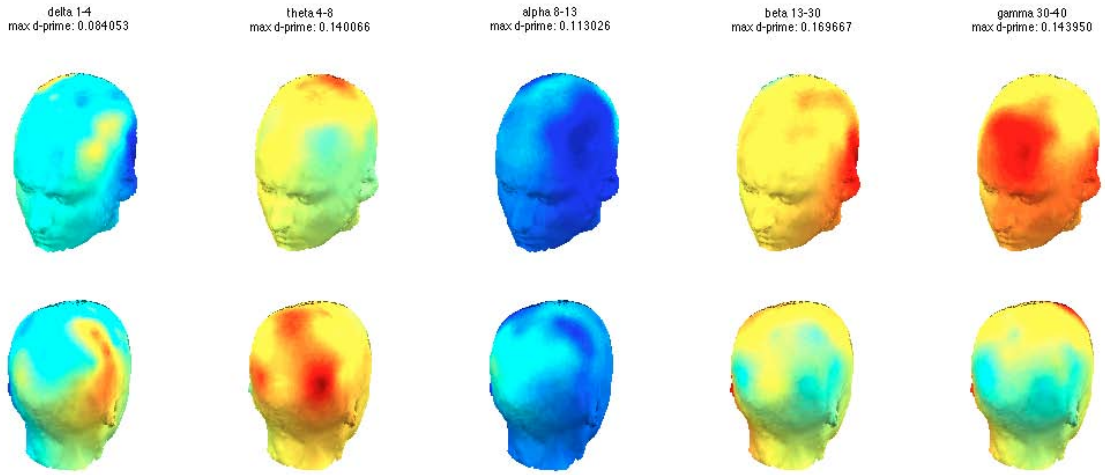

Figure 14: Above are scalp maps that display the gross differences in activity between difficulty conditions, averaged over aphasia subjects. Power was averaged over aphasia subjects for each scalp location, each band, and each condition. The difference between the average activity associated with each condition is shown, scaled to fill the available color range. Each band is labeled with a d' value, indicating the extent of the difference between activity patterns. We see a more generalized difference in gamma band activity than was seen for the controls.

the guess rate for the task), a pattern characteristic of over-fitting. In these cases, systems based on the spectrographic features did not exhibit the same perfect separation of training data, nor the poor performance on testing data. We took several steps to address the over fitting issue. First, we reduced the number of summary statistics used in the Hilbert feature set, from 5 to 3. This set exhibited the same over-fitting pattern seen using the original Hilbert feature set. Second, we employed the SFFS feature selection process to select a subset of scalp locations to be considered for classification. This method dramatically reduced the dimensionality of the feature space, and was able to achieve performance that indicated that over-fitting was not creating a substantial problem, with a mean AUROC of .74 over 10 subjects, with a standard deviation of .04. However, this did not represent an improvement over spectrographic features, for which the mean AUROC was .78. Furthermore, on an individual subject basis, the spectrographic features were consistently associated with better performance than the Hilbert features, as can be seen in Table 2.

This led us to conclude that the added dimensionality of the Hilbert feature set contributed to the difficulty of the feature selection and classification processes more than justified the benefit of encapsulating more complex signal features. Features such as Alpha spindles may still be useful when specifically coded for, but the potential variability

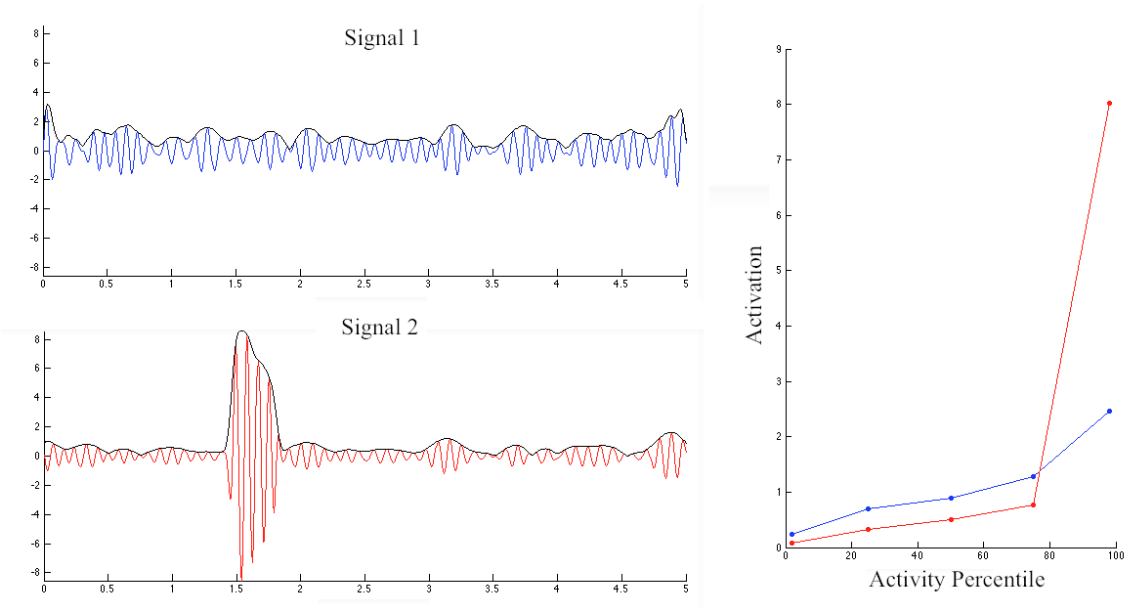

Figure 15: Features based on the Hilbert Transform make it possible to distinguish between signals that have the same average power in a window, but for which the distribution of activity is markedly different. The left side of the figure shows two synthetic signals with equal mean power in the Alpha band. The first signal (in blue) has a relatively uniform distribution of power over the 5 second window, where the second signal (in red) has a strong burst of activity for approximately .5 seconds. The Hilbert Transform was applied to each (with the resulting amplitude envelopes shown in black) and activation percentiles were calculated. The difference in activity for the $2^{nd}, 25^{th}, 50^{th}, 75^{th}$, and 98^{th} percentiles for each signal are shown in the right side of the figure. Notice that the 2^{nd} through 75^{th} percentiles show little difference in our summary statistic feature space, but the 98^{th} percentile shows a clear difference between the activity patterns of the two signals. This is the additional information we hoped to leverage by using the Hilbert transform.

Subject, Session	Spectrographic Features AUROC	Hilbert Features AUROC
1,1	0.74	0.70
1,2	0.75	0.68
2,1	0.87	0.82
2,2	0.82	0.76
3,1	0.77	0.75
3,2	0.81	0.80
4,1	0.73	0.70
4,2	0.76	0.74
...		
mean	.78	.74

Table 2: Spectrographic features consistently out-performed Hilbert features on a per-subject, per-session basis, when used as part of a processing pipeline that included the SFFS method for feature selection. When feature selection was not included, Hilbert features consistently led to over-fitting during training, and generalization performance near the guess rate.

between subjects makes us hesitant to stray from the machine learning based approach for which we've advocated in identifying important features.

5.2 Classifier

The system, as described in Section 1.7, used logistic regression for classification. Logistic regression is a method that assumes that the relationship between instances from a data set and their label are best represented in a probabilistic fashion. Some instances will be high likelihood examples of a particular label, others will be high likelihood instances of the opposite label, and some will be ambiguous. This variability in likelihood of belonging to one class or another is expressed with a logistic function that relates a single variable, the *decision variable*, to the likelihood of class membership. Data is associated with this decision variable through a linear combination of continuous-valued features of the instance. Put together,the probability of a particular label L is given by

$$P(L|X_i) = \frac{1}{1 + e^{-(B_0 + \sum_{j=1}^{n} B_i X_{ij})}}$$

for instance X_i and coefficients B.

A model is fit to a data set by identifying a linear combination of features that projects the data onto a single dimension in such a fashion as to maximize the agreement between

the actual labels for the known data and the predicted labels resulting from the model.

$$\max_{B} \prod_{i} P(L|X_i).$$

Logistic regression was a good initial choice for a classifier because it can characterize data from a variety of distributions, and is also not highly sensitive to outlier data, mislabeled training instances, or spurious statistics (compared to SVMs, need a good reference for this).

Fitting a logistic regression model to data can be somewhat time consuming, as it relies on Maximum Likelihood Estimation (MLE), a method that uses an iterated optimization approach such as Newton's Method. This is potentially troublesome, as some feature selection methods require that models be applied to data repeatedly to build a system for any particular subject. However, our method for evaluating the quality of a combination of extracted features, classifier, and selected features has consistently been the AUROC. In computing an AUROC, predicted classes for testing data are considered with all possible threshold values, which means that any monotonic transformation of the projection from data to decision variable will have no effect on the final AUROC. This allows us to disregard the logistic function that wraps the linear combination of features, $\frac{1}{1+e^{-z}}$, as well as the bias term, B_0, without changing the measured quality of the resulting classification system.

Linear Discriminant Analysis (LDA) can provide a similar projection from feature space onto a decision variable, although it is not appropriately applied to as wide a variety of data distributions. LDA identifies the optimal separating projection using the difference of means of data with a particular label, and the inverse of a covariance matrix. This is appropriate only when a normality condition and a variance condition are met by the data sets. In practice, we frequently used LDA, rather than Logistic Regression, with little or no degradation in performance, and a dramatic improvement in the runtime require to fit a model to data. Although requiring further investigation, it appears that the LDA computation may serve as an appropriate initialization state for the weights used in logistic regression. Essentially, a normality assumption, appropriate or not, is used to initialize the process of maximum likelihood estimation. In performing the optimization used for maximum likelihood estimation, the normality assumptions are no longer considered, allowing for the projection to correct for error associated with the assumptions. The projection found through LDA is often quite close to that found through pure MLE, making the final optimization very short, and speeding up logistic regression substantially.

5.3 Feature Selection

Although both logistic regression and LDA can perform well on our EEG data, they often do so only when combined with a feature selection method. Our primary method for feature selection was sequential forward floating search (SFFS), a heuristic by which a useful subset of available features is constructed by adding and culling features. SFFS relies on two main

data structures: a pool of available features and a working list of good features (which is initialized as an empty list). The process of adding to the working list first appears to be greedy. Each individual feature is evaluated based on how well a classifier is able to predict data labels on the basis of that feature alone. The feature associated with the best performance is added to the working list, and this working list is stored as the best known list of length 1. Next, all subsets made up of the current working list, and one additional feature from the pool, are evaluated. The subset that performs best is stored as the best known list of length 2. This greedy mechanism does not continue on uninterrupted. After each new feature is added to the working list, producing the best seen list of length n, feature sets of length $n-1$ are considered by withholding a single element and reevaluating. If a feature list of length $n-1$ is found that outperforms the previously best-performing list of length $n-1$, the culling process will continue back, next considering lists of length $n-2$ [37]. This process makes it possible to discover feature sets that outperform greedily constructed sets without reverting to an exhaustive search. However, this method is both a heuristic, meaning there are no guarantees about the quality of sets that result, and often very slow, due to the unpredictable number of culling steps.

In applying the SFFS algorithm to our data, the pool of available features was made up of electrode locations. Although each electrode location is associated with multiple frequency features, we considered this to be a reasonable decision as the frequency specific activations at a particular location are expected to form a sort of functional group. The evaluation of a subset of electrodes requires that a classifier be trained and evaluated using the feature set in question. In each situation for which SFFS was considered, the classifier used for evaluation matched the classifier being used for the pipeline under consideration. The quality of feature sets were compared on the basis of an AUROC value, the same metric by which we compared full processing and classification pipelines.

There are several drawbacks to relying on the SFFS heuristic. The method is essentially a walk through the space of possible feature combinations without any guarantees regarding the quality of the results, the time it will take to arrive at solutions, or a clear interpretation of the results for any particular domain. The difficulty in interpretation is especially significant when considering EEG data, as our manipulations of brain state are noisy. We would like to be able to support the reliability of the output of the machine learning methods by connecting them to prior expectations from literature. Although the relationship between the signal source and brain state may be consistent across sessions or even subjects, it is unlikely to be reflected in the selection of electrodes by the SFFS algorithm. As discussed earlier, EEG data relates meaningful signal sources to electrode recording in a diffuse manner, with several electrodes providing a measure of the same underlying source, and some underlying sources never accounting for the majority of the signal variation at any one electrode. Several electrodes may work in conjunction to characterize a particular signal source, but the SFFS algorithm would generally select only one, ignoring the others as essentially redundant information.

5.4 Alternative Classifiers and Feature Selections

To address concerns regarding over-fitting and feature selection, we considered several alternative classifiers that we hoped would constrain the dimensionality of the problem appropriately, or have an element of feature selection built into the model fitting process.

5.4.1 Hierarchical Logistic Regression

Our first attempt at an alternative classifier involved building an ensemble of weak learners that would attempt to characterize cognitive state with access to a limited subset of the available features, be those Hilbert features or spectrographic features. The collection of weak learners would then be integrated using a final classifier that would use the weak learner outputs as its own inputs. The imposed feature restriction was based on scalp topography, allowing each weak learner to make a decision based only on information gathered from a single electrode. In the case of the Hilbert features, this allowed each weak learner access to a vector of 25 features (5 frequency bands and 5 measures of activity) for each 5 second window of data. These learners were implemented with a logistic regressor, which outputs a probability of elevated cognitive load. These probabilities were then provided to a final logistic regressor that produces its own estimate of cognitive state. A diagram of this classifier can be seen in figure 16. Unfortunately, this division of the data did not provide improved classification results when using spectrographic or Hilbert features. Here, training data could be discriminated with a high degree of accuracy, but generalization failed with mean AUROC values of .55 and .54, respectively.

The site-specific, weak learners did not seem to have access to sufficiently rich information to make an estimate of cognitive load, providing uniformly poor classification performance, even on their training data; while the integrating classifier seemed to perform very highly on the training data and poorly on the testing data. This suggested that the information available at each layer should be shifted, with more information available to the lower layers, and fewer variables available at the higher. We then constructed a band-first hierarchical regressor, where the weak learners had access to data from all electrode sites and all activation percentiles, but only for a single band, as depicted in Figure 17. This reconfiguration did not perform well either, with over-fitting occurring in the first set of learners.

5.4.2 Artificial Neural Network

The division of features by band or by location seemed reasonable, but the reality may be that integrating over near-by locations, disparate locations, opposing frequency activations, or any number of other combinations may be much more indicative of cognitive load.

We wanted to continue investigating a hierarchical classifier structure as an alternative to distinct classifier and feature selection procedures. The hierarchical logistic regressor described above is an example of an artificial neural network (ANN), but one where the

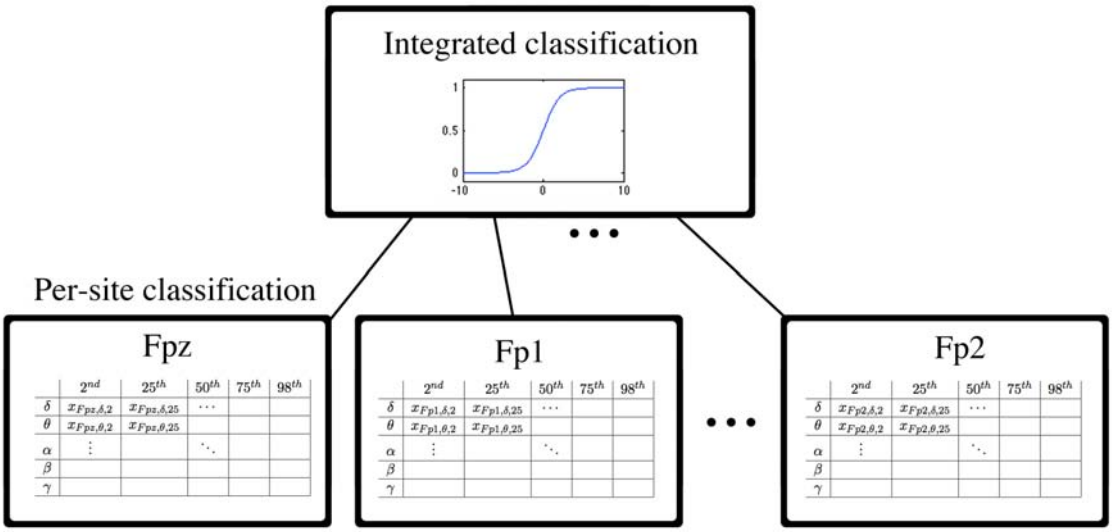

Figure 16: A hierarchical logistic regressor was trained, for which there were two layers. Nodes in the first layer represent logistic regressors that estimate the difficulty condition using all bands and activation percentiles, but for only a single scalp location (such as Fpz, Fp1, and Fp2, as shown). The second layer classifier integrates the outputs of each of these individual classifications, again using logistic regression. The goal of this restricted connectivity is to prevent any one classifier from having access to too much information, which can lead to over-fitting. In this case, the first layer classifiers had access to 5 bands and 5 percentiles, 25 total features for each window of EEG data. The second layer classifier had access to the outputs of the 64 site classifiers.

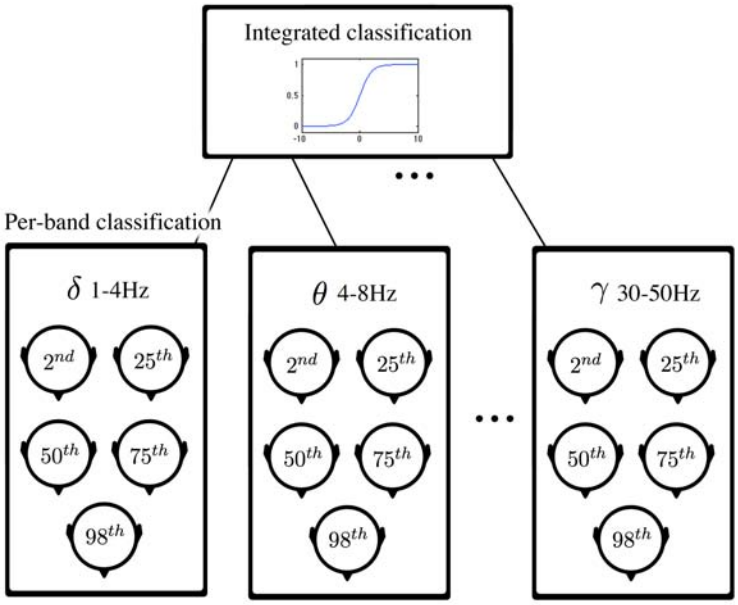

Figure 17: The site-first classifier showed low performance from the first layer, and over fit at the second. This suggested that more features should be made available to the first-layer classifiers, and fewer should be made available to the second layer. The above diagram depicts a band-first hierarchical regressor, for which first layer nodes have access to all 64 sites and 5 activation percentiles for a single frequency band, a total of 320 features. The second layer node has access to the outputs of the 5 band-specific classifiers.

training procedure was based on maximum likelihood estimation, and where the connectivity between the low-level learners and the high level learners was artificially imposed. In general, ANNs do not have these connectivity restrictions and are trained using back propagation, a gradient descent method that operates on error rates rather than probabilities and is not applied to low-level learners and high-level learners in isolation, as it was for the hierarchical logistic regressor. The net effect of these differences is that an ANN has the theoretical potential to emphasize particular features, consider many combinations of features, and essentially perform its own feature selection.

Our initial experiments using an ANN produced poor results, with AUROC values near the guess rate using Hilbert and spectrographic features. It is less clear that this was a case of over-fitting, as the discrimination between classes of training data did not reach the high-level seen using other methods. There are additional complications involved in employing back-propagation for training ANNs, as these networks have many parameters and can become stuck in local optima. This makes the choice of learning parameters and starting state important, and makes it less clear that something like over-fitting to the training data might be involved with this classifier. We manually tuned these parameters based on empirical results gathered from an unrelated classification problem of similar dimensionality. The ANN achieved good results with this practice problem, giving us confidence that our choice of learning and connectivity parameters was reasonable. Given this confidence, and the poor results on the brain state estimation problem, we did not further pursue an ANN-based classification approach.

5.4.3 Principal Components Analysis

Finally, we employed principal components analysis (PCA) in several ways. PCA is a method by which data can be transformed from an existing basis to a de-correlated, orthogonal basis using an eigendecomposition. The resulting features are usually arranged in order of the magnitude of their associated eigenvalues. For EEG data, this means that data is no longer indexed by scalp location, but by the components contribution to the total signal variance. The resulting ordering can be used in several ways. First, it can be used as a method for dimensionality reduction. Although the approach does not take data labels into account, it is often reasonable to assume that PCA components that account for a significant portion of signal variance are more appropriate for use in classifying signals than components that account for little. Consequently, using some number of the PCA components, in order of their eigenvalues, is generally a reasonable approach to feature selection. Second, PCA can be used for de-noising data before further processing. A user can transform their data to PCA space, suppress some set of features that are low variance (and assumed to be noise rather than signal), and then return the data to the original space. For our data, we employed the first strategy with some success using several combinations of feature sets and classifiers. However, these results never outperformed results from the same system using SFFS instead of PCA.

6 Results and Discussion

After performing analyses using several options for the components that make up a machine learning-based processing pipeline, we have largely confirmed that the initial implementation decisions that led to the cognitive load estimation system described in Mathan et al. are not immediately improved upon using our selected alternatives [27]. While better alternatives may exist, finding them will require additional data and an expansion of the considered options. In particular, the problem of over-fitting, which seemed to be a frequent cause of poor classification performance, may be addressed through a more explicit model of underlying processes. A model can provide a constraint on the dimensionality of the features that are provided to a classifier, addressing the issue of over-fitting, while encapsulating the most pertinent information. In terms of the machine learning approach that we've taken, this means using methods that identify useful feature transformations in a data driven manner.

6.1 Cognitive Load Estimation using Nominal Difficulty Labels

We now consider the result of applying our cognitive load estimation system to the new subject population and manipulation described in Section 2. Recall that our subjects had more pronounced deficits than in previous studies, where subjects were either healthy [2] or suffered from mild cognitive impairments [27]. Furthermore, recall that our manipulation of cognitive processing requirements was more subtle than in past studies, requiring listening and understanding, rather than the intense engagement produced by some traditional neurocognitive evaluation tasks [2] or by tasks that require active engagement under time-pressure.

Based on our previous investigations, we selected a cognitive load estimation system that consisted of the following components:

- high-pass filtered at 1Hz using a 4^{th}-order Butterworth filter
- myoelectric-contaminated segments removed based on outlier detection with respect to activity over 125Hz
- down-sampled the signal to 128Hz
- identified unreliable electrode locations based on frequent, extreme signal amplitudes
- removed ocular artifacts using the template-based approach previously described
- gathered average frequency power estimates from clinical bands using a 5 second Hamming window and a 50% overlap
- performed feature selection using SFFS with an LDA classifier and the AUROC for feature set evaluation

- performed final classification using logistic regression
- evaluated the system using 10-fold cross validation with folds consisting of contiguous data.

EEG data recorded while the subject was listening to a passage was processed by the system. The resulting predictor variables were derived from models trained for individual subjects, which were then aggregated over subject condition groups. Figure 18 shows the distribution of these predictor variables for each group, and for each nominal difficulty labeling. When presented with nominally difficult passages, the median estimated probability of elevated load is approximately .8 for control subjects, compared to approximately .2 when the same subjects are presented with passages of nominally low difficulty. For subjects with aphasia, the median probabilities for difficult and easy passages were approximately .95 and .05, respectively. For control subjects, this separation leads to a mean area under the ROC curve of .71, and for aphasia subjects, a mean area under the ROC curve of .72.

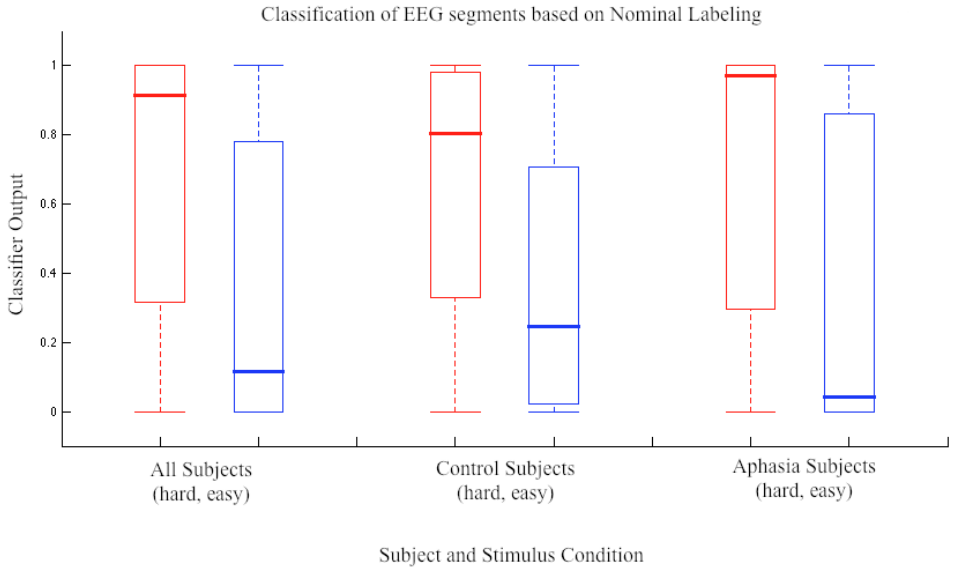

Figure 18: Above are box plots showing the relationship between the original stimulus manipulation and the logistic regressor outputs of the cognitive load estimation system. The classifier output is the predicted probability that the subject was under elevated load while performing the task. Each box plot shows the distribution of this probability prediction for a particular stimulus type (signified by color, with red for hard and blue for easy) for a particular subject group (shown in the x-axis label).

Although not as high of an AUROC value as was seen in the previous study [27], it does indicate the successful classification of high load activities and low load activities for both controls and aphasia subjects, albeit a noisy classification. Additionally, we see in Figure 18 that the separation of the median values is higher for aphasia subjects than for control subjects, despite a large overlap in distributions. This pattern would be consistent with an accurate classifier, but a significant number of mislabeled examples. This general pattern continues when considering individual subjects, as can be seen in the appendix.

6.2 Subject Responses

There are several issues that might, in effect, lead to a mislabeling of data gathered during trials. As discussed in Section 2, the experience of individual subjects for particular questions and for particular sessions may vary greatly, and the reported experience of subjects seems to support that this difference was more pronounced for aphasia subjects than for control subjects. Figure 19 shows the subjective assessment of passage difficulty divided by nominal difficulty labeling and subject group. Response values were normalized on a per-subject basis to fall into the continuous range [0,1], where 0 indicates low difficulty and 1 indicates high difficulty. Notice that many aphasia subjects reported that nominally easy passages were difficult for them to understand. It rarely occurred that nominally easy passages were reported as high difficulty by control subjects, who labeled approximately 25% of nominally difficult passages as easy. This pattern continues when responses are considered on a subject-subject basis, as can be seen in the appendix.

Additionally, the nature of our manipulation was subtle and may have contributed to the effective mislabeling of data. In previous experiments involving a continuous measure of cognitive load, the manipulations incurred a sustained elevation in cognitive workload, such as that created by the n-back [2], or a reading task with a time constraint [27]. Our listening task was constructed to more closely simulate a real world task. As such, it is likely to be the case that, even for passages that a subject later assessed to be difficult, much of the time listening was spent in a relaxed state. Because passages are labeled as "all-easy" or "all-hard", this essentially mislabels many sub-segments of any particular passage. This effect was plain during data collection, when subjects would occasionally note that they had essentially lost interest during a passage, but then reported that the passage had been difficult to understand. This is a reasonable response on the part of the subject, as a textual passage that has become difficult to follow becomes much less interesting. However, it also indicates that some of the EEG recordings were labeled as having been recorded during a high difficulty task, but were recored while the subject was in a relaxed state.

To investigate the of subject-to-subject differences in the reported difficulty of passages, we used a modified labeling on passages that was based on the reported difficulty provided by the individual subjects, as discussed in Section 2. Subjects were asked several questions after each passage, including how confident they were in their response to the

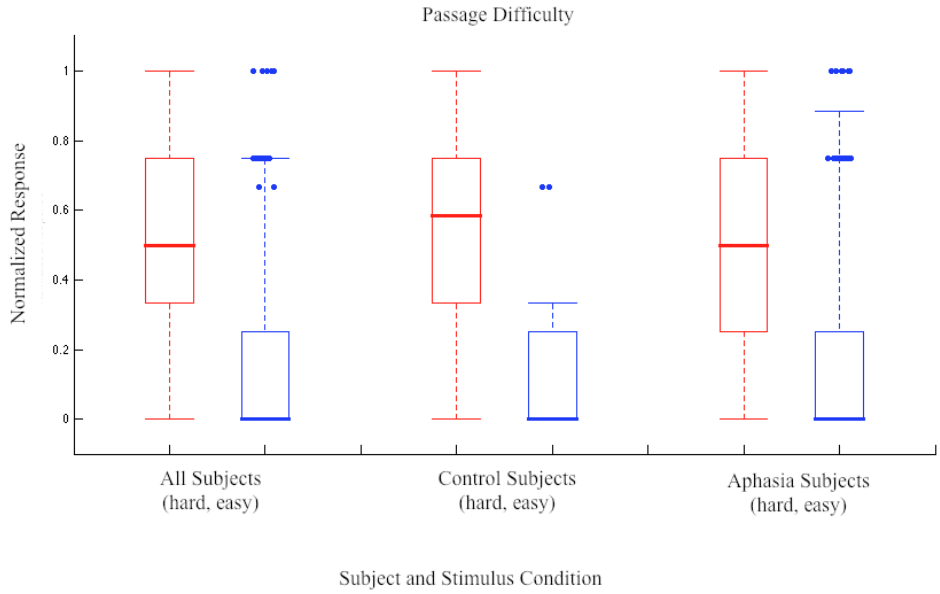

Figure 19: Above are box plots showing the relationship between the original stimulus manipulation and the subjectively reported difficulty of passages provided by subjects during the experiment. Responses were normalized to the continuous range [0,1] to adjust for different usages of the available scale. For example, some subjects reported using the full 1-5 scale explained at the start of the experiment, where others may not have reported a difficulty over 3. Notice that both controls and aphasia subjects disagree with the nominal passage labels, but usually in opposite directions from each other. Data that is further than 1.5 times inter-quartile range above the 75^{th} percentile, or below the 25^{th} percentile, are included individually (with an arbitrary horizontal offset).

comprehension question, how well they understood the passage, and how difficult they felt the passage was to understand. Subjects used these responses differently from each other, both in terms of which question seemed to draw out their experience of the task, as well as in terms of how they used the range of response values. Our decision to use a projection from these response values to a single subjective difficulty score is discussed in Section 2.

The relationship between this modified difficulty estimate and the original passage labeling can be seen in Figure 20. Under the assumption that subjects have accurately reported their experience of the passage, and that the perceived difficulty is indicative of the neurological correlates of mental workload, then this level of separation represents something of an upper limit on the quality of a cognitive load estimator trained on and predicting the nominal difficulty labels in our data. The distribution of these projected response values is clearly a good indication of the intended difficulty of the passage, and

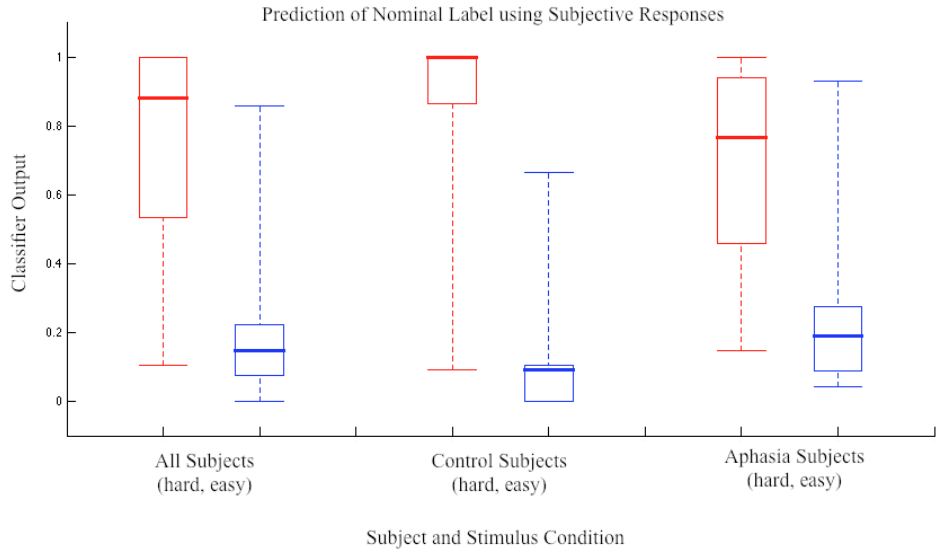

Figure 20: Above are box plots showing the relationship between the original stimulus manipulation and the predicted stimulus difficulty based on responses provided by subjects. When compared to the distributions in Figure 22, we see that these distributions are more tightly distributed about their median values. However, there is still substantial overlap in prediction values between nominally easy and nominally difficult tasks.

not unexpectedly, the projection that generated these distributions often heavily weighted the subjective difficulty score. However, we also see that the project is more effective in separating nominally easy and nominally difficult passages for control subjects. We hypothesize that, even when all responses are considered, aphasia subjects had a difficult time with some of the nominally easy passages and did not recognize the difficulty of some of the nominally difficult passages. This interpretation is supported by the actual response accuracy of subjects during testing. Table 3 shows that both control and aphasia subjects missed approximately 30% of the questions that followed nominally difficult passages. Controls missed only 6% of the questions that followed easy passages, while aphasia subjects missed approximately 16% of these questions. Control subjects were able to identify passages that were associated with incorrect responses, while aphasia subjects seemed to label these difficult passages, which led to response errors, as easy to understand. To further support this notion, it would be useful to go back to the data and perform an additional analysis on the perceived difficulty of passages that in-fact led to a response error, but we expect no clear story from these performance results. Note that two of the three highest performers on easy questions were aphasia subjects (6 and 11), as was the top performer on difficult questions (subject 3).

Subject, condition	difficult	easy	combined
1,c	0.844	0.975	0.917
2,c	0.625	0.925	0.792
3,a	0.875	0.875	0.875
4,c	0.750	0.925	0.847
5,c	0.656	0.947	0.814
6,a	0.844	0.950	0.903
7,a	0.594	0.750	0.681
8,a	0.562	0.900	0.750
9,a	0.469	0.816	0.657
10,a	0.844	0.900	0.875
11,a	0.750	0.950	0.861
12,a	0.688	0.850	0.778
all	0.708 (0.131)	0.897 (0.065)	0.813 (0.084)
control	0.719 (0.099)	0.943 (0.024)	0.843 (0.055)
aphasia	0.703 (0.150)	0.874 (0.068)	0.797 (0.095)

Table 3: The above table shows the accuracy of responses to difficult, easy, or all comprehension questions for individual subjects and for groups (control subjects, aphasia subjects, and over all subjects). The difference in the accuracy of responses following difficult questions and easy questions is pronounced for both control subjects and aphasia subjects. The difference in response accuracy between controls subjects and aphasia subjects seems insignificant, although the difference in accuracy for easy questions appears to be more so. However, many aphasia subjects performed as well or better than controls.

Having gathered our alternative difficulty labeling from subjective responses, we went back to the EEG-based measure of cognitive load to see if this alternative labeling resulted in an improvement in classifier performance. If it was the case that disagreement between the nominal labels and the perceived difficulty led to mislabeled data, then a classifier that was built on the self reported difficulty may be more effective in separating the data. The construction of the cognitive load estimation system consisted of the same parts described above. The distribution of predictor variables provided by the resulting classifier can be seen in Figure 21, or in the appendix for individual subjects.

We do not see a clear improvement from this relabeling, with similar distributions of classifier outputs to those seen using the nominal labels. This similarity was consistent for control and aphasia subjects, and for passages that were rated as high or low difficulty. This suggests that the effect of mislabeling whole passages was not the primary reason for low performance, relative to previous studies. This suggests that much of the data remains essentially mislabeled due to the variability of effort required during a task. If correct,

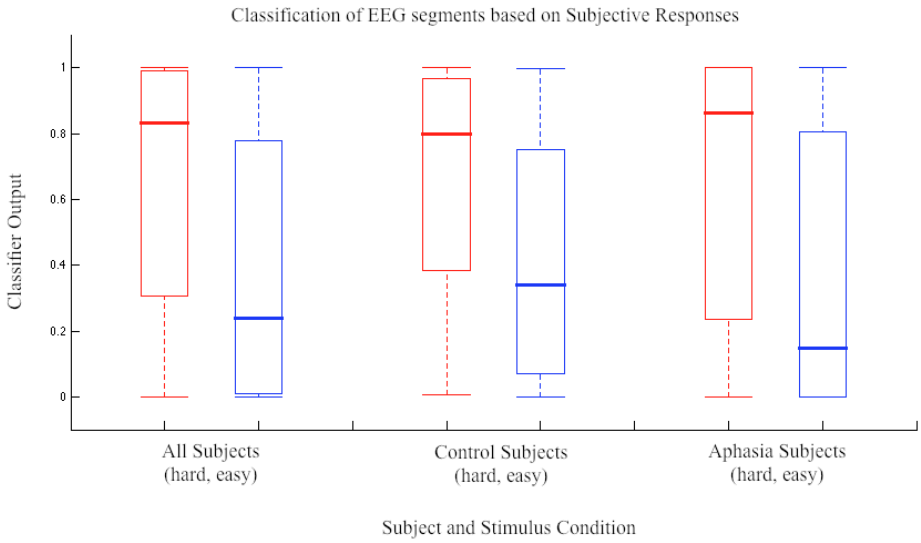

Figure 21: Above are box plots showing the relationship between the difficulty estimate gathered from subjective responses and the the logistic regressor outputs of the cognitive load estimation systems for individual subjects.

this problem results from what should be one of the strengths of the system. That is, it provides a continuous measure of cognitive effort, rather than a single report at the end of a task.

7 Conclusions

In this study, we found that subjective reporting and response accuracy are indicative of group differences between control and aphasia subjects. We also found that these measures present significant challenges, such as inconsistency between subjective task difficulty and the need to interpret outlier responses. Further, the real-time nature of the EEG measure makes it a viable tool for the development of novel rehabilitation techniques that may take advantage of live information regarding the effort required by a subject performing a task, without interruption.

However, this study has also made several of the limitations of the method more clear. Where earlier studies were able to separate periods of sustained high effort from periods of sustained low effort with a high degree of accuracy, our current study separated task conditions with lower accuracy. This limited result may be related to the mislabeling of data segments, as described in the previous section. It may also be the case that the models

derived from the machine learning steps were hindered by this inconsistency in the data.

To further develop an EEG-based system for monitoring cognitive effort, the effect of variability during effortful tasks should be investigated directly. This could be done by combining the analysis of sustained effortful cognitive activity with naturalistic tasks, such as those used in the current study. The neurophysiological signals associated with sustained cognitive load should be characterized, and a system based on these signals should be used to characterize effort during a task with variable, and known, differences in processing requirements. Once these relationships are understood, it may shed light on the temporal dynamics of cognitive effort during the types of naturalistic tasks that we considered here.

To continue improving cognitive load estimation in support of novel rehabilitation techniques, several additional considerations should be addressed. First, the processing of EEG should be further automated to reduce the necessity of technician involvement. The employment of robust statistics for identifying unreliable channels and segments of data described here should be evaluated on the basis of agreement with an expert analysis of EEG recordings. Our method for automatically separating ocular activity from cortical activity should be evaluated based upon its agreement with standard methods for removing ocular artifacts, including those that require the use of an EOG.

References

[1] A Behneman, C Berka, R Stevens, B Vila, V Tan, T Galloway, R R Johnson, and G Raphael. Neurotechnology to Accelerate Learning: During Marksmanship Training. *Pulse, IEEE*, 3(1):60–63, 2012.

[2] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(5 Suppl):B231–44, May 2007.

[3] BioSemi. ActiveTwo Mk2.

[4] BioSemi. ActiView.

[5] Rodney J Croft. *The removal of ocular artifact from the EEG*. Doctor of philosophy thesis, University of Wollongong, 1999.

[6] Rodney J Croft, Jody S Chandler, Robert J Barry, Nicholas R Cooper, and Adam R Clarke. EOG correction : A comparison of four methods. *Psychophysiology*, 42:16–24, 2005.

[7] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc, New York, New York, 2nd edition, 2001.

[8] P J Durka and K J Blinowska. Analysis of EEG transients by means of matching pursuit. *Annals of biomedical engineering*, 23(5):608–11, 1995.

[9] Piotr J Durka, Artur Matysiak, Eduardo Martínez Montes, Pedro Valdés Sosa, and Katarzyna J Blinowska. Multichannel matching pursuit and EEG inverse solutions. *Journal of neuroscience methods*, 148(1):49–59, October 2005.

[10] R J Echemendia, M Putukian, R S Mackin, L Julian, and N Shoss. Neuropsychological test performance prior to and following sports-related mild traumatic brain injury. *Clinical journal of sport medicine : official journal of the Canadian Academy of Sport Medicine*, 11(1):23–31, January 2001.

[11] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[12] P. Fayol, H. Carrière, D. Habonimana, and J.-J. Dumond. Preliminary questions before studying mild traumatic brain injury outcome. *Annals of Physical and Rehabilitation Medicine*, 52(6):497–509, July 2009.

[13] Robert J Ferguson, Brenna C McDonald, Andrew J Saykin, and Tim a Ahles. Brain structure and function differences in monozygotic twins: possible effects of breast cancer chemotherapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(25):3866–70, September 2007.

[14] F Franco-Marina, J J García-González, F Wagner-Echeagaray, J Gallo, O Ugalde, S Sánchez-García, C Espinel-Bermúdez, T Juárez-Cedillo, M Á V Rodríguez, and C García-Peña. The Mini-mental State Examination revisited: ceiling and floor effects after score adjustment for educational level in an aging Mexican population. *International Psychogeriatrics*, 22(1):72, 2010.

[15] Walter J Freeman. Hilbert transform for brain waves. *Scholarpedia*, 2(1):1338, 2007.

[16] Katerine A. R. Frencham, Allision M. Fox, and Murray T. Maybery. Neurophysiological Studies of Mild Traumatic Brain Injury: A Meta-Analytic Review of Research Since 1995. *Journal of Clinical and Experimental Neuropsychology*, 27(3):334–351, 2005.

[17] Alan Gevins and Michael E. Smith. Electroencephalography (EEG) in Neuroergonomics. In Raja Parasuraman and Matthew Rizzo, editors, *Neuroergonomics: The Brain at Work*, pages 15–31. Oxford University Press, Inc, New York, New York, 2007.

[18] Carlos Guerrero-mosquera and Angel Navia Vazquez. Automatic Removal of Ocular Artifacts from EEG Data using Adaptive Filtering and Independent Component Analysis. In *17th European Signal Processing Conference*, number Eusipco, pages 2317–2321, Glasgow, Scotland, 2009.

[19] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection 1 Introduction. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[20] D Hirtz, D J Thurman, K Gwinn-Hardy, M Mohamed, a R Chaudhuri, and R Zalutsky. How common are the "common" neurologic disorders? *Neurology*, 68(5):326–37, January 2007.

[21] Charles W Hoge, Dennis McGurk, Jeffrey L Thomas, Anthony L Cox, Charles C Engel, and Carl A Castro. Mild Traumatic Brain Injury in U.S. Soldiers Returning from Iraq. *The New England Journal of Medicine*, 358(5):453–463, 2008.

[22] Carrie A Joyce, Irina F Gorodnitsky, and Marta Kutas. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Society*, 41, 2004.

[23] Tzyy-ping Jung, Colin Humphries, Scott Makeig, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Extended ICA Removes Artifacts from Electroencephalographic Recordings. *Neural Information Processing Systems*, 10:894–900, 1998.

[24] Daniel Kahneman. *Attention and Effort*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1973.

[25] Mineichi Kudo and Jack Sklansky. Comparison of Algorithms that Select Features for Pattern Classifiers. *Pattern Recognition*, 33(1):25–41, 2000.

[26] Stephane G. Mallat and Zhifeng Zhang. Matching Pursuits With Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[27] Santosh Mathan, Andrew Smart, Trish Ververs, and Michael Feuerstein. Towards an Index of Cognitive Efficacy related cognitive decline. In *Engineering in Medicine and Biology*, 2010.

[28] T W McAllister, M B Sparling, L a Flashman, S J Guerin, a C Mamourian, and a J Saykin. Differential working memory load effects after mild traumatic brain injury. *NeuroImage*, 14(5):1004–12, November 2001.

[29] Michael Mccrea, Kevin M Guskiewicz, Stephen W Marshall, William Barr, Christopher Randolph, Robert C Cantu, James A Onate, and James P Kelly. Acute Effects and Recovery Time Following Concussion in Collegiate Football Players. *Journal of the American Medical Association*, 290(19):2556–2563, 2003.

[30] Michael Mccrea, James P Kelly, Christopher Randolph, Ron Cisler, and Lisa Berger. Immediate Neurocognitive Effects of Concussion. *Neurosurgery*, 50(5):1032–1042, 2002.

[31] Jed a Meltzer, Hitten P Zaveri, Irina I Goncharova, Marcello M Distasio, Xenophon Papademetris, Susan S Spencer, Dennis D Spencer, and R Todd Constable. Effects of working memory load on oscillatory power in human intracranial EEG. *Cerebral cortex (New York, N.Y. : 1991)*, 18(8):1843–55, August 2008.

[32] Neurobehavioral Systems. Presentation (software).

[33] S N Niogi, P Mukherjee, J Ghajar, C Johnson, R a Kolster, R Sarkar, H Lee, M Meeker, R D Zimmerman, G T Manley, and B D McCandliss. Extent of microstructural white matter injury in postconcussive syndrome correlates with impaired cognitive reaction time: a 3T diffusion tensor imaging study of mild traumatic brain injury. *AJNR. American journal of neuroradiology*, 29(5):967–73, May 2008.

[34] Paul L Nunez. Electroencephalogram. *Scholarpedia*, 2(2):1348, 2007.

[35] Fred Paas and Pascal W M Van Gerven. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Technology*, 38(1):63–71, 2003.

[36] Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. Recipes for the linear analysis of EEG. *NeuroImage*, 28(2):326–41, November 2005.

[37] P Pudil, J Novovieova, and J Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(June 1993), 1994.

[38] Philip Schatz, Jamie E Pardini, Mark R Lovell, Michael W Collins, and Kenneth Podell. Sensitivity and specificity of the ImPACT Test Battery for concussion in athletes. *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists*, 21(1):91–9, January 2006.

[39] C. D. Wickens. Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455, June 2008.

[40] Christopher D. Wickens. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177, 2002.

A Additional Results

Figures 22, 24, and 26 show several results broken down by individual subjects.

Figures 25, 26, 27, and 28 show the relationship between additional subjective questions that were asked after each passage and comprehension question.

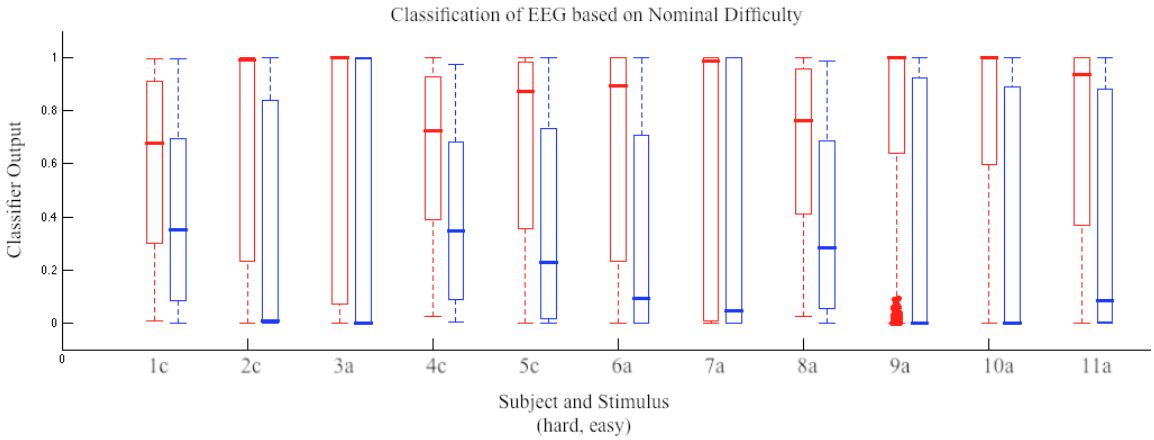

Figure 22: Above are box plots showing the relationship between the original stimulus manipulation and the logistic regressor outputs of the cognitive load estimation systems for individual subjects. Each box plot shows the distribution of classifier prediction for a particular stimulus type (signified by color, with red for hard and blue for easy) for a particular subject (shown in the x-axis label).

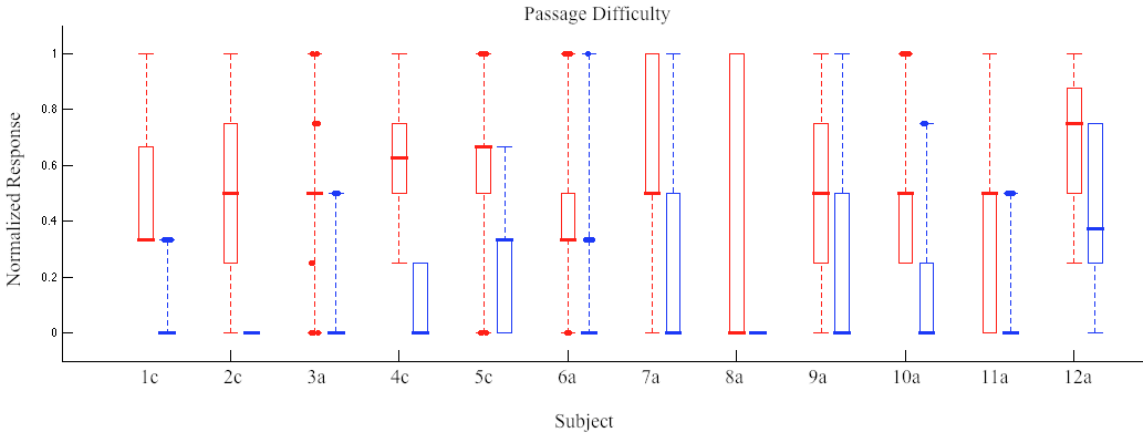

Figure 23: Above are box plots showing the relationship between the original stimulus manipulation and the subjectively-reported difficulty of passages provided by subjects during the experiment. The responses were normalized to the continuous range [0,1] to adjust for different usage of the available scale.

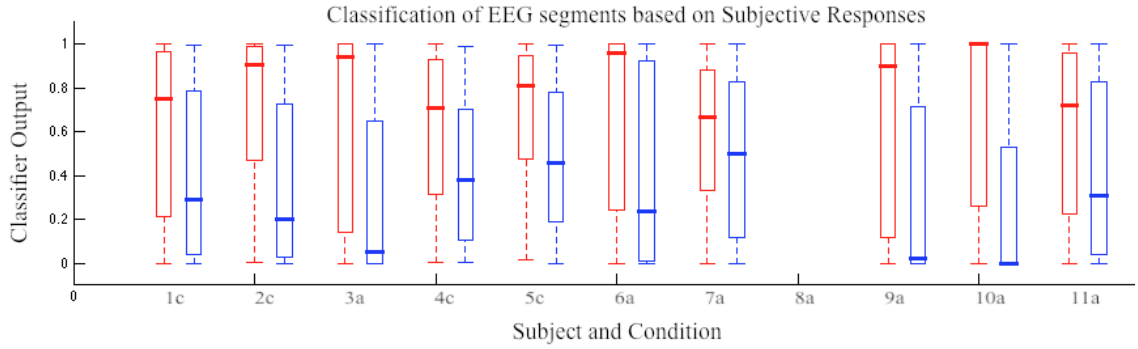

Figure 24: Above are box plots showing the relationship between the difficulty estimate gathered from subjective responses, and the logistic regressor outputs of the cognitive load estimation systems for individual subjects. Subject 8 provided uniform subjective responses, resulting in a failed difficulty projection, so was excluded. This could be addressed in the future using an alternative projection method.

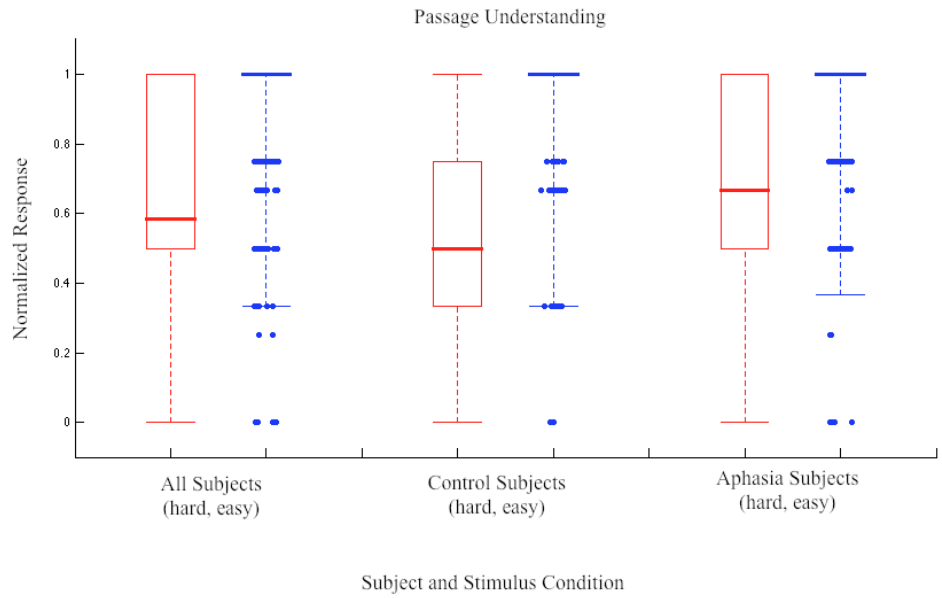

Figure 25: The above box plots show the distribution of responses to a question regarding a subject's understanding of the passage they had just heard. Distributions are divided by subject group and by the nominal difficulty of the passage that had just been heard.

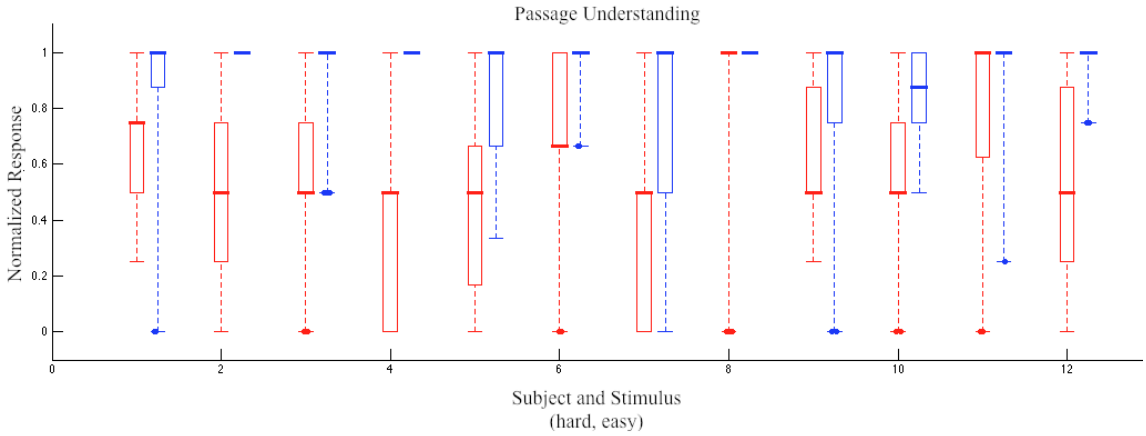

Figure 26: The above box plots show the distribution of responses to a question regarding a subject's understanding of the passage they had just heard. Distributions are divided by subject and by the nominal difficulty of the passage that had just been heard.

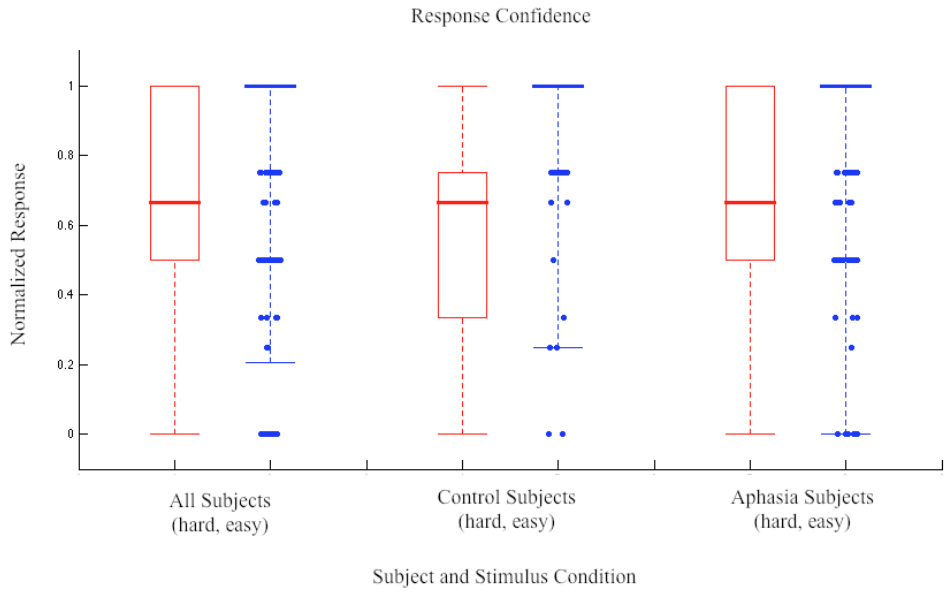

Figure 27: The above box plots show the distribution of responses to a question regarding a subject's confidence in their response to a comprehension question that followed a passage. Distributions are divided by subject group and by the nominal difficulty of the passage that had just been heard.

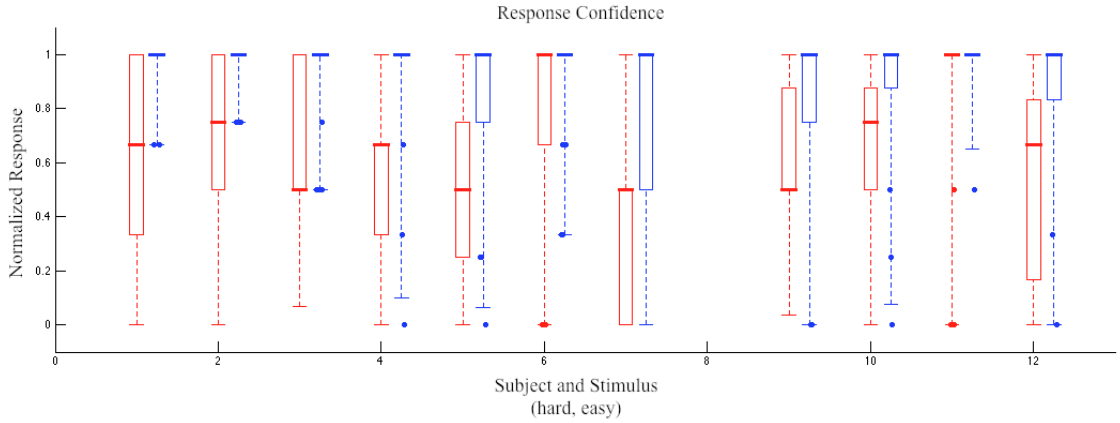

Figure 28: The above box plots show the distribution of responses to a question regarding a subject's confidence in their response to a comprehension question that followed a passage. Distributions are divided by subject and by the nominal difficulty of the passage that had just been heard.