Using a Low-cost EEG Sensor to Detect Mental States

Bao Hong Tan

CMU-CS-12-134

August 2012

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Prof. Jack Mostow, Chair
Prof. Tai Sing Lee

Submitted in partial fulfillment of the requirements
for the Degree of Master of Science

Involvement of the advisors for this work was supported in part by the National Science Foundation through Grant 1124240 and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628 to Carnegie Mellon University. The opinions expressed are those of the author and do not necessarily represent the views of the National Science Foundation, the Institute, or the U.S. Department of Education.

Keywords: single-channel EEG, mental states, intelligent tutoring, reading tutor, dry sensor, frequency band, power spectrum

Abstract

The ability to detect mental states, whether cognitive or affective, would be useful in intelligent tutoring and many other domains. Newly available, inexpensive, single-channel, dry-electrode devices make electroencephalography (EEG) feasible to use outside the lab, for example in schools. Mostow et al. [2011] used such a device to record the EEG of adults and children reading various types of words and easy and hard sentences; the purpose of this experimental manipulation was to induce distinct mental states. They trained classifiers to predict from the reader's EEG signal the type of the text read. The classifiers achieved better than chance accuracy despite the simplicity of the device and machine learning techniques employed.

Their work serves as a pilot study for this thesis and provides the data set for all analyses in this work. This thesis further explores the properties and temporal structure of the EEG signal with the aim of improving the accuracy of detecting mental states. The EEG signals associated with the word stimuli are analyzed for the existence of event-related potentials (ERPs) that could distinguish the word type, which in turn could be exploited in classification. The EEG signals for the sentence stimuli are subjected to various feature extraction methods and temporal manipulations. This thesis demonstrates the potential of exploiting the temporal structure of EEG signals in detecting mental states with low-cost devices.

Acknowledgements

I would like to extend my greatest appreciation and gratitude to my advisor, Jack, and my co-advisors Kai-min and José, all of whom have given me much support, encouragement and useful advice during the past one year of hard work.

I would like to thank Prof. Lee for serving on the committee and providing me with constructive criticism on this work. Thanks are also due to Deborah for handling the administrative work for the Fifth Year Masters Program.

I would also like to thank members of Project LISTEN who have helped me directly or indirectly in completing this work.

Lastly, I would like to thank my family and friends for their constant moral support and encouragement.

Contents

Chapter 1

Introduction

The ideal intelligent tutor could understand students' minds to identify their mental states such as knowledge, emotions, and thoughts, and decide what and how to teach at each moment [Mostow et al., 2011]. Mostow et al. [2011] investigated the use of electroencephalography (EEG) in Project LISTEN's Reading Tutor [Mostow and Beck, 2007], an intelligent tutor that helps children learn to read. They tested the EEG signal as a complementary source of information to identify students' mental states, as exploratory work towards the ultimate goal of using EEG to guide tutorial decisions in the Reading Tutor. This thesis builds on their work, and further explores the properties of the EEG signal in identifying mental states useful for automating tutorial decisions.

Figure 1.1: A child using the Reading Tutor at the Kofi Annan Technology Centre in Ghana. Photograph taken by G. Ayorkor Korsah in 2007.

1.1 Using EEG to detect mental states

The EEG signal is a voltage signal that arises from synchronized neural activity, that is, the coordinated firing of millions of neurons in the brain. It can be measured by non-invasively placing an electrode on or near the scalp, or for greater accuracy, by implanting an electrode in the skull (e.g., Donoghue et al. [2007]). As noted by Mostow et al. [2011], synchronized neural activity varies according to development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation. For example, rhythmic fluctuations in the EEG signal occur within several particular frequency bands, and the relative level of activity within each frequency band has been associated with brain states such as focused attentional processing, engagement, and frustration [Berka, 2007; Lutsyuk, 2006; Marosi, 2002], which in turn are important for and predictive of learning [Baker, 2010].

The idea of detecting mental states using EEG is not new. In fact, a brain-computer interface (BCI) system is specifically designed to detect the mental state of its user so as to carry out the

user's desired action. In other words, a BCI system allows the brain to communicate with the system directly through EEG signals. The study of BCI as a field emerged from the desire for new assistive technology, targeted at handicapped patients, especially those paralyzed [Cashero, 2011]. Some BCI systems, such as speller programs and mouse cursor controllers, exploit evoked responses from the brain when the user is presented with a stimulus. Such a response is known as an event-related potential (ERP) which is the measured brain response that is the direct result of a specific sensory, cognitive, or motor event [Luck, 2005]. An ERP consists of one or more components that appear as deflections in the EEG signal. BCI systems could detect the elicited ERP components to perform the user's desired action.

Even though it recorded students' EEG signals in Mostow et al.'s study, the Reading Tutor cannot be considered as a complete BCI system, as students still have to interact with it through the computer keyboard, mouse, and microphone (for speech recognition). Moreover, using EEG to automate tutorial decisions is not intended to be an explicit function students can utilize, although there is nothing to stop them from doing so. There might be future plans to allow other aspects of the tutor program to be controlled through EEG, but the current goal of using EEG to automate tutorial decisions still remains a challenge.

Far from being a BCI system currently, the desirable (future) version of Reading Tutor nevertheless shares similar problems in data acquisition and processing with existing BCI systems that use scalp level EEG. The EEG signal recorded at the scalp level is inherently noisy since the voltage signal has to pass through the skull and brain tissues. Furthermore, the strength of the voltage signal is typically on the order of microns of voltage, weak enough to be easily corrupted by the slightest level of noise. Besides this undesirable nature of EEG, environmental and usage conditions also play a role in EEG recording. For instance, a BCI system could be deployed in an informal setting such as the home and school where it is difficult to obtain the highest quality EEG signals due to muscle and eye artifacts generated by the user and background noise generated by the environment, and possibly irrelevant brain processes. Many BCI systems have been successfully developed despite these unfavorable environmental and usage conditions, thus we hope that the Reading Tutor can also perform reasonably well under the same conditions.

1.2 Low-cost EEG devices

Multi-electrode, medical grade EEG systems have long been used in hospitals and laboratories. But the recent availability of low-cost EEG devices makes it feasible to take this technology from the laboratory into informal environments such as schools and homes. The benefits of such devices are affordability and ease of use. Several types of low-cost EEG devices exist commercially in the market today. For example, the Emotiv EPOC is a 14-channel device that requires the application of saline solution on the user's scalp to increase the quality of the EEG signal recorded. Another device, used in this thesis, is the NeuroSky Mindset™ (2009), an audio

headset equipped with a single, dry EEG sensor. It uses Bluetooth® technology to transfer signal samples wirelessly to the host computer.

Figure 1.2: The Mindset device. Picture taken from www.neurosky.com

The Mindset measures the voltage between an electrode resting on the forehead and two electrodes each in contact with an ear to serve as a ground reference. More precisely, the position on the forehead is Fp1, as defined by the International 10-20 system [Jasper, 1958]. Furthermore, unlike other multi-channel devices, especially those used in laboratories, the Mindset requires no gel or saline for recording, and requires no expertise to wear. The Mindset also includes proprietary noise cancellation technologies that eliminate known noise frequencies from sources such as muscle movements and electrical devices. Notch filters eliminate electrical noise from the power source, which varies from 50 to 60 Hz depending on the geographical location [NeuroSky, 2012b].

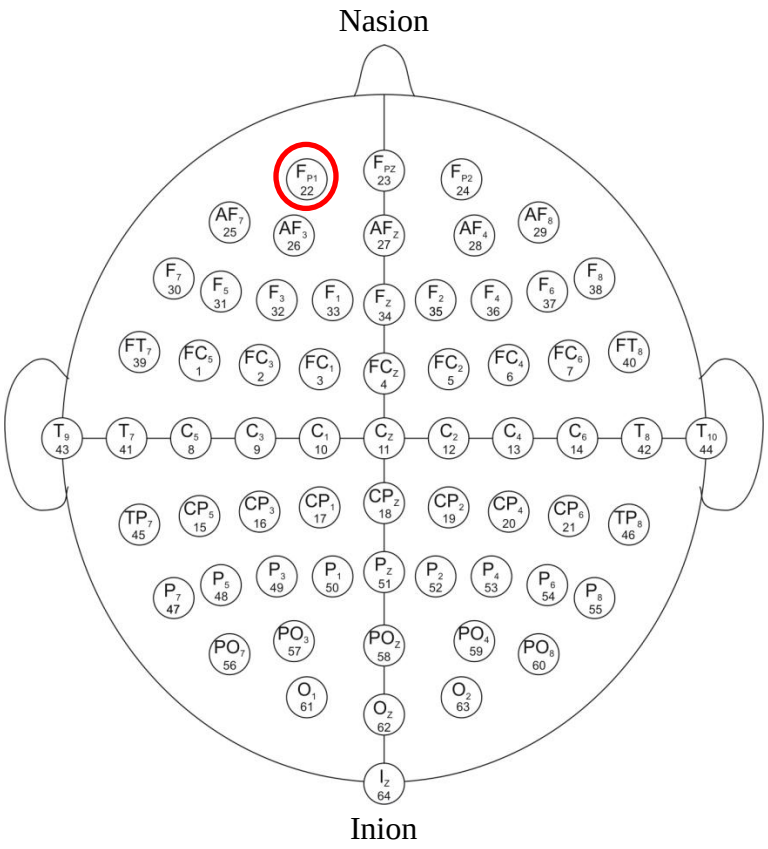


Figure 1.3: Placement for a 64-electrode system using the International 10-20 standard. Reference points are nasion, which is the point between the forehead and the nose (top of diagram); and inion, which is the bony lump at the base of the skull on the midline at the back of the head (bottom of diagram). Original picture taken from Sharbrough et al. [1991]. A bold circle added at the top indicates Fp1.

Besides recording the EEG signal, the Mindset provides *attention* and *meditation* signals claimed to measure the corresponding mental states. These two signals are derived from the EEG signal using proprietary techniques. Despite the limitations of a single, dry sensor, and working with untrained users, the Mindset produced attention and meditation signals that distinguished between the two mental states with 86% accuracy [NeuroSky, 2009b]. Another study by NeuroSky shows that the EEG signal collected by the Mindset is similar to that of the Biopac system, wet-electrode equipment widely used in medical and research applications [NeuroSky, 2009a]. In fact, the Mindset is more noise-resistant than Biopac in low frequency bands due to the use of fixed and short (less than 10 inches) transmission wires carrying the EEG signal. The length of wires in the Biopac system was 3 feet long. As explained in section 1.4.1, there are several independent studies showing that the Mindset can provide information useful enough for detecting mental states.

1.3 Pilot study

This thesis extends the work by Mostow et al. [2011], who assessed the feasibility of collecting useful information about cognitive processing and mental states with the Mindset. According to them, the ability to record longitudinal EEG signals in authentic school settings is important for several reasons. One is that they can analyze learning over intervals longer than a lab experiment, in contrast to short-term memory effects. Another is that they can study data generated by children's *"in vivo"* behavior at school, rather than their more constrained behavior in unfamiliar laboratory settings under intense adult supervision. Children would also feel more comfortable wearing the single, dry-electrode Mindset than multi-, wet-electrode devices.

Mostow et al. [2011] conducted a pilot study in which participants wore the Mindset while using the Reading Tutor in a school setting. The Reading Tutor displays text, listens to the student read aloud, and logs detailed longitudinal records of its multimodal tutorial dialogue to a database [Mostow and Beck, 2009]. More simply put, the Reading Tutor is an intelligent tutor that helps children learn to read. The study was designed to find out whether EEG can detect when reading is difficult and whether EEG can detect lexical features. They used easy and hard passages to simulate difficulty, and various types of isolated words to provide the lexical features. The words and passages can be considered stimuli to induce different mental states.

They used the Mindset in the pilot study to record the EEG of adults and children reading passages and isolated words, both silently and aloud. The participants sat in front of a computer and small body movements were unavoidable while the EEG signal was being recorded. They trained logistic regression classifiers to predict from a reader's EEG signal the difficulty of the passage or the type of the word read. The classifiers achieved statistically above-chance accuracy despite the simplicity of the machine learning technique employed. Their results demonstrated that the stimuli can induce detectable, distinct mental states. Also, the EEG signal generated by a user in a school setting is reliable enough for mental states to be detected.

Using the same data set, the main goal of this thesis is to investigate how the temporal structure of the EEG signal can lead to an improvement in classification accuracy. Interested readers can refer to Chapter 2 for detailed information on the pilot experiment and the EEG data set used in the pilot study as well as this thesis.

1.4 Related work and state of the art

This section presents further information on EEG and low-cost devices as well as some prior studies that are related to the work in this thesis.

1.4.1 Low-cost devices

There are relatively fewer studies done using the Mindset and other types of low-cost EEG devices, than using laboratory-grade equipment. One of the reasons could be that low-cost devices were not available until several years ago, dating back to the development of the NeuroSky technology in 1999, and the release of the first consumer based EEG device in 2007.

Another reason could be that low-cost devices are still deemed unsuitable for high-quality research applications. In fact, Chi et al. [2010] noted that dry electrodes, such as the one in the Mindset, are limited to niche, nonmedical/scientific applications like toys and fitness monitoring. There are even fewer studies done using a single-channel EEG device. Most existing literature based on the Mindset was published just a few years ago, primarily on the proprietary attention and meditation signals.

Crowley et al. [2010] conducted two psychological tests, Stroop's color-word inference test [Stroop, 1935] and the Tower of Hanoi test [Hinz, 1989], to induce stress in subjects while the the Mindset produced the attention and meditation signals. They conducted Stroop's test as a preliminary study before achieving an accuracy of 78.04% in identifying a subject's stress level in the Tower of Hanoi test. As they were unable to correlate isolated moments of human error in Stroop's test with a precise change in the attention or meditation signals, they had to rely on the overall trend in the signals to detect when a subject undergoes a change in these emotions. Both tests demonstrated the Mindset's suitability as a minimally invasive means of measuring the attention and meditation level of a subject. Another group of researchers, Haapalainen et al. [2010], collected data from multiple bio-sensors and compared their ability to assess cognitive load. The attention signal was the third best source of information to use, after heat flux (rate of heat dissipation from body) and electrocardiogram (ECG), with an average classification accuracy of 60.2% across the various cognitive experiments they conducted.

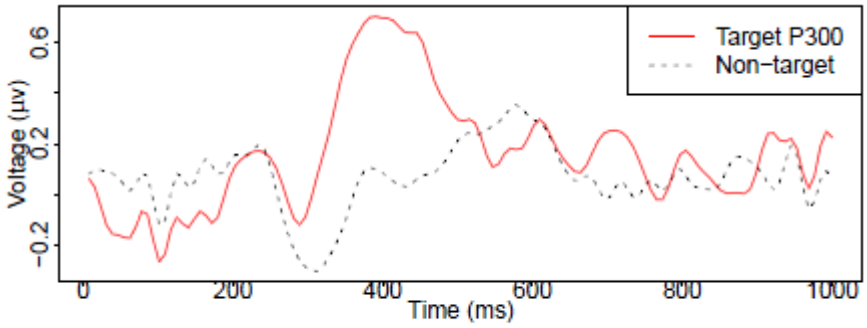
The successes achieved by these researchers suggest that the quality of EEG signal recorded is sufficiently high to provide useful derivative information, in this case the attention and meditation signals. However, in order to make the results replicable and independent of the underlying recording device, this thesis avoids dependence on proprietary technologies and instead uses the EEG signal and other information and features derived from it through non-proprietary methods and algorithms.

Luo and Sullivan [2010] used a custom-built, dry, single-channel EEG sensor to study steady-state visual evoked potential (SSVEP), a type of event-related potential (ERP) induced by a flashing visual stimulus. Luo et al. conducted an experiment in which each participant was allowed to look at any one out of four possible lights flashing at 9, 10, 11 and 12 Hz. The sensor was placed at PO2, located at the back of the head where the visual cortex is. The raw EEG signal was time-locked to each stimulus onset and was preprocessed using a finite impulse response (FIR) filter with a cut-off frequency of 40 Hz to remove high frequency noise. They segmented the processed EEG signal into trials corresponding to each flash of light under the assumption that the participant looked at the 9 Hz light. Then they computed the correlation coefficients of the signal samples between every possible pair of the trials, and used the median value of the coefficients as a feature. The segmentation and computation steps were repeated for the other lights to obtain 4 features in total. Using the features, they were able to determine which of the four lights each participant was looking at with an average accuracy of 87.5% across the participants. This study shows that a single-channel sensor can collect raw EEG signal

of sufficient quality to conduct ERP studies. The main difference between Luo et al.'s study and this thesis is the stimuli used. In this work, each stimulus is either a sentence or an isolated word, and has a longer and more varied duration than that of a single flash of light. Furthermore, a flash of light is known *a priori* to induce an ERP unlike the sentence stimuli.

1.4.2 Event-related potential

As with traditional EEG based research, this thesis also investigates the presence of event-related potentials that could be exploited to distinguish mental states. Some BCI systems such as the P300 speller developed by Farwell and Donchin [1988] also make use of ERPs. The P300 is an ERP component that is observable as a positive deflection peaking at 300 milliseconds after the onset of a suitable stimulus. The P300 speller allows a user to type a letter at a time, by flashing rows and columns of letters for the user to choose from. Like Luo et al., Farwell and Donchin made use of the ERP component elicited (P300) by a flashing stimulus. Figure 1.4 below shows P300 along with much noise even after averaging twenty trials together. With our small data set, we would expect plots with a similar level of noise.



(f) 20 averaged

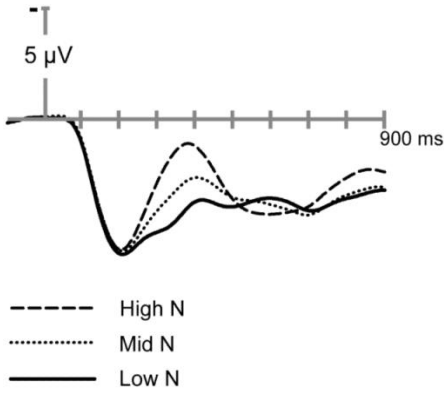
Figure 1.4: P300 after averaging 20 trials together. Plot taken from Cashero [2011].

The exact ERPs to be investigated come from published studies and visual inspection of averaged EEG signals in the data set. In the field of linguistics, N400 is the most commonly encountered ERP component that is strongest over the centro-parietal part of the scalp (e.g., position Pz, see Figure 1.3). Like P300, N400 is an observable negative deflection in the EEG signal peaking at 400 milliseconds after the onset of the stimulus. Much research has been done to find out how lexical features affect the occurrences and amplitudes of ERP components, including N400, but an in-depth treatment of this subject is beyond the scope of this thesis.

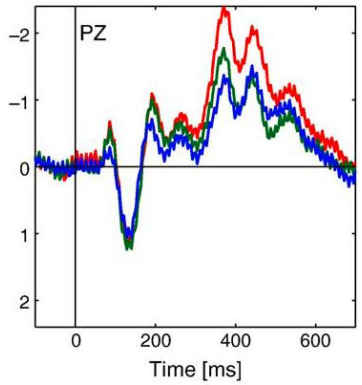
Dambacher et al. [2006] studied the effects of frequency, predictability and position of words on P200 and N400. P200 is a positive deflection peaking at 200 milliseconds post-stimulus. Their experiment using open-class words ("nouns, verbs, adjectives, and 'ly' adverbs") shows that P200 and N400 can appear together. In too many studies to cite here, N400 is reported as a functionally specific marker of lexico-semantic processing, often elicited as a response to semantically inappropriate words in sentences [Friederici et al., 1993; Kutas and Hillyard, 1983].

Also, N400 has long been observed in response to pronounceable pseudo-words, such as LORK [Laszlo and Federmeier, 2011]. Laszlo et al. studied whether the orthographical or lexical properties of an input word would have any effect on the amplitude of N400 elicited by that input. Their work shows that N400 can even be elicited by meaningless, non-words (illegal letter strings) with minimum orthographic neighborhood size (number of real words that differ in spelling by only one letter). However, they also explained that N400 is dependent on the context of the sentence containing the non-word and is not always observable. Laszlo et al.'s work is highly relevant to this thesis because the stimuli used in the pilot study included pseudo- and non-words. We use the popular method of Analyses of Variance (ANOVA) in the analyses of ERPs in this work.

Figure 1.5 below shows examples of P200 and N400 taken from Laszlo et al. (left) and Dambacher et al. (right), both of which obtained from averaging across hundreds of trials. Both plots show three main differences. First, though taken from the same location of the scalp (central parietal), the N400 in Laszlo et al.'s study stays positively below the horizontal axis while the N400 in Dambacher et al.'s study crosses the horizontal axis in the negative direction. In fact, N400 need not be negative in absolute terms, as it is negative-going relative to the reference electrode [Kutas and Federmeier, 2011]. Second, the N100 component appears in Dambacher et al.'s plot (though not mentioned in the study), but not in Laszlo et al.'s. The first two differences could be attributed to the different stimuli used (neighbor frequency vs. word frequency), and unlikely the location of the reference electrode or the re-referencing method used after data collection. In both Laszlo et al.'s and Dambacher et al.'s studies, electrodes were referenced online to the left mastoid and then digitally re-referenced offline to the average of the left and right mastoids. The last difference is the magnitude of P200 amplitude (5 vs. 2 μV) which could be due to the equipment and electrical amplifier used. Overall, as evident in the two plots below, N400 could appear with various physical properties, in the presence or absence of N100.



(Laszlo et al. 2011)



(Dambacher et al. 2006)

Figure 1.5: Examples of P200 and N400 after averaging across trials. *N* refers to the orthographic neighborhood size. The vertical axis represents the amplitude of the signal and is inverted with the negative polarity on top. The horizontal axis represents time in milliseconds.

1.4.3 Functional connectivity

Functional connectivity is a general term that has no specific definition but has similar meaning throughout the literature. It is usually considered as the temporal correlations between spatially distinct parts of the brain [Liang et al., 2006; Supekar et al., 2008], or between spatially remote neurophysiological events [Friston and Büchel, 2006]. In other words, functional connectivity is a measure of how distinct brain regions interact with one another in the course of a neurophysiological event. As explained by Friston and Büchel [2006], functional connectivity is simply a statement about the observed correlations, it does not comment on how these correlations are mediated. For example, correlations can arise in signals recorded by electrodes placed on different sites of the scalp, and in phasic coupling of neural assemblies. A neural assembly is a group of neurons that synchronize their activity to perform a specific computation [Fingelkurts et al., 2006].

Functional connectivity can be measured and analyzed using various statistical and even graphical methods (e.g., Smit et al. [2008] and Stam et al. [2009]). Much of the existing literature analyzed the connectivity between distinct brain regions. The two most popular methods used to capture activity in multiple brain regions simultaneously are functional magnetic resonance imaging (fMRI) and multi-electrode EEG recording. fMRI is an MRI procedure that captures changes in blood flow in the entire brain at once [Huettel et al., 2008], and is widely used in studies that examine brain activity in patients who suffer from brain diseases such as Alzheimer's (e.g., Supekar et al. [2008]) and schizophrenia (e.g., Liang et al. [2006]). Analyzing the pattern of functional connectivity between brain regions in healthy vs. non-healthy subjects is useful for detecting the disease of interest.

EEG coherence is another specific form of functional connectivity commonly used in EEG studies to examine functional interactions between brain regions (e.g., Weiss and Rappelsberger [2000] and Srinivasan et al. [2007]). Srinivasan et al. suggested that relating various measures of neural dynamics to functional brain state is the central goal of EEG studies. Srinivasan et al. also suggested that coherence, a correlation coefficient that estimates the consistency of relative amplitude and phase between any pair of signals at each frequency, is one of the most promising measures of neural dynamics. Essentially, coherence measures how well two given EEG signals relate to each other in each of their constituent frequency bands. Coherence can be computed by the magnitude squared coherence function [Kay, 1988] using for example, Welch's overlapped average periodogram method [Rabiner and Gold, 1975; Welch, 1967]. Srinivasan et al. noted that the coherence measure is distinct from *synchrony* [Singer, 1999], which usually refers to signals oscillating at the same frequency with identical phases, in the neuroscience literature; coherence is instead a measure of synchronization between two signals based mainly on phase

consistency. For example, two signals may have different phases but as long as the phase difference is constant, they are said to be coherent.

EEG coherence provides important information on the interactions between neural systems operating in each frequency band [Srinivasan et al., 2007]. According to Nunez and Srinivasan [2005], each EEG electrode signal is a superposition of signals from several brain sources operating at various or even a combination of frequencies. The contribution of each source depends on source and electrode locations and the spread of electrical current through volume conduction [Nunez and Srinivasan, 2005]. (Volume conduction refers to the process of electrical signals and electromagnetic field propagating through brain tissues and the scalp.) Thus, EEG amplitude in each frequency band can be related to the synchrony of the underlying current sources [Nunez and Srinivasan, 2005], since the more synchronized the sources are, the higher the amplitude is. It must be noted that signals become attenuated over distance. Depending on the goals of an experiment, volume conduction can be undesirable since each electrode typically records a mixture of signals from multiple, rather than one, physically nearby sources. Thus, volume conduction defeats the intention of recording an uncontaminated version of the signal generated by the part of the brain at which an electrode is specifically placed. However, looking from a positive perspective, it alleviates the lack of data from other parts of the brain caused by the use of a single EEG channel. As explained earlier, volume conduction permits signals from multiple parts of the brain to superpose at the single electrode on the forehead (Fp1), thereby enriching the signal recorded at that electrode with "extra information", albeit the overall information obtained would not be as much as that from multiple electrodes well spread around the scalp. This hypothesis of extra information is supported by Srinivasan et al. [2007], who found that volume conduction can elevate EEG coherence at all frequencies for moderately separated (< 10 cm) electrodes; a smaller elevation is observed with widely separated (> 20 cm) electrodes. This suggests that a brain region at the back of the head can have an impact on the signal recorded by an electrode at the front.

Due to the limitation of a single EEG channel, the coherence measure between channels cannot be applied in this thesis. Instead, we analyze the (Pearson) correlations between frequency bands. Analyzing the correlations between frequency bands in a given EEG signal provides insight into how well various brain regions synchronize with one another in terms of their activities, although it is not possible to identify the contributing brain regions from just a single channel. Liang et al. [2006] also used correlation coefficients in analyzing fMRI time series from different brain regions. They computed the correlation coefficients between each pair of brain regions and compared the coefficients from schizophrenia patients to that from healthy subjects. Their results identified abnormalities in functional connectivity.

1.5 Goals and contributions

This thesis extends the pilot study [Mostow et al., 2011] by exploring the properties and temporal structure of sentence and word trials, and investigates how to improve the accuracy of

detecting mental states. The phrase "mental state" refers to thoughts and emotions. For example, the "hard" mental state refers to the mental state that occurs when the user is reading difficult text. We break down the overall goal of exploiting temporal structure to improve the accuracy of detecting mental states, into the following questions along with the proposed approaches:

1. **How can we model the temporal structure of word and sentence trials?** For word trials, the temporal model for each word type consists of the time windows in which ERPs are elicited. For sentence trials, we propose two temporal methods, relative segmentation and truncation, as a means to model the temporal structure.

2. **How can we represent the properties of the EEG signal?** For word trials, we use the mean ERP amplitudes in each time window. For sentence trials, we use several feature extraction methods, based on the EEG signal itself and frequency bands derived from the EEG signal.

3. **What level of performance can we achieve in detecting mental states, using the proposed temporal models?** For word trials, we evaluate the performance as the accuracy achievable by classification tasks using the mean ERP amplitudes as features. For sentence trials, we also evaluate the performance of the temporal models through classification tasks using the proposed features.

The expected contributions are:

1. Evaluation of the proposed temporal models for word and sentence trials in detecting mental states, based on the proposed features.
2. Evaluation of EEG and frequency band features in detecting mental states, and their performance with respect to the temporal models.

Note that the performances of the temporal models and the features are intricately tied to each other, since, for example, the proposed features might not be compatible with the temporal models.

The results in this work also provides insight into the level of classification accuracy one could expect with a low-cost, dry, and single-channel EEG sensor in an informal environment.

Though this thesis is largely exploratory work towards the ultimate goal of automating tutorial decisions, the methods and results presented in this thesis could still be useful in other relevant domains.

1.6 Overview of thesis

Figure 1.6 below shows the experimental workflow in this thesis.

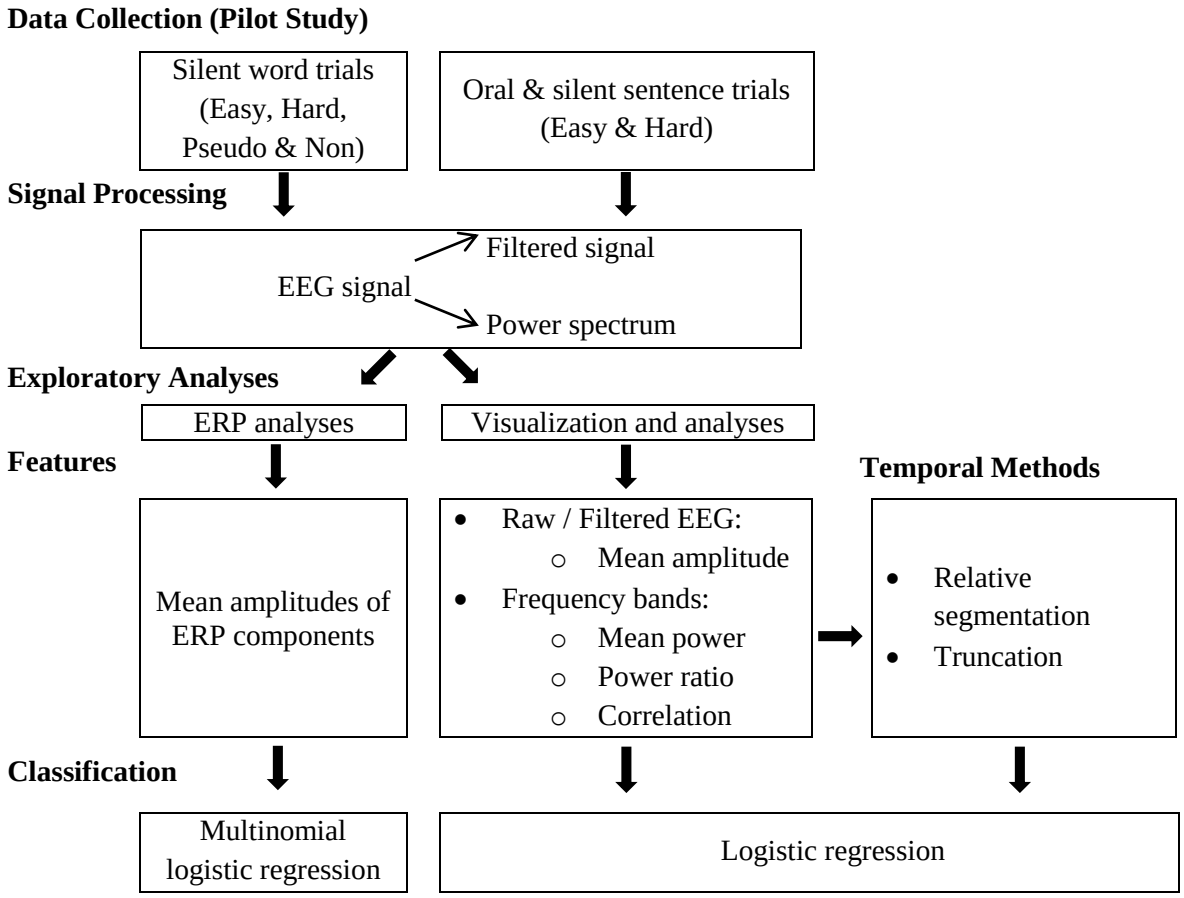


Figure 1.6: Experimental workflow in thesis.

This thesis is organized according to the workflow. Chapter 2 describes the pilot study experiment, the resulting data set used in this thesis, and details of signals processing. Chapter 3 describes the methods, features and classifiers used in the analyses. Chapter 4 presents analyses of word trials and classification results. Chapter 5 presents analyses of sentence trials, in particular, exploratory analyses of the correlation feature, classification results, and temporal analyses. Chapter 6 concludes this work with a discussion of the results followed by insights and avenues for future work.

Chapter 2

Data

This chapter provides an overview of the pilot data set, including how the data set was collected and processed, as well as the stimuli used.

2.1 Pilot experiment and data set

Mostow et al. [2011] collected the data set in December 2010. The NeuroView (version 4.2.2) software from NeuroSky saved the digitized EEG signal samples transferred wirelessly from the Mindset device. The experimental protocol was implemented in the Reading Tutor's homegrown language for scripting interactive activities. 10 adults (approximately 20–60 year old) and 11 children (9 and 10 year old) participated in the experiment in a school setting. A few other participants user-tested the protocol or had no EEG data. All of the children were native English speakers while some of the adults were not (it is not known how many exactly).

2.2 Stimuli

There were two types of stimuli: passages each consisting of several sentences, and isolated words. Each subject read the same passages and words. There were 6 *easy* and 6 *hard* passages altogether for both the silent and read-aloud conditions. The easy passages were from texts classified by the Common Core Standards[1] at the K-1 level while the hard passages were from practice materials for the Graduate Record Exam[2] and the ACE GED test[3]. The sentences in a passage form a coherent story; each sentence, instead of the entire passage, is the unit of analyses.

	Sentence		Passage	
	Min/Max	Average	Min/Max	Average
Easy	3–18	7	65–79	70
Hard	2–35	20	62–83	76

Table 2.1: Number of words in easy and hard sentences and passages.

It must be emphasized that the passages were not designed to intentionally induce mental states. Initially, we hypothesized that since text induces comprehension, the difficulty of a sentence also induces distinct mental states. This hypothesis has been verified in the pilot study, in which trained classifiers could predict from a subject's EEG signal the difficulty of the sentence read, with statistically above-chance accuracy.

Besides passages, there were 20 real words, 20 *pseudo*-words, and 10 *non*-words altogether for both reading conditions. The real words were all 2-syllable 7-letter words; 10 of them were easy

[1] http://www.corestandards.org
[2] http://majortests.com/gre/reading_comprehension.php
[3] http://college.cengage.com/devenglish/resources/reading_ace/students

and the rest were hard, to see if the EEG data reflected difficulty in word reading. Prior work [Bizas, 1999] had found distinct EEG indicators of visual-spatial, orthographic, phonological, and semantic operations in reading. The easy words had a Kucera-Francis (K-F) frequency of 30 or more (mean = 84) and an age of acquisition (AOA) below 315 on a scale from 0 to 700 (mean = 254.4) [Coltheart, 1981]. The hard words had a K-F frequency below 10 (mean = 3.4) and an AOA above 450 (mean = 555.5). Examples of the easy words are "chicken" and "brother"; hard words, "cologne" and "brocade".

Pseudo-words were 3-letter pronounceable strings, chosen to vary in their number of orthographic neighbors (words that differ in spelling by only one letter), since EEG data (specifically, event-related potentials) are sensitive to neighborhood size [Laszlo and Federmeier, 2011]. Mostow et al. included the non-words to see if they could detect when readers saw unfamiliar words. Each non-word was an illegal 3-letter string, also with varying orthographic neighborhood sizes, also from the same study. They varied the orthographic neighborhood size of the pseudo-words and non-words from 0 neighbors to 22 neighbors, to enable analysis of its effects.

2.3 Procedure

Each subject wore the Mindset with the front electrode resting on his or her forehead and with the ear cups properly in contact with the ears to provide ground reference. The experimenter ensured that the device was correctly worn and the EEG signal received properly in the software. To simulate the environment typical Mindset users would be in, no effort was made to ensure the electrode was exactly at the Fp1 position, and subjects were allowed to have body movements, including eye blinks. Consequently, the recorded EEG signals could include muscle and ocular (eye) artifacts, especially when the signals were recorded during the read-aloud condition. Each subject took part in only one session. His or her EEG signal was recorded continuously while he or she read the stimuli presented by the Reading Tutor. There was no break in recording, except on the rare occasions when the headset needed adjustment for proper recording of the EEG signal. Each subject was instructed to read at his or her own pace and to click on a button shown on the computer screen when ready to advance to the next word or sentence.

Table 2.2 below shows the number of passages and isolated words read by each subject in each condition, in a single session. Each session consisted of two phases. In the first phase, the subject was asked to read aloud. The Reading Tutor displayed 3 easy and 3 hard passages in an alternating order, one sentence at a time. Although instructed to read aloud, the readers (especially children) did not always read the displayed sentences correctly. Each passage was followed by a multiple-choice cloze question (formed from the next sentence in the passage) to ensure that readers were reading for meaning. Next, 10 real and 10 pseudo-words were displayed one at a time, in a randomized order, for the subject to read aloud as well. The second phase resembled the first, but conducted in a silent reading condition along with different text stimuli. The Reading Tutor displayed the other 3 easy and 3 hard passages, again in an alternating order.

Similarly, the other 10 real and 10 pseudo-words, as well as the 10 non-words, were displayed one at a time, in a randomized order. Note that the read-aloud condition omitted the non-words because they are unpronounceable.

	Passages		Words			
	Easy	Hard	Easy	Hard	Pseudo	Non
Oral	3	3	5	5	10	N/A
Silent	3	3	5	5	10	10
Total	6	6	10	10	20	10

Table 2.2: Number of passages and isolated words read by each subject in each condition, in a single session.

2.4 Signal processing

Table 2.3 below shows the signals in the data set. The first 4 signals were computed and delivered by the Mindset device to the NeuroView software while both the filtered signal and power spectrum were computed offline using Matlab (version 7.12). "Rate" refers to the rate of reporting. Since there is no way to obtain the raw voltage signal from the Mindset, the term "EEG signal" is used throughout this thesis to refer to the processed voltage signal, in order to avoid confusion. The EEG signal was delivered in an analog digital conversion (ADC) unit and had to be converted to voltage with a formula provided by NeuroSky [NeuroSky, 2012a], before using it in the analyses.

Signal	Rate (Hz)	How it was obtained
EEG	512	Applied notch, low- and high-pass analog filters to the raw voltage signal (sampled at 512 Hz)
Quality	1	Proprietary techniques
Attention		
Meditation		
Filtered	512	Applied a Butterworth bandpass digital filter of 0.1–20 Hz to the processed EEG signal
Power spectrum of 1 Hz bands from 1 to 100 Hz	8	Applied short-time Fourier transform (STFT) to the processed EEG signal with a window size of 1 second and an overlap of 448 samples

Table 2.3: Signals in the data set and how they were obtained

The quality signal represents the level of noise, including muscle artifacts, present in the EEG signal. The value of each sample from the quality signal ranges from 0 to 200; 0 indicates an acceptable noise level while 200 indicates an absence of contact of the electrode with the skin. It must be emphasized that a value of 0 does not indicate a perfectly clear signal. In fact, it is possible that some muscle artifacts are not detected and hence not accounted for in the quality signal. The value of each sample from either the attention or meditation signal ranges from 0 to 100; 0 indicates that no actual value is available (perhaps due to excessive noise), while 100

indicates a strong presence of the respective mental state. The filtered signal and the power spectrum were obtained by applying the corresponding computation technique to an entire session of recording. A bandpass filter of 0.1–20 Hz was applied offline to remove high-frequency noise generated by non-neural activity. The upper limit of 20 Hz is chosen due to quantitative evidence from Whitham et al.'s [2007] finding that EEG frequencies above 20 Hz could be contaminated by electromyography (EMG), or muscle, activity. In general, the filtered signal is very similar qualitatively to the unfiltered EEG signal; the main difference is that the filtered signal appears smoother visually.

Table 2.4 below shows the definitions of the frequency bands used in this thesis. To obtain the alpha band, for example, we average the 1 Hz bands in the alpha frequency range across the frequencies.

Name	Frequency Range (Hz)
Delta	1–4
Theta	5–8
Alpha	9–13
Gamma	31–100

Name	Frequency Range (Hz)
Low Beta	14–16
Mid Beta	17–20
High Beta	21–30
Beta	14–30

Table 2.4: Definitions of the frequency bands.

To enable finer-grain analyses, we sub-divided the beta band in a fashion similar to how Rangaswamy et al. [2002] divide the beta band into three smaller bands with ranges of frequency very similar to those defined here. We also included the beta band itself as an additional source of information. According to Baumeister et al. [2008], beta activity is associated with anxious thinking and active concentration. Specifically, the low beta band is associated with a relaxed yet focused mental state, while the high beta band is associated with alertness. Despite the fact that the beta, high beta and gamma bands could be contaminated with EMG activity according to Whitham et al., we still used features derived from these bands. As we presume EMG activity to arise frequently in oral trials, the rhythmic activity in the beta and gamma bands could be indicative of the strength of EMG activity which could in turn be predictive of the difficulty of the stimulus, if for example, hard sentences cause less EMG activity to be generated. Indeed, another study by Whitham et al. [2008] suggests that fast rhythms in the gamma band are predominantly due to EMG activity. They observed higher power in the gamma band when healthy subjects performed mental tasks involving limb- or eye-movement than when the subjects performed other tasks; in contrast, paralyzed subjects had comparable power in the gamma band regardless of the type of task performed. Because paralyzed subjects generate little or no EMG activity at all, they concluded that while electrical rhythms in the gamma band are inducible by mental activity, such rhythmic activity are largely due to EMG unrelated to cognitive effort.

In the pilot study, Mostow et al. [2011] derived features from the EEG, attention, meditation and filtered signals and frequency bands. However, in order to make the results replicable and independent of the underlying hardware, this thesis does not derive any feature from the proprietary attention and meditation signals. This thesis merely uses the proprietary quality signal as an aid to reject low-quality EEG signals and to detect missing data, rather than to distinguish mental states. If the analyses and methods in this thesis were used for signals recorded by another type of hardware, we presume that there is a way to assess the quality of a signal, perhaps using a feature of the hardware itself.

2.5 Overview of data set

A *trial* refers to the display of a stimulus, either an individual sentence or an isolated word, by the Reading Tutor. The EEG and filtered signals and frequency bands for each trial were extracted from the respective subject's session. The NeuroView software recorded timestamps at a 1-second resolution and they had to be linearly interpolated to provide precise timestamps at the millisecond resolution for each discrete sample. Interpolation was done by dividing the time range for an entire session evenly by the number of samples. There is a slight loss of precision in timing for each sample but the loss is assumed to be negligible.

A trial was considered *good* if the quality signal indicated a value of 0 for the duration of the entire trial, that is, the level of noise present in the trial was acceptable; otherwise, the trial was excluded from analyses in this thesis. Though this measure ensures good trials are as clean as possible, they could still be contaminated as muscle artifacts might be considered as part of an acceptable level of noise. 3 adults and 1 child, out of the 10 adults and 11 children participants, were excluded from analyses because they had large amounts of missing or poor-quality data, as determined by consistently non-zero samples from the quality signal. Consequently, the analyses involved only the data for 7 adults and 10 children.

Table 2.5 below shows the distribution of trials across the subjects, reading conditions, and stimulus type. In each cell, the first number refers to the number of good trials, the second the number of rejected trials and the third the sum of the first two. Note that no statistic is shown for non-words in the oral reading condition as they are unpronounceable.

Oral Reading

	Sentence		Word			
	Easy	Hard	Easy	Hard	Pseudo	Non
Adults	189 / 70 / 259	61 / 39 / 100	38 / 12 / 50	39 / 11 / 50	75 / 25 / 100	N/A
Children	188 / 98 / 286	37 / 73 / 110	46 / 9 / 55	42 / 13 / 55	87 / 23 / 110	N/A

Silent Reading

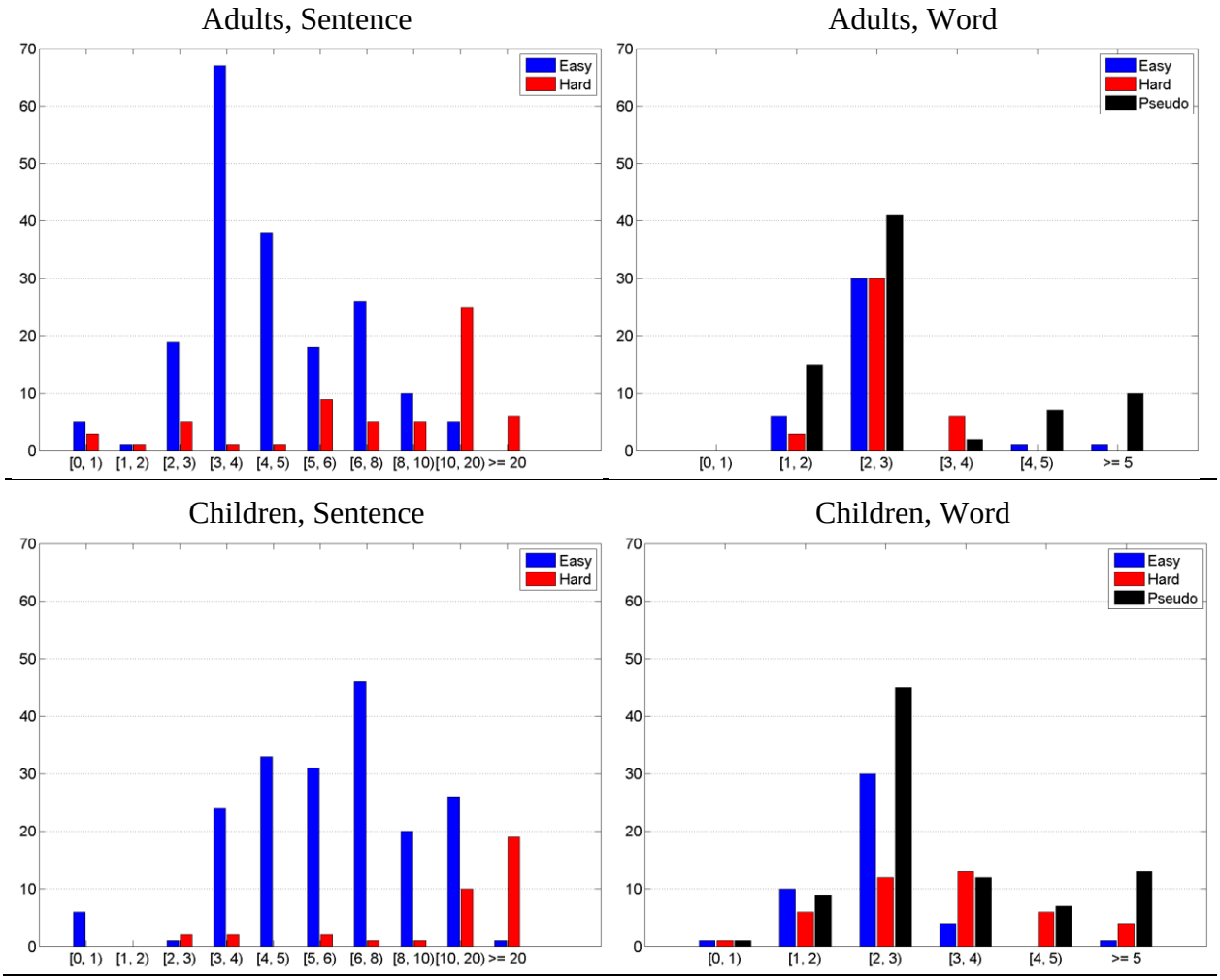
	Sentence		Word			
	Easy	Hard	Easy	Hard	Pseudo	Non

Adults	257 / 103 / 360	89 / 41 / 130	30 / 20 / 50	30 / 20 / 50	60 / 40 / 100	60 / 40 / 100
Children	289 / 84 / 373	89 / 45 / 134	31 / 19 / 50	34 / 16 / 50	71 / 29 / 100	72 / 28 / 100

Table 2.5: Distribution of good/rejected/total trials for oral and silent reading

Figure 2.1 below shows the distribution of duration (in seconds) of good trials for oral and silent reading conditions. The Y-axis in each histogram represents the frequency and the X-axis shows the intervals that the duration of each trial is grouped into. Adhering to conventional mathematical notation, "[0, 1)", for example, refers to the interval that includes 0 but not 1. Hard sentences lasted longer than easy sentences because they contained more words. It is evident from the skewed distribution of the duration of easy sentences that adults generally read faster than children. Subjects probably had to spend more time thinking about how to pronounce pseudo- and non-words, resulting in longer duration. The distributions for words, regardless of the reading condition, are similar for both adults and children.

Oral Reading



Silent Reading

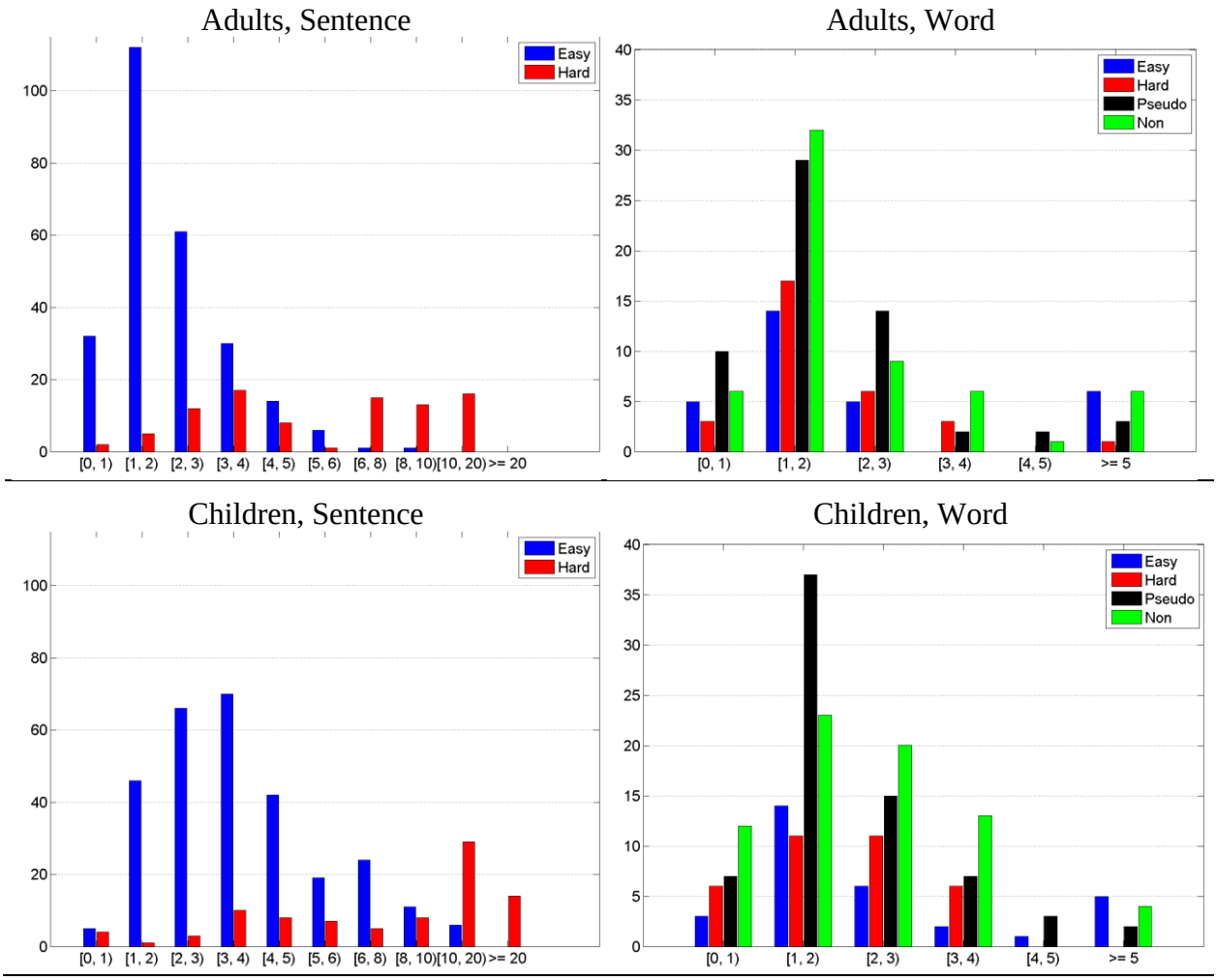


Figure 2.1: Distribution of duration (seconds) of good trials for oral and silent reading. The Y-axis represents the absolute frequency and the X-axis represents the intervals of duration in seconds.

Chapter 3

Methods

This chapter provides a detailed explanation of all methods and classifiers used in the analyses and experiments, as well as the reasons behind using those methods.

3.1 Artifact removal

It is widely recognized in literature that pervasive EEG artifacts associated with eye movements, cardiac signals, muscle noise, and electrical line noise pose a challenge for EEG interpretation and analysis. Furthermore, this thesis deals with EEG signals recorded in an informal, rather than a tightly-controlled and shielded laboratory environment. Measures designed to minimize eye moments and muscle artifacts are typically employed in experiments in a laboratory. For example, in Dambacher et al.'s study [2006], participants were instructed to look at the "fixation cross" that would be displayed on the computer screen right before each stimulus; all stimuli were centered on screen to minimize eye moments as much as possible. On the other hand, in an informal setting, it would be difficult to implement such measures in practice and to ensure participants follow instructions exactly, especially without any supervision, as in the case of using the Reading Tutor. EEG signals recorded in an informal setting thus have a greater chance to be contaminated with artifacts. The mechanism built into the Mindset device to gauge the quality of the EEG signal might not be entirely fool-proof, allowing contaminated signals to pass off as being clean. Also, rejection of contaminated trials leads to undesirable data loss as shown by section 2.5. It is thus of paramount importance to remove undesirable artifacts. While artifact removal methods can be applied to EEG signals recorded during silent reading, they most likely would not work as well on signals recorded during oral reading, since such signals would presumably be full of muscle artifacts. Thus, artifacts would be left intact, embedded in the EEG signals and be used in the analyses in this thesis.

3.1.1 Blind source separation

Blind source separation (BSS) methods are commonly used to remove artifacts from multi-channel data. BSS methods assume each recorded channel is a mixture of uncorrelated spatial source components. Source separation is best illustrated with the "cocktail party problem", in which multiple microphones are placed at various locations in the room to record the voices in the party. The sound signal captured by each microphone is a mixture of voices contributed by the participants in the room. BSS attempts to separate the recorded channels (sound signals) into the individual source components (voices). A survey of relevant literature reveals independent component analysis (ICA) and principal component analysis (PCA) as the two most popular BSS methods (e.g., Jung et al. [1998] and Wallstrom et al. [2004]). PCA decomposes the channels into uncorrelated components that are spatially orthogonal but not necessarily mutually

statistically independent. In contrast, ICA does not impose the orthogonality requirement but does form components that are mutually statistically independent.

The usual workflow of removing artifacts can be summarized in three steps. First, apply a BSS method to decompose recorded channels into source components. Second, remove source components contaminated with artifacts. Identification of artifactual components typically requires manual visual inspection for anomalies (e.g., amplitude exceeding a threshold), though automatic methods have been developed (e.g., Joyce et al. [2004] and Wallstrom et al. [2004]). Third, recombine the uncontaminated source components to form a clean version of the recorded channels. Wallstrom et al., motivated by the fact that ocular (eye) activity creates significant artifacts in EEG signals [Fisch, 1991], developed a method that relies on signals recorded by electrooculography (EOG) electrodes which can be placed above and/or below the eyes to monitor eye movements. Their method automatically and effectively selects principal components contaminated by ocular artifacts, as determined by the degree of correlation with observed EOG signals.

Unfortunately, since BSS methods require multiple channels as input, it is inappropriate to apply these methods directly to single-channel data without any modification to these methods. Furthermore, it is difficult to remove ocular artifacts reliably without EOG. The bandpass filter is probably the only appropriate method to apply, though it can only systematically remove high-frequency components, artifacts or not.

3.2 Event-related potential (ERP)

The conventional method to measure an ERP component is to measure the strength of the observed deflection, which in turn is usually determined as the mean amplitude of the EEG signal in the time window containing the component [Dambacher et al., 2006; Friederici et al., 1993; Laszlo and Federmeier, 2011]. An appropriate time window is determined visually from grand averages produced by averaging trials together. A large enough number, usually at least a hundred trials, are needed so that random noise is canceled out, leaving behind mostly ERP components. In many studies, various time windows are considered, even for the same ERP component such as N400. For example, Laszlo and Federmeier used the window 0.25–0.45 second for N400 while Dambacher et al. used 0.3–0.5.

3.2.1 Analyses of variance (ANOVA)

The ANOVA method is commonly used to analyze the effect of factors on the mean amplitude of an ERP component. Examples of a factor are the electrode location and the orthographic neighborhood size. The simplest form of ANOVA is the one-way model with a single independent variable (IV) which can take on two or more levels (or discrete values) [NIST, 2012]. In particular, a one-way model with an IV that has fixed levels, instead of randomly sampled ones, is also known as a fixed-effects model. Only fixed-effects models are used in this thesis. The one-way model determines through an F test whether the IV has an effect on a dependent variable (DV). The null hypothesis is that two or more sample groups, each labeled

with a different IV level, have an equal mean. To illustrate how this model works, suppose the IV is the type of a word stimulus, easy, hard, pseudo- or non-word and the DV is the mean amplitude of N400 in each trial for a predefined time window. Each sample group is associated with a unique word type and contains the mean amplitudes of N400 induced by word stimuli of the corresponding type.

More IVs can be added to extend the one-way model. A two-way model, for instance, has an additional null hypothesis, besides the usual null hypotheses on the individual IVs, that there is no interaction between the two IVs. An interaction between two IVs suggests that the effect of one depends on the level of the other. Interested readers can refer to NIST [2012] for detailed information on ANOVA.

3.2.2 Multiple comparison

The one-way ANOVA model simply tests the null hypothesis that all sample groups have the same mean. Rejection of the null hypothesis indicates that at least one of the groups has a mean different from the others; in other words, there is an overall effect from the IV. In such a case, it is necessary to conduct post-hoc tests to find out exactly which pairs of sample groups have means that differ from each other within the pair. The intuitive method of performing a regular *t*-test for each possible pair of the sample groups tends to artificially increase the chances of discovering significant results [Abdi, 2007]. A *multiple comparison* method is thus needed for such simultaneous comparisons. The Tukey-Kramer method is the recommended post-hoc test in an imbalanced one-way ANOVA design if the sample groups have homogenous (equal) variances [Stoline, 1981]. An imbalanced design consists of sample groups with different sizes, which is the case for the analyses in this thesis. The Tukey-Kramer method is proven to be conservative in imbalanced ANOVA designs, that is, it has a low risk of incorrectly deducing a difference when in fact there is no significant difference. However, if the variances are heterogeneous, we use another method involving a Bonferroni correction [Abdi, 2007] instead. Levene's test [Levene, 1960] is used to test if sample groups have equal variances because it is less sensitive to departures of sample groups from normality and hence more robust, when compared to Bartlett's test [Snedecor and Cochran, 1989].

3.3 Feature groups

This section describes various groups of features that could be extracted from the EEG signal for sentence trials.

3.3.1 Pilot study baseline

The features used in the pilot study are the mean amplitudes of the EEG, filtered, meditation and attention signals, as well as the mean power of each 1 Hz band from the power spectrum (1–100 Hz). The features are computed across the duration of a trial without regard to any temporal structure. In this thesis, we use all features except those derived from the proprietary meditation and attention signals to compute the baseline results.

3.3.2 Mean EEG amplitude

In this thesis, the mean amplitude is the simplest feature of all. We compute the mean amplitude across all sample points in the EEG and filtered signals as features. Thus, a trial will have two mean amplitude features, one each for the EEG and the filtered signals.

3.3.3 Mean power

The power of a frequency band is computed from the complex-valued Fourier coefficients which in turn are obtained from the Fourier transform represented by the following:

$$f_k = \int_{-\infty}^{\infty} x(t) e^{-2\pi i t k} dt$$

where *t* is time, *k* is the desired frequency in Hz, *i* is the imaginary unit, and *x(t)* is the value of the continuous signal at time *t*. But since the EEG signal is digitized, we used the discrete version of the Fourier transform, specifically the fast Fourier transform (FFT) implementation in Matlab (version 7.12).

As the Fourier transform is formulated for a stationary signal, that is, a signal that has unchanging frequency components, the time domain is absent. To compensate for the absence of the time domain, we applied the common practice of computing the Fourier coefficients for overlapping windows of EEG samples, a method also known as the short-time Fourier transform (STFT). STFT is applied to a continuous EEG signal recorded in a session. We computed the coefficients for each 1 Hz of frequency from 1 to 100 Hz. We chose 100 Hz as the upper limit since the low-pass filter applied to the raw voltage signal already discarded frequencies above 100 Hz. The window size is 512 samples with an overlap of 448 samples, yielding a sampling rate of 8 Hz for each 1 Hz frequency band. In comparison, the EEG signal has a sampling rate of 512 Hz.

Given a Fourier coefficient *a+bi*, the power is calculated as $|a + bi|^2/512 = (a^2 + b^2)/512$. The mean power of an 1 Hz band is then the average power over the entire trial. To compute for example the mean power of the alpha band, the power in each 1 Hz band in the alpha frequency range is first averaged across each individual band, before averaged over the entire alpha band. The mean power of a frequency band measures the average activity in that band.

3.3.4 Frequency ratio 1

The frequency ratio 1 of, for example, alpha and beta, is the mean power of alpha divided by that of beta. The mean power is computed across the entire trial. More precisely,

$$Ratio1(x, y) = \frac{\sum_{t=1}^{n} x(t)}{\sum_{t=1}^{n} y(t)}$$

where *x(t)* and *y(t)* are power samples of two frequency bands at time *t*, and *n* is the number of samples.

This feature group is useful for measuring the overall average activity in a band with respect to that of another. For example, as the subject pays more attention to a stimulus, his or her beta band activity might increase while the alpha band activity decreases.

3.3.5 Frequency ratio 2

The frequency ratio 2 is a more sensitive version of ratio 1. Ratio 2 of, for example, alpha and band, is the mean ratio of alpha to beta power at the sample level, across time. More precisely,

$$Ratio2(x, y) = \frac{1}{n}\sum_{t=1}^{n}\frac{x(t)}{y(t)}$$

where the notations are the same as before.

3.3.6 Correlation coefficient

The Pearson product-moment correlation coefficient is a measure of the linear dependence (correlation) between two variables. We use this to measure the correlation between frequency bands, for example, between alpha and beta. The value of the coefficient itself is used as a feature; it ranges from -1 to 1 with both limits indicating perfect negative and positive correlation, respectively, and 0 indicating no correlation. With two groups of *n* samples each for variables *X* and *Y*, the sample correlation coefficient *r* is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The sampling distribution of *r* is not normal. Fisher's *z*, also known as Fisher transformation [Fisher, 1915; Fisher, 1921], is used to convert the coefficient *r* to an approximately normally distributed variable *z*, via the formula,

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{arctanh}(r)$$

where ln and arctanh are the natural logarithm and inverse hyperbolic functions, respectively.

z is only approximately normally distributed only if *X* and *Y* follow a bivariate normal distribution, and the (X_i, Y_i) pairs in the formula for *r* are independent. Fisher transformation is particularly useful when there is a need to perform hypothesis testing on samples of *r* using for example, the *t*-test which requires test samples to be normally distributed in the first place. Fisher transformation can also be used for constructing confidence intervals on *r*. Liang et al. [2006] also used the Fisher transformation before using the correlation coefficients for *t*-tests.

3.4 Temporal methods

The duration of the EEG signal recorded for a stimulus varied widely across subjects due to differences in reading abilities such as reading speed. For example, hard sentences usually resulted in longer EEG signals due to them containing more (and difficult) words. The varying durations of sentences require techniques that can deal with such differences, but at the same time, not exploit the duration property of EEG signals, since duration is highly correlated with the difficulty of the sentence. Thus, designing good temporal techniques to effectively capture the brain responses to words in a sentence can be challenging, especially for silent sentence trials, since it is difficult to know which word the user is reading at a given time. Capturing the brain response to each individual word in an oral sentence trial is possible with the aid of recorded speech from the user. But even so, the utterance of a word could be as short as a few hundred milliseconds, too short to even allow capturing a brain response such as N400. Nevertheless, this method of using recorded speech might still be possible but we shall leave it as future work.

3.4.1 Relative segmentation

Relative segmentation refers to segmenting the EEG signal or frequency band for a trial into equal parts such that each part has an equal duration. It is termed as "relative" since each part would have a duration relative to the total duration of a trial. This method is convenient for dealing with the widely varying durations of sentence trials. In particular, we consider 3-segmentation, in which the EEG signal is segmented into 3 equal parts. Using 3-segmentation is advantageous as the parts can be easily operationalized as the start, middle and end of a trial, and vice versa. Features can then be computed for each part separately, allowing us to investigate which part is useful for distinguishing between easy and hard sentences.

3.4.2 Truncation

Another way to deal with the varying durations is to truncate the trial, that is, to use information only from the first few seconds. This method also allows us to examine how much information since the start of a trial would be needed to distinguish between easy and hard sentences. For example, if using only the first second of a trial could produce a high classification accuracy, there would be no need to use information beyond that.

3.5 Classification

After having discussed the features and temporal methods, we next describe our classification methodology which is based largely on that of the pilot study.

3.5.1 Logistic regression

We chose regularized logistic regression as it reduces the risk of overfitting the training data, which is important when each training data point is represented by a high-dimensional vector of features. Overfitting is undesirable as it increases the generalization error, that is, it decreases the performance of the trained classifier on unseen test examples. In particular, ℓ_1 regularized logistic regression is chosen over ℓ_2 because it has been shown to be superior to ℓ_2 [Ng, 2004]. As proven by Ng, ℓ_1 regularized logistic regression has a logarithmic sample complexity,

meaning that the classifier can still effectively perform feature selection even if there are exponentially more irrelevant features than training examples. This property is certainly beneficial for exploring the usefulness of many features in our small pilot data set.

3.5.2 Cross-validation

Classification accuracy is computed as the percentage of trials classified correctly; chance performance is one over the number of classes. Classifiers are trained separately, and their classification accuracies reported, for each combination of subject group (adults, children or both) and reading condition (oral or silent). Two types of classifiers are trained and tested in a leave-one-out cross-validated fashion. *Reader-specific* classifiers are trained on a subject's data from all but one trial (sentence or word), and tested on the held-out trial. Each trial is held out in turn until each one is tested exactly once. As noted by Mostow et al. [2011], cross-validating across stimuli (e.g., passages) with multiple successive observations (e.g., sentences) avoids improper exploiting of statistical dependencies, such as temporal continuity, between observations of a subject on the same stimulus. The classification accuracy for a subject is the percentage of trials from that subject classified correctly. The overall classification accuracy is then computed as the average across all subjects. *Reader-independent* classifiers are trained on the data from all but one subject, and tested on the held-out subject. Again, each subject is held out in turn, and the resulting accuracies are averaged to cross-validate across subjects.

3.5.3 Classification accuracy

One problem in computing the classification accuracy is class size imbalance, as evident in section 2.5. This issue arises because there are more easy sentences than hard ones and more non-words than real words. It is made even worse by discarding low-quality trials. Consider the following scenario in which 10 easy and 5 hard sentence trials from a subject are tested in total. If there is a bias in the classifier, causing it to predict all trials as easy, the classification accuracy would be computed as $\frac{10}{15}$ = 66.7%, which is artificially high considering the fact the classifier did not predict any as hard at all. To deal with this problem, we report the average per-class accuracy. Suppose *x* out of 10 easy trials and *y* out of 5 hard ones are predicted correctly. The average per-class accuracy would then be $\frac{1}{2}(\frac{x}{10} + \frac{y}{5})$.

3.5.4 Rank accuracy

Rank accuracy is used in place of classification accuracy for words where appropriate, for example, when the multinomial logistic regression classifier is used. Rank accuracy is evaluated as "the average percentile rank (normalized between 0 and 100) of the correct category if categories are ordered by the value of the regression formula" (or the category likelihood) [Mostow et al., 2011]. It is a more sensitive criterion than classification accuracy for evaluating performance on multi-category tasks such as decoding mental states from brain data [Mitchell et al., 2004]. Chance performance is 50% regardless of the number of categories. Like the classification accuracy, we also report the average per-class rank accuracy.

3.5.5 Resampling

Class size imbalance also affects the training of classifiers besides classification accuracy. A common solution is to resample the training data to obtain equal-size sets of training data. There are various resampling methods, two of which are random undersampling and random oversampling. In the context of binary classification, random undersampling refers to selecting a subset of data points randomly from the larger class so that the subset has the same size as the smaller class; both the subset and the smaller class can then be used for training. In contrast, random oversampling refers to selecting data points randomly with replacement from the smaller class so that the set of selected data points has the same size as the larger class.

Undersampling and oversampling have their disadvantages. "Random undersampling can potentially remove certain important examples, and random oversampling can lead to overfitting" [Chawla et al., 2004]. Another undersampling method, *truncation,* refers to reducing the larger class to the temporally earliest *k* data points. All three resampling methods were used in the pilot study. However, we only use random undersampling in this thesis for three reasons. First, adhering to only one resampling method reduces the number of results so as to simplify analyses. Second, though oversampling balances class size, it is unclear what effects the duplicated data points would have in training a logistic regression classifier. Third, random undersampling is arguably similar to truncation, and in fact, random undersampling might be fairer as it has no temporal bias.

3.5.6 Significance

To determine whether a classifier performed significantly better than chance, the overall classification or rank accuracy for each subject is first computed to yield a distribution of *N* accuracies where *N* is the number of subjects. Treating this distribution as a random variable, the hypothesis that the mean of this distribution exceeds chance performance is then tested using a one-tailed *t*-test at the 95% confidence level. Counting *N* subjects instead of observations is conservative in that it accounts for statistical dependencies among observations from the same subject [Mostow et al., 2011].

3.5.7 False discovery rate

We use the False Discovery Rate (FDR) [Benjamini and Hochberg, 1995] method to account for multiple comparisons, for example, when multiple *t*-tests are performed to check for the significance of classification results.

Let $P_1 \leq P_2 \leq \ldots \leq P_m$ be the sorted *p*-values yielded by the *m* corresponding null hypotheses, H_1, $H_2, \ldots, H_m$. Then let *k* be defined as,

$$\mathrm{k} = \underset{i}{\operatorname{argmax}} \left\{ P_i \leq \frac{i\alpha}{mC_m} \right\}$$

where *i* is an integer from 1 to *m,* and α is the significance level, 0.05 in our case. If the hypotheses being tested are all independent, C_m = 1. But since there are usually arbitrary

dependencies amongst the hypotheses being tested, we use $C_m = \sum_{i=1}^{m} \frac{1}{i}$ which is the strictest form of FDR procedure available. If *k* exists, reject all hypotheses which have *p*-values smaller than or equal to P_k; otherwise, reject no hypothesis.

Chapter 4

Word Analyses

ERP studies aim to discover any significant deflection in averaged EEG signals. In order to unmask any potential deflection, it is important that the signals be free from muscle and ocular artifacts and are as clean as possible. Since trials in the oral reading condition are presumably full of muscle artifacts, if a deflection does occur in the EEG signal, it would be near impossible to attribute its source to muscle movement or coordinated neural activity. Therefore, it would be advantageous to study averaged trials in the silent reading condition instead. With less muscle artifacts in silent trials, ocular artifacts such as eye blinks could still pose a problem, since they could be mistaken for an ERP component due to their amplitudes being larger than that of regular EEG signals. However, eye blinks should have occurred rarely during a short word trial, especially during the first second of stimulus onset. Moreover, the effect of such a rare occurrence would be reduced, if not canceled out, when averaging across trials. This chapter only considers word trials since ERP components are unlikely to be discovered in sentence trials. The onset of each word in a sentence should be precisely time-locked to the EEG signal, so that the part of the signal associated with each word could be isolated and averaged separately, allowing any ERP component to survive the averaging process. An ERP component might arise in between two words, but again, without time-locking, it would unlikely survive averaging. For silent sentence trials, it is challenging to do the required time-locking as it is difficult to know precisely what word the subject was reading at a given time, since each sentence was presented entirely at once. We could track eye movements to gauge which word the subject is reading at a given time, but we shall leave it as future work. Also, we could time-lock the EEG signal to each word in oral sentence trials with the aid of recorded speech from the subject, but again, it is beyond the scope of this chapter.

Averaging across trials is tricky due to the wide range of possible duration each trial has. As Figure 2.1 in Chapter 2 shows, even an isolated word trial can last from less than 1 second to more than 5 seconds. Intuitively, a word trial should be short since it should not take that long to read a word. Many related studies investigating N400, some of which discussed earlier in Chapter 1, considered only the first second of the EEG signal after the onset of the stimulus; this is in line with the generally accepted fact that semantic processing happens within the first second. To ensure the analyses are as reliable as possible, trials lasting beyond 5 seconds are deemed unsuitable for use. One reason is that there might be a delay after the stimulus was shown, but before the subject started to perceive it, in which case the first second of the EEG signal would not contain information relevant to processing the word. Another is that a subject might have spent some time thinking about an unfamiliar pseudo- or non- word after he or she was exposed to it, before moving on to the next word (recall that subjects controlled when the

next word was shown); in this case, the EEG signal might not have relevant information in the first second.

Since it is difficult to verify which if either reason is correct, and in fact, both could apply to some instances, word trials of similar duration are grouped together to ensure the ERP investigation is as reliable as possible. More specifically, they are divided into three overlapping groups according to their duration. A trial in the first group has duration between 0.6 and 5 seconds and in the second, between 1 and 5 seconds. The third group has no restriction on the duration and holds all word trials, and exists for comparison. Note that a trial can be in both the first and second groups as long as it satisfies the criteria. The upper limit of 5 seconds is chosen as majority of the trials lasted between 0 and 5 seconds; 1 second is too stringent as it will discard most of the trials. The lower limits of 0.6 and 1 second are minimum criteria for the mean amplitudes of the N400 peak and the post N400 effect, respectively, to be computed. In here, the post N400 effect refers to the part of the EEG signal immediately after the occurrence of N400.

4.1 Grand averages

Figure 4.1, Figure 4.2 and Figure 4.3 below show the grand averages of trials from the first (0.6–5 second) and third (all) group for each word type, pooled by adults, children or both, in the silent reading condition. The grand average for the third group is shown as a grey dashed line in each plot, imposed with the grand averages from the first group. Grand averages of trials from the second group are omitted for brevity as they are similar to that of the first group, with near perfect correlation coefficients of 0.98 or 0.99 within the first second, regardless of the word type and the pooling method. However, the correlation coefficients of the grand averages between the first and third or between the second and third groups range from only 0.4 to 0.7. Evidently, the addition of trials longer than 5 seconds distorts the grand averages, suggesting that such trials should not be used for analyses. Despite the relatively poor correlation, the grand averages for the third group have similar waveforms as for the first, as shown in the plots below.

Each grand average is obtained by collapsing across the bandpass-filtered signals corresponding to the trials. Using filtered signals is a common practice to eliminate high-frequency noise, for example, it was done so in Laszlo et al.'s study [2011]. At each time step, the sum of the sample points across the trials is divided by the number of trials that contributed to that time step. The number of trials remains the same for each time step up till 0.6 second in the first group, and decreases very gradually up till 1 second. The peak amplitude of the grand averages at about 12 μV is larger than that found in many studies, except for Luo et al.'s whose dry sensor device also recorded ERP components peaking at the same order of magnitude of about 10 μV. As a comparison, the peaks as shown below are about 2 times stronger than the peaks of 5 μV found in Laszlo et al.'s study.

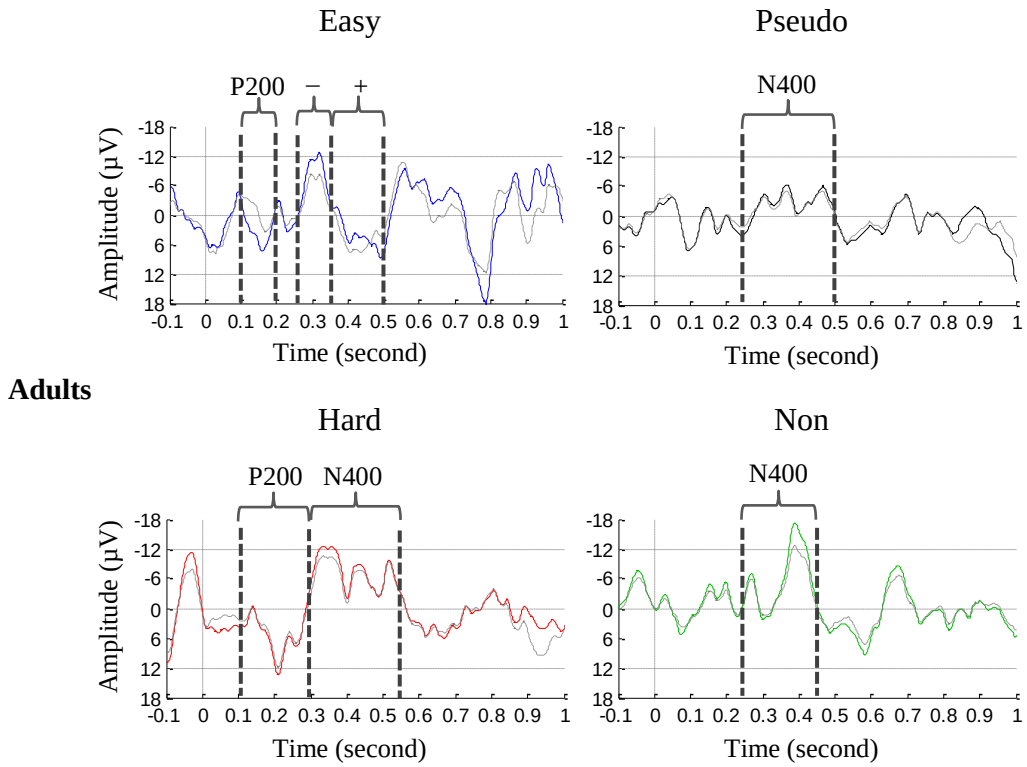


Figure 4.1: Grand averages of silent word trials, pooled by adults, in the silent reading condition. The vertical axis (inverted) shows the amplitude in micro-volts while the horizontal axis shows the time in seconds. The vertical dotted line at time 0 marks the onset of the stimulus. + and – indicate positive and negative deflections, respectively.

Notice that there are no clear-cut ERP components. Despite the amount of noise still present in the grand averages, several potential ERPs can still be observed. For hard words pooled by adults, there is a possible P200 between 0.1 and 0.3 second, and a possible N400 between 0.3 and 0.55 second. For easy words, a potential but weak P200 can be observed quite clearly between 0.1 and 0.2 second; there are also two deflections between 0.25 and 0.35 second and between 0.35 and 0.5 second. A potential N400 could exist between 0.25 and 0.5 second for pseudo-words. Another possible N400 can also be observed between 0.25 and 0.45 second for non-words.

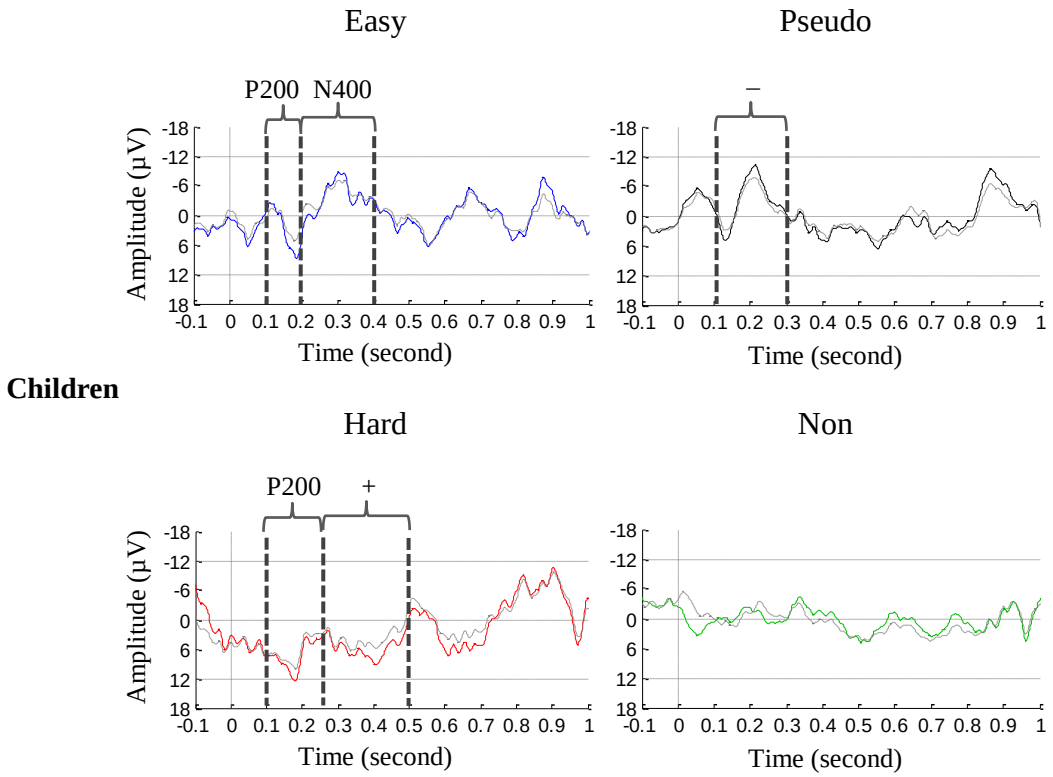
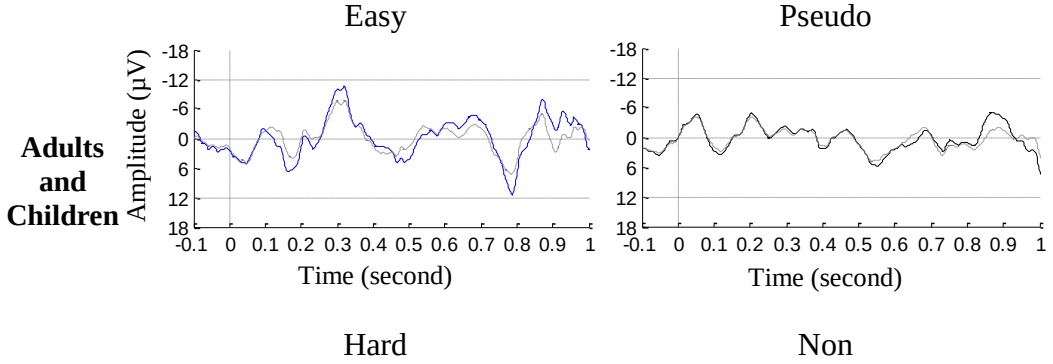


Figure 4.2: Grand averages of silent word trials, pooled by children, in the silent reading condition.

When pooled by children, N400 can be seen between 0.2 and 0.4 second for easy words, and a potentially weak P200 might exist between 0.1 and 0.2 second. It is worth noting that in adults or children, the grand average for easy words looks similar. There is a positive peak right before 0.2 second, a strong negative peak at 0.3 second, and finally a positive peak again at 0.55 second. For hard words, a possible P200 as well as an unexpected positive deflection is seen between 0.1 and 0.25 second and between 0.25 and 0.5 second, respectively. For pseudo-words, there is a negative deflection between 0.1 and 0.3 second. The grand average for non-words seems random. The absence of N400 could be due to the lack of lexico-semantic processing required to induce N400.



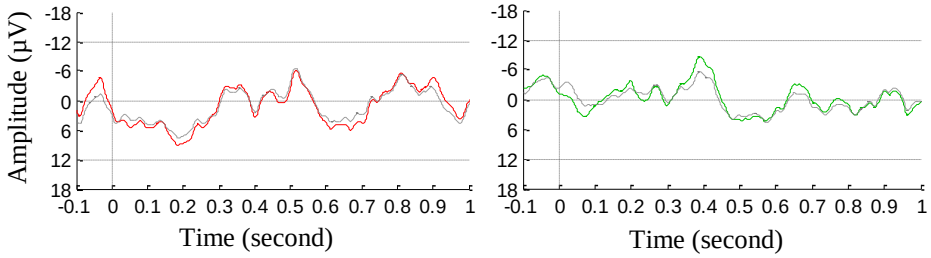


Figure 4.3: Grand averages of silent word trials, pooled by both adults and children, in the silent reading condition.

When pooled by both adults and children, N400 survived, albeit weakened, for hard and non-words. The N400 in the grand average for hard words is weakened by the positive deflection produced by children in the window between 0.25 and 0.5 second.

A possible reason for the weak N400 observed is that the Mindset sensor was placed at the frontal position (Fp1) where the N400 effect might be weak, when N400 has in fact maximal effect at centro-parietal sites (e.g. Pz). In addition, the low signal to noise ratio, coupled with the possible lack of enough trials, makes both P200 and N400 difficult to see. N400 might even be absent for non-words, since according to Laszlo et al., the presence of N400 for non-words depends on the sentence context. In the pilot experiment, all of the words were shown individually and were not part of any sentence. Furthermore, as Figure 4.4 below illustrates, the mean amplitude of N400 decreases with increasing orthographic neighborhood size. The mean amplitude peaks at a neighborhood size of 0 and decreases gradually as the size increases to 25. Since the neighborhood size of pseudo- and non-words in the stimuli set varied from 0 to 22, it is harder to observe the maximal effect of N400 in the pilot data set.

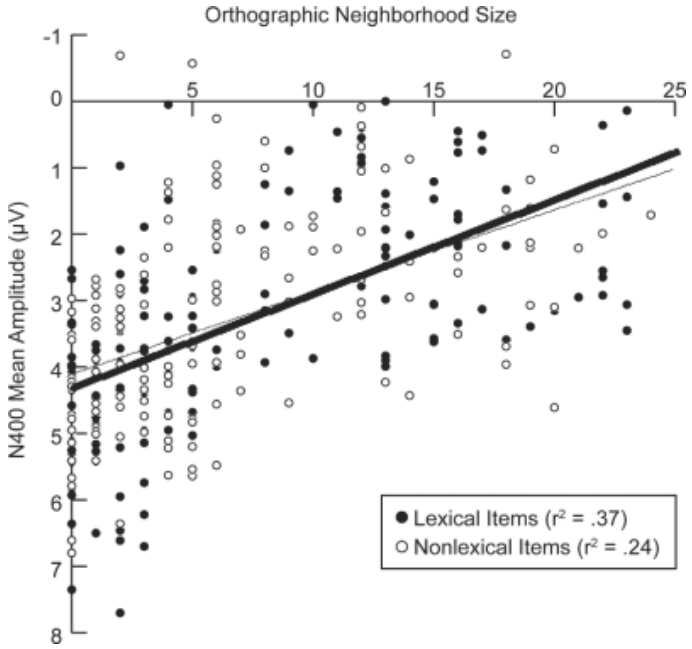


Figure 4.4: The relationship between N400 mean amplitude (from the window 0.25–0.45) and orthographic neighborhood size (N). Plot is taken from Laszlo et al.'s study. Lexical items (pseudo-words) are represented by filled circles and non-lexical items (non-words) by empty circles. Single regression trend lines for the relationship between N400 mean amplitude and N are also plotted for each item type. The function relating N400 amplitude to N is nearly identical for the two item types.

4.2 Existence of ERPs

The next step after observing the grand averages is to check for the existence of ERPs. We first establish a baseline to which the mean amplitudes of ERPs will be compared to. Then we define time windows to represent the ERPs in each word type.

4.2.1 Pre-stimulus baseline

We cannot apply the standard practice of subtracting the 0.1 second pre-stimulus mean amplitude from the signal for each trial here, due to the amount of noise present in the pre-stimulus window as evident even in the grand averages. Subtracting the non-zero mean amplitude of noise from the signals could negatively affect the ERP analyses, causing real ERP components not to be recognized and vice versa. The pre-stimulus noise could have arisen from random background brain processes, the effects of the previous stimulus seen by the subject, or the clicking of the mouse to advance to the next word. (Recall that there is no resting break between stimuli.) We hypothesize that the noise ideally cancels out so that the pre-stimulus baseline is approximately zero. We use the value of zero as many studies such as Laszlo et al.'s show a pre-stimulus baseline of zero or near zero in the grand averages, although the quantity is usually not explicitly presented.

Since the amplitude of the pre-stimulus window does not depend on the stimulus itself, to test the hypothesis that the pre-stimulus baseline is approximately zero, we performed a *t*-test on the mean amplitudes in the pre-stimulus window pooled from across all word types and subject groups. The test confirms the mean amplitude in the pre-stimulus window does not deviate significantly from zero ($p = 0.874$, mean = -0.01).

4.2.2 Results

To test for the existence of ERPs, the mean amplitude across the sample points in the following windows is computed for each trial: from 0 to 0.1 for pre-P200 effects, 0 to 0.2/0.25/0.3 for pre-N400 effects, 0.1 to 0.2/0.25/0.3 for P200, 0.2 to 0.4/0.5, 0.25 to 0.45/0.5 and 0.3 to 0.55 for N400, and 0.4/0.45/0.5/0.55 to 1 for post-N400 effect. Other windows to test for observable, non-standard deflections include 0.25–0.35 and 0.35–0.5. These windows are defined according to the visual observations described earlier. In each window, we performed a one-sample two-tailed t-test to validate the null hypothesis that the mean amplitudes from the trials in that window are drawn from a normal distribution of zero mean. If the null hypothesis is rejected, the mean amplitudes are drawn from a distribution with non-zero mean. We used the baseline of zero mean as discussed in the previous subsection.

Table 4.1 below shows the results of the one-sample two-tailed *t*-tests on the mean amplitudes in the pre-defined windows. For brevity, only significant as well as some interesting insignificant results are shown. When accounting for multiple comparisons for the entire table, none of the results are significant. Hopefully, with more data in the future, the grand averages would show deflections sufficiently clear that the *t*-test results would be significant even after accounting for multiple comparisons.

A general observation is that significant deflections present in adults' EEG are absent in the children's, and vice versa, for each specific combination of window and word type. Though not tested, one possible explanation is that the cognitive processes in the children subjects differed sharply from that of the adults, perhaps due to differences in reading and language abilities, hence resulting in EEG with different characteristics.

On the adults' side, the results in windows 0.10–0.30 and 0.30–0.55 show that P200 and N400 are present for hard words. Furthermore, there is no pre-P200 and post-N400 effect. For non-words, N400 exists in the windows 0.20–0.40 and 0.25–0.45, though the latter window has a larger mean amplitude (-4.2 vs -6.1). For pseudo-words, N400 has a marginal presence in the window 0.20–0.50 (p = 0.058), and not shown below, in 0.25–0.50 (p = 0.051); a post-N400 effect also exists in 0.50–1.00. The post-N400 effect is not uncommon as shown in many other studies, but since it is usually not concretely quantified in existing literature, a comparison cannot be made here. However, it can still be used as a feature in classification. Overall the results agree with the visual observations.

There are fewer significant results for children. A negative deflection is unexpectedly present in place of P200 within the window 0.10–0.25 for pseudo-words. A positive deflection is also unexpectedly present in place of N400 within the window 0.30–0.55 for pseudo-words. There is no post-N400 effect. N400 almost exists for easy words within the window 0.25–0.45 (p = 0.069).

Effect	Window	Word Type	Pool	p	95% C.I.		Mean	Trials
P200	0.10–0.20	Hard	Adults	0.195	-2.4	11.0	4.3	26
			Children	0.087	-1.3	18.7	8.7	27
			Both	**0.030***	0.6	12.4	6.5	53
		Non	Adults	0.182	-5.9	1.2	-2.4	46
			Children	0.085	-6.3	0.4	-2.9	55
			Both	**0.028***	-5.1	-0.3	**-2.7***	101
	0.10–0.25	Hard	Adults	0.103	-1.2	12.7	5.7	26
			Children	0.195	-3.9	18.1	7.1	27
			Both	**0.047***	0.1	12.8	6.4	53
		Pseudo	Adults	0.328	-1.6	4.7	1.6	53
			Children	**0.040***	-6.3	-0.2	**-3.2***	57
			Both	0.415	-3.1	1.3	-0.9	110
	0.10–0.30	Hard	Adults	**0.045***	0.1	9.5	4.8	26
			Children	0.281	-5.5	18.2	6.3	27

			Both	0.078	-0.6	11.8	5.6	53
		Pseudo	Adults	0.453	-1.5	3.3	0.9	53
			Children	**0.032***	-5.4	-0.3	**-2.8***	57
			Both	0.255	-2.8	0.8	-1.0	110
N400	0.20–0.40	Non	Adults	**0.038***	-8.3	-0.2	-4.2	46
			Children	0.319	-2.1	0.7	-0.7	55
			Both	**0.021***	-4.3	-0.3	-2.3	101
	0.20–0.50	Pseudo	Adults	0.058	-3.7	0.1	-1.8	53
			Children	0.862	-2.6	3.0	0.2	57
			Both	0.383	-2.4	0.9	-0.8	110
	0.25–0.45	Easy	Adults	0.490	-8.3	4.1	-2.1	22
			Children	0.069	-14.7	0.6	-7.0	22
			Both	0.059	-9.4	0.2	-4.6	44
		Non	Adults	**0.020***	-11.2	-1.0	-6.1	46
			Children	0.466	-2.2	1.0	-0.6	55
			Both	**0.016***	-5.6	-0.6	-3.1	101
	0.30–0.55	Hard	Adults	**0.014***	-13.3	-1.6	-7.5	26
			Children	0.410	-5.9	14.1	4.1	27
			Both	0.590	-7.5	4.3	-1.6	53
		Pseudo	Adults	0.277	-4.4	1.3	-1.5	53
			Children	**0.047***	0.1	5.8	**2.9***	57
			Both	0.466	-1.3	2.8	0.8	110
Post-N400	0.40–1.00	Pseudo	Adults	**0.003***	0.5	2.1	1.3	43
			Children	0.809	-3.0	3.8	0.4	51
			Both	0.392	-1.1	2.7	0.8	94
	0.50–1.00	Pseudo	Adults	**0.001***	0.9	3.4	2.1	43
			Children	0.948	-3.9	3.6	-0.1	51
			Both	0.390	-1.2	3.0	0.9	94

Table 4.1: One-sample two-tailed t-test results. *p*-values < 0.05 are highlighted in bold with an asterisk. Mean values that have an unexpected sign are bolded with an asterisk.

4.3 ANOVA analyses

Even though the results in the previous subsection indicate which ERPs are present, further tests are still needed to investigate the effectiveness of using the ERPs to classify a trial according to the word type. We performed a 2-way between-subjects ANOVA test to find out the effect of window and word type on the mean amplitudes in each window, for all word trials. We also performed a 3-way between-subjects ANOVA test by including an additional factor of subject group (adults or children) to investigate any difference between children and adults. Table 4.2 below shows the results. We are particularly interested in the interaction between word type and window since this would give an indication of how feasible it is to detect the mental states associated with each word type. We hypothesized that there is a significant interaction between window and word type.

Pool	Factor	F	p
Adults	Window	$F(17) = 4.3$	**0.000***
	Word Type	$F(3) = 2.8$	**0.036***
	Window × Word Type	$F(51) = 0.9$	0.685
Children	Window	$F(17) = 2.1$	**0.006***
	Word Type	$F(3) = 8.0$	**0.000***
	Window × Word Type	$F(51) = 1.9$	**0.0002***
Both	Window	$F(17) = 5.0$	**0.000***
	Subject Group	$F(1) = 2.8$	0.092
	Word Type	$F(3) = 6.2$	**0.0004***
	Window × Subject Group	$F(17) = 1.4$	0.141
	Window × Word Type	$F(51) = 1.8$	**0.000***
	Subject Group × Word Type	$F(3) = 5.1$	**0.002***
	Window × Subject Group × Word Type	$F(51) = 0.9$	0.705

Table 4.2: Results of ANOVA tests with window, word type and subject group as factors. p-values < 0.05 are highlighted in bold with an asterisk.

For adults, there is a significant main effect of window and word type on the mean amplitudes. However, it is surprising that there is no significant interaction between window and word type, which means the effect of word type does not depend on the window and vice versa. It might be the case that only a very few combinations of word type and window differ in their sample mean. For children, there are also significant main effects of window and word type, as well as a significant interaction between window and word type. For both adults and children, there is a significant interaction between window and word type, which could be largely contributed by children.

We performed further ANOVA tests to find out exactly which word type has differing mean amplitude, and in which window. In each window and each subject group (adults, children or both), we performed a one-way between-subjects ANOVA test to compare the effect of word type on the mean amplitude. In each analysis, the null hypothesis is that the averages of the samples (mean amplitudes) from each sample group (word type) are the same. A p-value < 0.05 indicates at least one pair of sample groups with significantly different averages, at the 95% confidence level. Each ANOVA test is also followed by a Tukey-Kramer test to compare the averages of the sample groups in a fair way that compensates for multiple comparisons. The Tukey-Kramer test is suitable as results from the Levene's tests for each window and subject group combination indicate that the distributions of mean amplitude for each word type have the same variances at the 95% confidence level.

Table 4.3 shows the results from the ANOVA and Tukey-Kramer tests. The fourth and fifth columns show the F-test results and the p-values, respectively, from the ANOVA tests. The last column shows the pairs of sample groups with significantly different averages, as a result of the

Tukey-Kramer tests; a dash indicates no significant pair. The 95% confidence interval of the difference in mean between the first and second items in each significant pair is shown in parentheses; '>' indicates the first item having a larger mean than the second. Windows with insignificant results are omitted from the table for brevity.

Effect	Window	Pool	*F*	*p*	Significant Pairs
P200	0.10–0.20	Adults	$F(3, 143) = 2.0$	0.111	-
		Children	$F(3, 158) = 3.3$	**0.022***	Hard>Pseudo (1.3, 19.5), Hard>Non (1.3, 22.1)
		Both	$F(3, 305) = 5.0$	**0.002***	Hard>Pseudo (1.3, 13.3), Hard>Non (2.6, 15.4)
	0.10–0.30	Adults	$F(3, 143) = 2.9$	**0.039***	Hard>Non (1.0, 13.1)
		Children	$F(3, 158) = 2.2$	0.088	-
		Both	$F(3, 305) = 4.0$	**0.008***	Hard>Pseudo (1.0, 12.2), Hard>Non (1.5, 13.2)
Other	0.25–0.35	Adults	$F(3, 143) = 0.4$	0.748	-
		Children	$F(3, 158) = 2.4$	**0.047***	Easy<Non (-20.6, -0.2)
		Both	$F(3, 305) = 2.1$	0.106	-
N400	0.25–0.45	Adults	$F(3, 143) = 0.7$	0.527	-
		Children	$F(3, 158) = 2.7$	**0.046***	Easy<Pseudo (-15.1, -1.4)
		Both	$F(3, 305) = 1.1$	0.333	-

Table 4.3: Results of the ANOVA and Tukey-Kramer tests. *p*-values < 0.05 are bolded with an asterisk.

For adults, there is only one significant pair in the window 0.10–0.30, which could partly explain the insignificant interaction between window and word type as found earlier. For the window 0.1–0.2, the Hard>Pseudo and Hard>Non pairs in children survived after pooling adults and children together, indicating that the addition of adults did not dilute the significance of these two pairs. In fact, these two pairs might already be "almost significant" as suggested by the *p*-value of 0.111 in adults. The same type of effect can be observed in the window 0.1–0.3, where the Hard>Non pair survived after pooling adults and children together, and an additional Hard>Pseudo pair even became significant.

At a high level, the results indicate that it might be difficult to classify trials using the mean ERP amplitudes. The small number of significant pairs suggests that in each window, the distributions of the mean amplitude from the word types highly overlap, posing a great challenge to separate the trials using a classifier.

4.4 Classification results

In this section, we present and discuss results of classification tasks using the mean amplitudes of ERPs as features.

4.4.1 4-way classification

To find out whether the mean amplitudes in each window would help in identifying the four mental states associated with the word types, we performed a 4-way classification using the ℓ_1-

regularized multinomial logistic regression classifier on trials with duration between 0.6 and 5 seconds. All suitable windows except for the pre-stimulus window are included, regardless of whether they produced any significant results in earlier ANOVA tests, because the regularized classifier should choose the features useful for determining the word type. Our hypothesis is that the classification performance would not be extremely better than chance, since the ANOVA analyses shown in Table 4.3 indicate that the pairs of word types are not easily separable.

Rank Accuracies

	Reader-specific		Reader-independent	
	ERP features	Baseline	ERP features	Baseline
Adults	54% (0.248)	51%	**57% (0.0034) *#**	49%
Children	48% (0.761)	49%	47% (0.7608)	51%
Both/adults	54%	-	58%	-
Both/children	48%	-	58%	-
Both	52% (0.151)	50%	**58% (0.0033) *#**	49%

Classification Accuracies

	Reader-specific		Reader-independent	
	ERP features	Baseline	ERP features	Baseline
Adults	30% (0.206)	23%	**31% (0.0083) *#**	25%
Children	24% (0.572)	23%	24% (0.576)	25%
Both/adults	30%	-	34%	-
Both/children	24%	-	35%	-
Both	27% (0.280)	24%	**35% (0.0039) *#**	24%

Table 4.4: Rank and 4-way classification accuracies of classifying silent word trials. *p*-values are shown in parentheses. An asterisk indicates above-chance significance at the 5% level with a *t*-test. An additional hex indicates significance after accounting for multiple comparisons.

Table 4.4 above shows the rank and classification accuracies and the associated *p*-values in parentheses, along with the results obtained using the baseline features. Each rank or classification accuracy is the average across subjects. Each *p*-value is obtained by performing a one-sample *t*-test on the accuracies contributed by the subjects. "Both/adults" shows the average accuracy across only adult subjects while using training trials from children as well; "both/children" is likewise. The reader-specific average accuracy for "both/adults" is the same as that for just adults since testing and training trials are taken from the same subject, but the reader-independent average accuracy is different because while the testing trials are taken from the same subject, the training trials are from both adults and children.

There are still significant results even though the classification performance is not extremely better than chance. Adults and "both" show reader-independent results significantly better than the chance performance of 50% (rank accuracy) or 25% (classification accuracy), even after

accounting for multiple comparison using Bonferroni correction ($\alpha = 0.05/6 = 0.0083$) and false discovery. For adults, the improvement in reader-independent rank accuracy from that of reader-specific, resulting in statistical significance, could be explained by the fact that more beneficial training trials from other adults are available for the subject in each iteration of cross-validation. The benefit offered by training trials from other adults indicates that there are some common characteristics between adults' signals. This comes as no surprise as existing ERP studies have long demonstrated that ERPs can be observed across subjects. However, it is surprising that this kind of improvement in reader-independent accuracy is not evident in children (48% vs. 47%). This difference suggests that more training trials from other children yield little or no improvement in the reader-independent rank accuracy of a child subject, but do yield slight improvement to both/adults (54% vs. 58%). In contrast, when more training trials are taken from adults, there is an improvement in both/children's reader-independent rank accuracy (48% vs. 58%), albeit not tested for significance. Overall, adults can benefit from children's trials but not as much as children from adults', and children benefit the least from other children's trials. This interesting behavior can also be observed in the classification accuracies, which is not surprising as classification and rank accuracy are correlated. A further investigation of this behavior is needed in the future.

It is evident the ERP features perform better than the baseline features. None of the baseline results are significant as the *p*-values range from 0.417 to 0.663 for rank accuracies and from 0.479 to 0.742 for classification accuracies. Note that the baseline accuracies reported here differ from those reported in Mostow et al. [2011] due to the removal of the proprietary attention and meditation signals.

4.4.2 Pairwise classification

The results shown in the previous subsection indicate that it is difficult to distinguish amongst the four word types, although there are some significant but modest results. To further understand the difficulty of this 4-way classification task, we performed pairwise classification using the same set of ERP features to investigate which pairs are more separable than the others.

Table 4.5 below shows the classification accuracies for each pair of word types per each subject group. There are 12 *t*-tests (6 pairs × 2) performed per each subject group and trial group combination. After accounting for Bonferroni correction ($\alpha = 0.05/12 = 0.0042$) and false discovery, there are two significant results as indicated by a hex sign. Note that the 1–5 seconds trial group uses additional features from windows that end beyond 0.6 second, while the other trial group excludes such features. This could explain why certain pairs, for example, hard/pseudo in adults, have higher classification accuracies in the 1–5 seconds trial group.

Adults

	0.6–5 seconds		1–5 seconds	
Pair	Reader-specific	Reader-independent	Reader-specific	Reader-independent
Easy/Hard		65% (0.014) *		
Easy/Pseudo				
Easy/Non				
Hard/Pseudo				60% (0.037) *
Hard/Non		60% (0.0021) *#		61% (0.043) *
Pseudo/Non				61% (0.049) *

Children

	0.6–5 seconds		1–5 seconds	
Pair	Reader-specific	Reader-independent	Reader-specific	Reader-independent
Easy/Hard				
Easy/Pseudo				
Easy/Non				65% (0.047) *
Hard/Pseudo				
Hard/Non				
Pseudo/Non			62% (0.023) *	

Adults and Children

	0.6–5 seconds		1–5 seconds	
Pair	Reader-specific	Reader-independent	Reader-specific	Reader-independent
Easy/Hard				
Easy/Pseudo	59% (0.012) *			
Easy/Non	60% (0.0091) *	60% (0.017) *		58% (0.048) *
Hard/Pseudo		56% (0.021) *		
Hard/Non		63% (0.0036) *#		63% (0.011) *
Pseudo/Non				

Table 4.5: Classification accuracies for each pair of word types per subject group. *p*-values are shown in parentheses. An asterisk indicates significance using the *t*-test at the 5% level. An additional hex indicates significance after accounting for multiple comparisons.

4.4.3 Per-window classification

The previous subsection shows the accuracies of classifying each pair of word types using all the suitable windows as basis for features. To investigate which window works best for each pair of word types, we used all suitable windows as features individually. Table 4.6, Table 4.7 and Table 4.8 below show the results along with the best performing window; for brevity, only ones significant at the 5% level are shown. None of the results are significant after accounting for multiple comparisons, although there are a few with very low *p*-values (0.0006 and 0.0007) in

the table for both adults and children. To conserve space, the names of the word types are abbreviated as: easy (E), hard (H), pseudo (P) and non (N).

Adults

	0.6–5 seconds				1–5 seconds			
	Reader-specific		Reader-independent		Reader-specific		Reader-independent	
	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)
E/H							.50-1.0	70 (.008) *
E/P			.25-.35	61 (.037) *			.50-1.0	63 (.036) *
E/N	.10-.20	60 (.014) *			.10-.25	59 (.0312) *		
H/P			.00-.30	63 (.008) *			.00-.30	62 (.007) *
H/N			.00-.30 .25-.50	62 (.066) 58 (.0006) *				
P/N	.10-.20	62 (.033) *	.10-.20	59 (.018) *			.10-.20	61 (.004) *

Table 4.6: Classification accuracies for each pair of word types in adults. p-values are shown in parentheses. For brevity, only results significant at the 5% level, as indicated by an asterisk, are shown. To conserve space, the word types are abbreviated as: easy (E), hard (H), pseudo (P) and non (N).

For adults, the accuracies range from 58% (hard/non) to 70% (easy/hard). The results generally agree with the grand averages shown in Figure 4.1, in that there are visual differences within each pair in the grand averages in the best performing window. For example, consider the easy/non pair in the 0.6–5 seconds trial group. In the grand average for easy words, there is a positive deflection in the best performing window 0.10–0.20 while there is a negative deflection in the grand average for non-words. Also consider the easy/pseudo pair. In the grand average for easy words, there is a clear positive deflection in the best performing window 0.25–0.35 while there is no such deflection in the same window for pseudo-words.

The easy/hard pair has no significant result here even though it has an accuracy of 65% ($p < 0.05$) when all suitable windows are used. This suggests that multiple windows provide more information to separate easy from hard trials. Also some pairs like easy/pseudo are separable here only when single windows are used. This suggests that the high classifier model complexity as a consequence of using all suitable windows causes the learned classifier to overfit and hence not generalize well to unseen test trials.

The 0.6–5 seconds and the 1–5 seconds trial groups share most of the best performing windows. The most noticeable differences are for the easy/hard and easy/pseudo pairs which have the 0.50–1.0 window as best performing. The easy/non pair does not have exactly the same best performing windows (0.10–0.20 and 0.10–0.25) in the two trial groups but nevertheless both windows are still similar.

Children

	0.6–5 seconds				1–5 seconds			
	Reader-specific		Reader-independent		Reader-specific		Reader-independent	
	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)
E/H	.25-.35	59 (.025) *	.25-.45	63 (.044) *			.20-.40	64 (.009) *
E/P	.25-.50	62 (.017) *	.20-.40	69 (.004) *			.25-.50	68 (.024) *
E/N	.20-.50	59 (.022) *					.25-.45	64 (.045) *
H/P	.10-.25	56 (.021) *						
H/N			.10-.25	58 (.015) *				
P/N								

Table 4.7: Classification accuracies for each pair of word types in children.

For children, the accuracies range from 56% (hard/pseudo) to 69% (easy/pseudo). The results are again generally in agreement with the grand averages shown in Figure 4.2. The pseudo/non pair is not distinguishable which is unsurprising as their grand averages are seemingly random.

Adults and Children

	0.6–5 seconds				1–5 seconds			
	Reader-specific		Reader-independent		Reader-specific		Reader-independent	
	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)	Win.	Acc. (%)
E/H			.25-.35	65 (.0007) *			.40-1.0	65 (.003) *
E/P	.25-.35	57 (.006) *	.25-.35	61 (.001) *			.25-.35	61 (.002) *
E/N	.10-.20	57 (.041) *	.00-.20	59 (.017) *				
H/P	.30-.55	54 (.007) *					.30-.55	61 (.010) *
H/N			.30-.55	61 (.0006) *			.30-.55	60 (.007) *
P/N			.00-.20	54 (.013) *			.00-.20	54 (.043) *

Table 4.8: Classification accuracies for each pair of word types in both adults and children.

For both adults and children together, all pairs of word types are distinguishable with varying degrees of accuracy. Again, it is noticeable that reader-independent accuracies are usually higher than the corresponding reader-specific accuracies. The easy/hard pair shows the highest accuracy (65%). This pair is also distinguishable in children but not adults. For adults, this pair has a best performing window of 0.25–0.45 which is the same as that in children, but with an accuracy of only 59% (p=0.108). Indeed, in that window, the grand averages for easy and hard words in adults both have a negative deflection instead of deflections of opposing polarity, which could contribute to the difficulty of distinguishing that pair.

4.4.4 Summary

The 4-way classification task shows some modest results. When we performed the pairwise classification, we obtained a deeper understanding of which pairs are more distinguishable from others.

Although the results in the per-window classification are not significant when accounting for multiple comparisons, they generally agree with the grand averages. Furthermore, the easy/hard and hard/non pairs in both adults and children together have very low *p*-values (0.0007 and 0.0006 respectively) which cannot be ignored. They indicate the potential to distinguish between the mental states that arise when the subject is reading an easy vs. a hard word.

Chapter 5

Sentence Analyses

In this chapter, we first present preliminary analyses of the correlation feature (section 3.3.6) and then show how it performs in the sentence trials classification task along with other features. In the classification task, since the features for each sentence trial are computed without any consideration of the temporal structure of the EEG signal and power spectrum, we next investigate how relative segmentation and truncation affect classification accuracy.

5.1 Feature analyses

The mean amplitudes of the EEG and filtered signals as well as the power ratio between a pair of frequency bands are commonly used as features. Using the correlation between a pair of frequency bands recorded by the same electrode as a feature is less common. Hence we analyze how frequency bands correlate in adults and children in oral and silent reading, so as to understand how this feature could be useful in distinguishing between the EEG signals for easy and hard sentences.

5.1.1 Correlation

Figure 5.1 below shows the across-subject average Pearson correlation coefficients for each pair of frequency bands for sentence trials in either reading condition. The correlation coefficient for each pair of frequency bands for a sentence trial is obtained by applying the correlation formula to the entire duration. The coefficients are presented in the form of heatmaps, with red representing 1, black 0, and blue -1. The shade of the color in each cell in a heatmap represents the degree of correlation: the higher the intensity, the larger the correlation. To conserve space, the names of the frequency bands are abbreviated as follows: delta (D), theta (T), alpha (A), low beta (LB), mid beta (MB), high beta (HB) and gamma (G). The beta band is excluded from analyses since it correlates highly with low, mid and high beta.

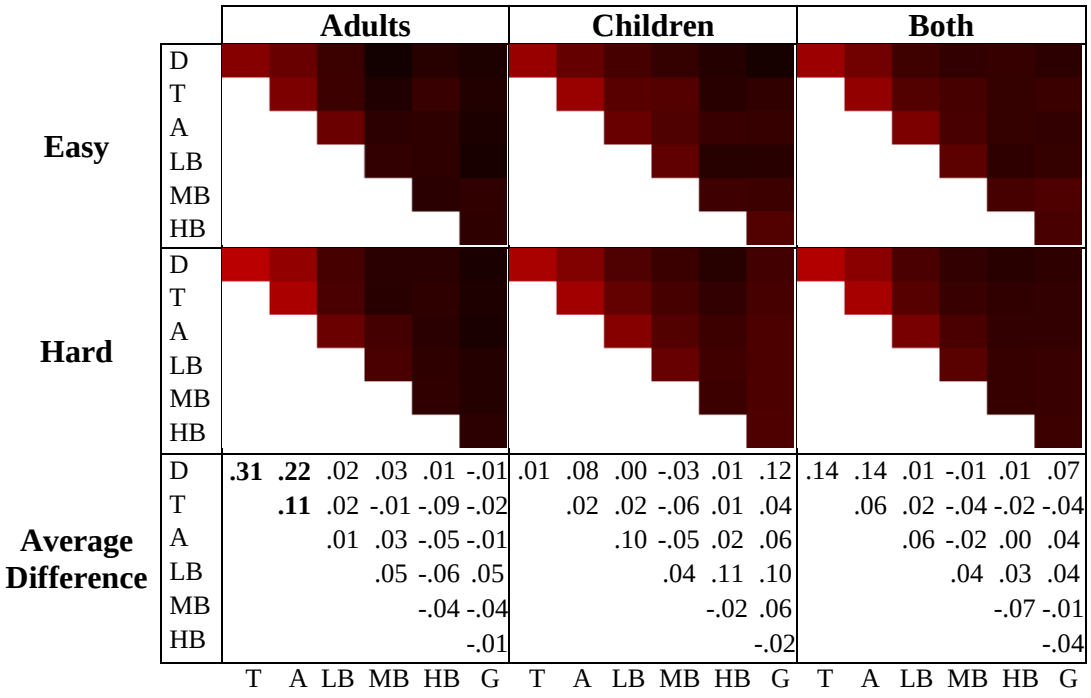
Within each subject and for each reading condition, undersampling is used to select easy sentence trials since there are more easy than hard ones. Undersampling in this case prevents any bias that could be introduced by the imbalanced data set. For example, the correlation coefficients for easy sentences might be randomly distributed between -1 and 1, and by averaging over a larger number of trials, the coefficients could appear smaller than for hard sentences due to cancellation. Only with undersampling can we be sure that a comparison in coefficients between easy and hard sentences is fair.

The average heatmap is calculated for each subject first before producing the grand average across subjects as shown in the figure below. Calculating the average heatmap for each subject first allows a sanity check to ensure the individual subjects' averages are consistent. If all the sentence trials from all subjects were pooled together to produce the grand average first, the

individual subjects' differences in their trials would be hidden, leaving no room for a consistency check.

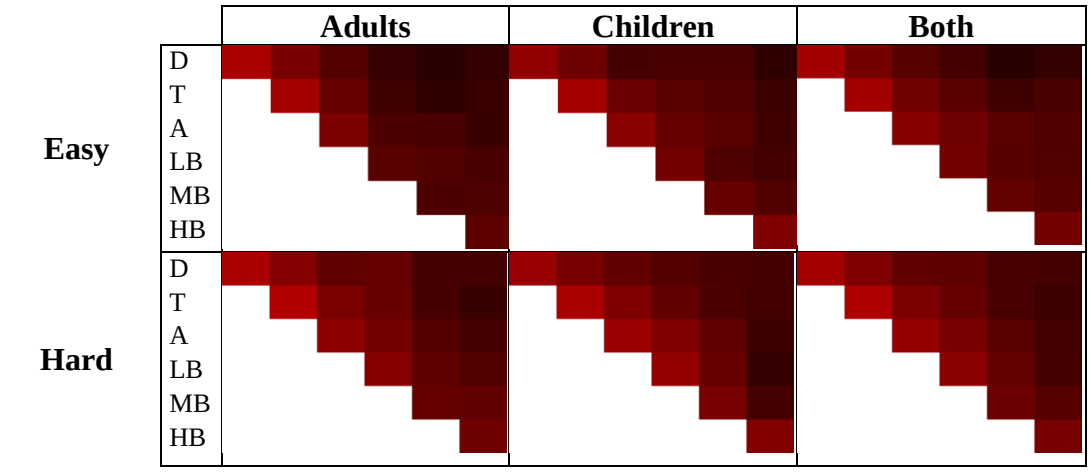
Within each subject, the difference in correlation coefficients for a pair of frequency bands is calculated as the subject's hard average minus easy average. The figure below shows the average differences across subjects. Actual numbers are presented to provide a concrete sense of the average differences.

Sentence/Silent reading



Average Difference	Adults T	Adults A	Adults LB	Adults MB	Adults HB	Adults G	Children T	Children A	Children LB	Children MB	Children HB	Children G	Both T	Both A	Both LB	Both MB	Both HB	Both G
D	**.31**	**.22**	.02	.03	.01	-.01	.01	.08	.00	-.03	.01	.12	.14	.14	.01	-.01	.01	.07
T		**.11**	.02	-.01	-.09	-.02		.02	.02	-.06	.01	.04		.06	.02	-.04	-.02	-.04
A			.01	.03	-.05	-.01			.10	-.05	.02	.06			.06	-.02	.00	.04
LB				.05	-.06	.05				.04	.11	.10				.04	.03	.04
MB					-.04	-.04					-.02	.06					-.07	-.01
HB						-.01						-.02						-.04

Sentence/Oral reading



Average Difference

D	.02	.04	**.16**	**.20**	-.01	.07	-.09	.04	.09	.08	.06	-.06	-.04	.04	.13	.14	.02	.01
T		**.13**	**.18**	**.18**	-.01	.03		.08	.06	-.01	.03	-.04		.11	.12	.09	.01	.00
A			**.12**	**.13**	-.05	.01			.11	.03	**.14**	-.05			.11	.09	.04	-.02
LB				.06	-.06	-.01				.08	.04	-.13				.06	-.01	-.07
MB					-.03	.06					**.14**	-.04					.05	.01
HB						.04						.09						.06
	T	A	LB	MB	HB	G	T	A	LB	MB	HB	G	T	A	LB	MB	HB	G

Figure 5.1: Average (across subjects) correlation coefficients for each pair of frequency bands for sentence trials in both reading conditions are shown in the form of heatmaps, with red representing 1, black 0, and blue -1. The shade of the color in each cell indicates the degree of the correlation. The frequency bands are abbreviated as follows: delta (D), theta (T), alpha (A), low beta (LB), mid beta (MB), high beta (HB) and gamma (G). The difference in values of the correlation coefficients for a pair of frequency band is calculated as that of hard minus easy. The average difference across subjects is presented here. Values referenced in the discussion are bolded.

In the silent reading condition, adults have larger correlation coefficients for hard sentences than easy ones. This effect is especially pronounced for the delta/theta, delta/alpha, and theta/alpha pairs in adults, as shown by the average differences, while the same effect seems more evenly distributed amongst the pairs in children.

In the oral reading condition, the correlation coefficients are generally larger than that in silent reading, as shown by the higher intensities of red in each cell in the heatmaps. In adults, the low beta and mid beta bands have relatively larger correlation coefficients with the delta, theta, and alpha bands; theta-alpha, too, has large correlation coefficients. In children, there are fewer pairs with relatively larger correlation coefficients: high beta with alpha and mid beta.

Table 5.1 below shows the pairs of frequency bands that have different correlation coefficients in easy and hard sentence trials for each reading condition. Within each subject, an unpaired two-sample *t*-test compared the correlation coefficients of each pair of frequency bands in easy and hard trials. The last row of the table shows the results of an unpaired two-sample *t*-test that compared the per-subject average correlation coefficients in easy and hard trials. Each adult and child contributed $\binom{7}{2} = 21$ coefficients since there are 7 distinct frequency bands in consideration. Before any *t*-test is performed, the correlation coefficients are first subjected to Fisher transformation so that the converted values are approximately normal. All results significant at the 10% level are shown here, with 10% chosen to include results that are "close enough" for discussion purposes. When accounting for multiple comparisons within each subject and between subjects, none of the results are significant.

For the silent reading condition, even though the grand average heatmap above indicates that these delta-theta, delta-alpha, and theta-alpha pairs are likely to be significant in most adults, in reality, they are not. The delta-theta pair is significant ($p < 0.05$ uncorrected) in only 2 out of 7 adults and the delta-alpha and theta-alpha pairs are significant ($p < 0.05$) in only 1.

Sentence/Silent reading

Subject	Significant pairs	*p*	Subject	Significant pairs	*p*
Adult-02	Delta-Theta	**0.015***	Child-11	Theta-Alpha	**0.016***
Adult-03	Delta-Theta Delta-Alpha Theta-Alpha	**0.018*** **0.029*** **0.022***	Child-12	Alpha-Gamma	0.084
Adult-04	LowBeta-MidBeta MidBeta-Gamma	**0.019*** **0.044***	Child-14	Delta-LowBeta	0.076
Adult-08	Delta-Alpha Theta-Alpha	0.068 0.072	Child-16	Delta-MidBeta Theta-HighBeta Alpha-HighBeta	0.055 0.054 **0.041***
Adult-10	MidBeta-Gamma	0.054	-	-	-
Between 7 adults	Delta-Theta Delta-Alpha Theta-Alpha	**0.029*** 0.085 0.062	Between 10 children	Delta-Gamma Alpha-LowBeta	0.060 0.088

Sentence/Oral reading

Subject	Significant pairs	*p*	Subject	Significant pairs	*p*
Adult-03	Delta-HighBeta	0.081	Child-14	Delta-LowBeta Alpha-HighBeta	**0.049*** **0.015***
Adult-04	Theta-Alpha Theta-LowBeta	**0.005*** 0.093	Child-15	Alpha-HighBeta	0.088
Adult-06	Delta-MidBeta Theta-LowBeta Alpha-MidBeta	0.050 0.073 **0.031***	Child-19	Delta-MidBeta Delta-Gamma	0.067 0.062
Adult-07	Delta-MidBeta Delta-Beta	**0.035*** **0.021***	Child-20	Delta-Theta	0.074
Adult-10	Theta-Alpha Theta-LowBeta Alpha-LowBeta Alpha-MidBeta	0.077 0.088 0.074 0.088	-	-	-
Between 7 adults	Delta-LowBeta Delta-MidBeta Theta-Alpha Theta-MidBeta Alpha-MidBeta	0.071 0.082 **0.041*** 0.085 **0.049***	Between 10 children	-	-

Table 5.1: Pairs of frequency bands that have different correlation coefficients in easy and hard sentence trials at the 10% significance level. *p* values < 0.05 are bolded with an asterisk. Note the subject index numbers are not assigned sequentially.

For the oral reading condition, the grand average heatmap for adults generally agrees with the *t*-test results as the heatmap shows that low beta and mid beta bands have relatively larger correlation coefficients with delta, theta and alpha, which is indeed the case here. However, most of the pairs in the table are not significant even at the uncorrected 5% level.

5.1.2 Classification results

We performed an easy vs. hard sentence trial classification to find out how each feature performs, more specifically, the single most predictive feature for each subject group (adults, children or both) and reading condition. We tested each feature one at a time, that is, we represented each trial by a single feature at a time. As a comparison, we also included one at a time, the baseline features (section 3.3.1), specifically, only the mean power of each 1 Hz band, since the other baseline features are already found in the mean EEG feature group. Before performing the classification task, we expected the correlation features to perform as accurately as other features.

Table 5.2, Table 5.3 and Table 5.4 below show the results that are significant at the 1% level, as indicated by an asterisk; others are omitted for brevity. Each cell shows the lowest and highest accuracies achieved by any feature in the corresponding feature group, and the best performing feature along with its associated *p*-value in parentheses. The last row of each table shows the performance of baseline features as a comparison. When accounting for multiple comparison within each feature group, subject group and reading condition, only the mean power of alpha in adults and mean amplitude of filtered in both adults and children together are significant (as indicated by a hex). When accounting for multiple comparisons across feature groups, none of the results are significant.

Adults

	Oral		Silent	
Feature Group	Reader-specific	Reader-independent	Reader-specific	Reader-independent
Mean EEG / Filtered				
Mean Power		Alpha 50-66% (.001) *#		
Ratio 1			Theta-Beta 37-59% (.008) *	HighBeta-Theta 41-60% (.007) *
Ratio 2			Theta-LowBeta 41-63% (.001) *	MidBeta-Theta 38-59% (.007) *
Corre.		Delta-MidBeta 44-62% (.009) *	Delta-Theta 38-60% (.007) *	Delta-Alpha 46-62% (.003) *
Baseline				

Table 5.2: Single feature classification accuracies for adults. For brevity, only results significant at the 1% level are shown. Each cell shows the best performing feature with its associated *p*-value in parentheses, and the lowest and highest accuracies obtained by any feature in the respective feature group. An asterisk indicates significance at the 5% level while a hex indicates significance after accounting for multiple comparisons in each feature group.

In the oral, reader-independent condition for adults, the mean power of alpha is the single most predictive feature. The fact that it is significant within its own feature group for the reader-independent condition but not for the reader-specific condition again suggests that more data helps in improving the classification accuracy. Our expectation that the correlation features would achieve a similar level of accuracy as other features is indeed confirmed. For example, the correlation of delta-theta in the silent, reader-specific condition is the best performing feature in the correlation feature group. Correlation of delta-theta also performed similarly well in the silent, reader-independent condition, with an accuracy of 60% and *p*-value of 0.004 (not shown in table). These observations are in agreement with prior analyses of the correlation feature. However, the correlation of delta-alpha performed surprisingly well even though its difference between easy and hard sentences is not significant at $p=0.085$ between adults. Another interesting point is that the best performing features in the ratio 1 and 2 groups involve the theta and beta family bands.

Children

	Oral		Silent	
Feature Group	Reader-specific	Reader-independent	Reader-specific	Reader-independent
Mean EEG / Filtered				
Mean Power				
Ratio 1	Gamma-HighBeta 30-64% (.009) *			
Ratio 2	Theta-LowBeta 31-69% (.009) *			
Corre.				
Baseline				

Table 5.3: Single feature classification accuracies for children.

There are far fewer significant results for children than for adults, suggesting that there is less consistency across children or even across individual trials; in other words, it is difficult to have a single predictive feature across subjects or individual trials. With reference to earlier analyses of the correlation feature, the correlation of alpha-lowbeta has an accuracy of 58% and a *p*-value of 0.018; the low accuracy is consistent with the fact that it has a *p*-value of only 0.088 when its difference between easy and hard sentences is tested for significance.

Adults and Children

	Oral		Silent	
Feature	Reader-specific	Reader-independent	Reader-specific	Reader-independent

Group				
Mean EEG / Filtered				Filtered 54-54% (.008) *#
Mean Power		Gamma 40-59% (.005) *		
Ratio 1				
Ratio 2	Theta-Beta 38-62% (.009) *		Theta-Beta 44-59% (.003) *	Gamma-HighBeta 42-59% (.009) *
Corre.		Delta-MidBeta 46-60% (.002) *		
Baseline				

Table 5.4: Single feature classification accuracies for both adults and children.

For both adults and children together, it is interesting that the ratio-2 of theta-beta is the best feature in both the oral and silent reader-specific conditions. Moreover, we see that the ratio-2 of theta-lowbeta is also the best feature in the oral, reader-specific condition for children and silent, reader-specific condition for adults. These observations suggest that this ratio is stable even in the presence of muscle artifacts generated during oral reading.

5.1.3 Summary

There are interesting and unexpected results in the feature analyses which warrant further investigation in the future. In the analyses of the correlation feature, we observe that in silent reading, the delta-theta, delta-alpha and theta-alpha correlation coefficients are higher for hard sentence trials than for easy ones. This difference in the correlation coefficients allowed us to exploit them as features in the classification task to achieve an accuracy of 60–62% in adults. The correlation coefficients for the oral reading condition are also higher than for the silent condition, an observation also worth investigating in the future, although it does not seem to be directly related to the classification task of distinguishing between easy and hard sentences. Lastly, the classification task shows the ratio-2 of theta-beta/lowbeta to be stable in the presence of muscle artifacts in the oral reading condition.

5.2 Temporal analyses

We next investigate how the temporal structure of sentence trials affects classification accuracy.

5.2.1 Relative segmentation

We performed a classification task in which the EEG signal and power spectrum for each trial were divided into 3 equal segments. Again, each feature was tested one at a time, for each individual segment. The aim is to find out which feature works best in each segment, and also which segment would be useful in the classification task. Also, with a noisy data set, another important aim is to look for a single feature that works consistently well, for example, across all

segments, or across both the segmented and unsegmented cases. Such consistent features warrant future investigations of their predictive information. Table 5.5, Table 5.6 and Table 5.7 below show the classification accuracies for each segment. For brevity, only the best performing feature in each segment is shown, along with its accuracy and *p*-value; interesting results are also included for discussion purposes. When accounting for multiple comparisons in the entire table, there is no significant result.

Adults

Silent Reading/Reader-specific

Seg-1			Seg-2			Seg-3		
Feature	Acc.	*p*	Feature	Acc.	*p*	Feature	Acc.	*p*
Ratio1-T-B	60%*	0.019	Ratio1-G-LB	60%*	0.006	Ratio1-G-LB	58%*	0.007
Silent Reading/Reader-independent								
Ratio2-T-LB	62%*	0.010	Ratio1-T-HB	61%*	0.049	Ratio2-MB-T	61%*	0.005
Oral Reading/Reader-specific								
Ratio1-T-D	58%*	0.033	Ratio1-D-MB	64%*	0.001	Ratio1-T-A	61%*	0.002
Oral Reading/Reader-independent								
Ratio1-A-D	66%*	0.005	Ratio2-G-A Corre-D-MB	63%* 61%*	0.009 0.006	Mean-MB Mean-LB	63%* 62%*	0.037 0.005

Table 5.5: Classification accuracies for each segment, in adults. For brevity, only the top performing feature along with its accuracy and *p*-value are shown. An asterisk indicates significance at the 5% level.

For adults, there are some results that are in agreement with prior classification results in section 5.1.2. The ratio-1 of theta-beta in the silent, reader-specific condition is the best predictive feature in the first segment, as well as over the entire trial (Table 5.2). The accuracy here is 60% which is comparable to that of 59% found earlier. However, this ratio does not perform as well for the other two segments, with accuracies less than 52%. This suggests that this feature works best during the first segment, that is, the start of a sentence. The ratio-1 of gamma-lowbeta works consistently well across the second and third segments. In the silent, reader-independent condition, the ratio-2 of midbeta-theta is the best predictive feature for the third segment as well as over the entire trial. The accuracy here is slightly higher here as well (61% vs. 59%), although not statistically significant. In the oral, reader-independent condition, the correlation of delta-midbeta is the best predictive feature for the second segment as well as over the entire trial, although it has a slightly lower accuracy here (61% vs. 62%). In the third segment, interestingly, the mean power of low beta (14–16 Hz) is the second best predictive feature and is close frequency-wise to the alpha band (9–13 Hz). This suggests there is common activity in these two bands that is predictive. A future investigation could find out how well the two bands combined would perform. The mean power of mid beta (17–20 Hz), with an accuracy of 63%, also seems

to share predictive information with alpha. In adults, although some features achieved better accuracy in certain segments than over the entire trial, the differences are too small to be considered conclusive. Hence, it is not certain that segmentation helps for adults.

Children

Silent Reading/Reader-specific

Seg-1			Seg-2			Seg-3		
Feature	Acc.	*p*	Feature	Acc.	*p*	Feature	Acc.	*p*
Ratio1-LB-D	63%*	0.005	Ratio2-MB-D	62%*	0.013	Ratio2-MB-A	58%*	0.024

Silent Reading/Reader-independent

Mean-A	61%*	0.049	Ratio2-A-T	61%*	0.008	Ratio1-LB-D	59%*	0.013

Oral Reading/Reader-specific

Mean-D	66%*	0.015	**Ratio1-B-MB** Ratio2-T-LB	**66%*** 67%*	**3.7e-5** 0.013	Ratio1-MB-B	67%*	0.008

Oral Reading/Reader-independent

Ratio2-T-D	70%*	0.006	Corre-LB-MB	64%*	0.023	Mean-B	68%*	0.013

Table 5.6: Classification accuracies for each segment, in children.

For children, in the oral, reader-specific condition, the ratio-1 of beta-midbeta is especially interesting due to its low *p*-value of 3.7e-5. A closer inspection reveals that almost all the children subjects have accuracies between 65–71%; only one has 54%. The aggregated confusion matrix, computed by summing up across the individual matrices, is shown below. The confusion matrix shows that it is easier to predict a given hard sentence trial correctly, that is, hard trials have a higher recall than for easy trials (90% vs. 44%).

	Actual easy	Actual hard	Total
Predicted easy	82	8	90
Predicted hard	64	27	91
Total	146	35	

The ratio-2 of theta-lowbeta is the best predictive feature in the second segment as well as over the entire trial, but it has lower accuracy here (67% vs. 69%).

Adults and Children

Silent Reading/Reader-specific

Seg-1			Seg-2			Seg-3		
Feature	Acc.	*p*	Feature	Acc.	*p*	Feature	Acc.	*p*

Ratio1-A-B	59%*	0.011	Ratio1-D-T	58%*	0.013	Ratio1-MB-LB	57%*	0.011
Silent Reading/Reader-independent								
Mean-D Ratio1-A-T	62%* 58%*	0.017 0.0003	Ratio2-G-HB	60%*	0.002	Ratio2-G-HB	57%*	0.024
Oral Reading/Reader-specific								
Ratio2-T-LB	58%*	0.022	Ratio1-D-MB	64%*	0.001	Ratio1-T-MB	61%*	0.002
Oral Reading/Reader-independent								
Ratio2-MB-A Corre-D-MB	65%* 60%*	0.0003 0.0003	Ratio2-HB-LB	61%*	0.001	Mean-LB Mean-A	62% * 60% *	0.006 0.005

Table 5.7: Classification accuracies for each segment, in both adults and children.

In the silent, reader-specific condition, there is no common best predictive feature with results in section 5.1.2. However, in the silent, reader-independent condition, the ratio-2 of gamma-highbeta is the best predictive feature in segments 2 and 3 with accuracies of 60% and 57% respectively. It is also the best predictive feature in the silent, reader-independent condition in the unsegmented case with an accuracy of 59%. This suggests that this ratio does not work as well in the first segment. In the oral, reader-specific condition, again there is no common best predictive feature with results for the unsegmented case, although ratio-2 of theta-lowbeta (58%), the best predictive feature in the first segment, is similar to ratio-2 of theta-beta (62%), the best predictive feature in the unsegmented case. Lastly, in the oral, reader-independent condition, correlation of delta-midbeta is the second best predictive feature in the first segment (60%), which is also the best predictive feature in the unsegmented case (60%).

5.2.2 Truncation

We truncated each sentence trial to the first *t* seconds, where *t* ranged from 0.5 to 5 seconds (inclusively) with an interval of 0.5 second. The aim is to investigate which part of the trial contains sufficient useful information for the easy vs. hard classification. As truncation can also be seen as discarding potentially uninformative data after a certain threshold, it would also allow us to investigate whether it leads to improving performance with various threshold values. Table 5.8 below shows the distribution of truncated sentence trials. Each cell is the number of trials available that has a duration of at least *t* seconds. Knowing the distribution aids in understanding the classification results later on. It is evident that the number of hard trials remains fairly constant as duration increases, while the number of easy trials falls more sharply. There are also about 3 or 4 times as many easy trials as hard ones, except for children in the oral reading condition, where this ratio is 5 times instead, due to rejection of many low-quality hard trials.

Adults

Duration ≥ *t* seconds		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Oral	Easy	184	184	184	183	179	164	132	97	73	59

	Hard	57	57	57	56	55	51	51	50	49	49
Silent	Easy	251	221	156	113	73	52	38	22	12	8
	Hard	86	85	83	80	76	68	62	52	50	45

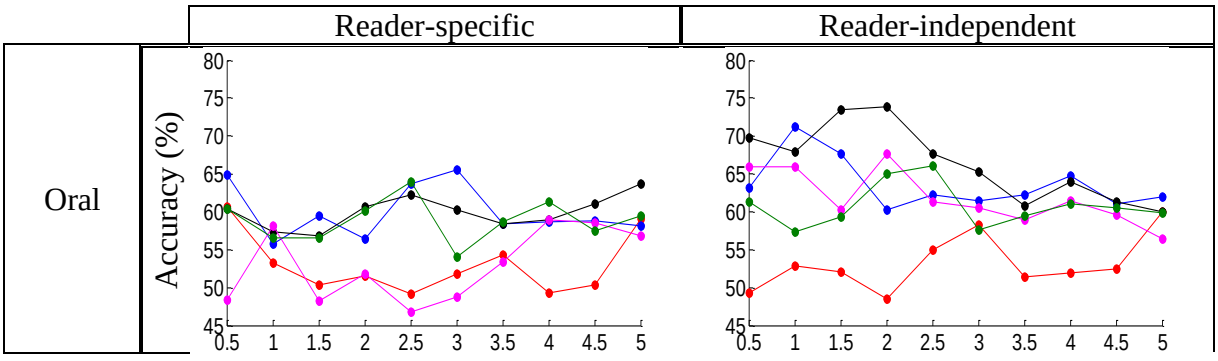
Children

Duration ≥ *t* seconds		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Oral	Easy	180	180	180	180	179	179	170	155	141	122
	Hard	37	37	37	37	37	35	34	33	33	33
Silent	Easy	284	284	264	238	202	172	140	102	81	60
	Hard	85	85	85	84	83	81	81	71	69	63

Table 5.8: Distribution of truncated sentence trials.

We performed classification tasks with truncated trials, again using one feature at a time from each feature group. The aim here is to investigate which part of the trial contains sufficient useful information for the classification task, and how the feature groups perform with respect to the truncation threshold. As we omit the individual best performing features for brevity, we discuss the feature groups in whole. However, we still report if there is any consistent trend in the individual best performing features. Also, the results for both adults and children together are omitted as it would be more interesting to focus on just adults and children separately since past experiments in this chapter have shown that adults and children have vastly different results. Figure 5.2 and Figure 5.3 below show the classification results. Each line in the plots below represents the highest accuracy achieved by any feature in the feature group and is color coded as follows: Red — Mean EEG/Filtered, Blue — Ratio 1, Black — Ratio 2, Magenta — Mean Power, Green — Correlation. There is no feature from any feature group in any reading condition that achieved significant classification accuracy after accounting for multiple comparisons.

Adults



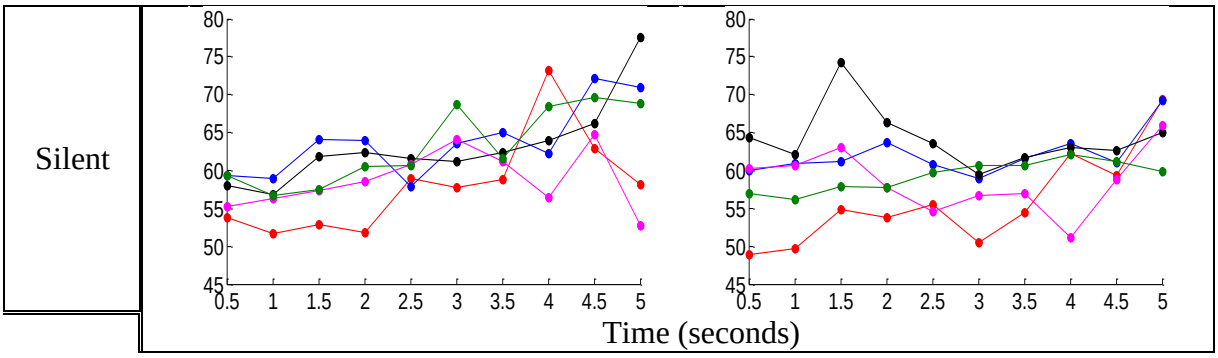


Figure 5.2: Classification results for truncated sentence trials, in adults. Each line in the plots represents the highest accuracy achieved by any feature in the feature group and is color coded as follows: Red — Mean EEG/Filtered, Blue — Ratio 1, Black — Ratio 2, Magenta — Mean Power, Green — Correlation.

For adults, it must be emphasized that the changes in accuracy as the truncated duration increases could be due to the reduced number of trials available. Therefore, results after t=4 must be considered with caution. Indeed, the highest accuracy of 77% in these four plots found at t=5 in the silent, reader-specific condition could be due to only 8 easy trials being available. We observe some interesting trends. For example, the mean EEG feature group generally achieves a lower accuracy than others, except for two cases. First, at t=4 in the silent, reader-specific condition, it achieved a spike in accuracy (74%, p=0.0001), highest amongst the groups for that truncation threshold. Second, in the silent, reader-independent condition, its accuracy rises to match that of the other feature groups as the truncation threshold increases. The first case could be attributed to random variation; the second possibly due to more between-subjects training data. One possible common reason is the reduced number of trials. One hypothesis for the general low performance is that the EEG and filtered signals are still too noisy to be reliable. Further, the observation that the mean EEG group barely surpasses 55% in the oral reading condition (specific & independent) lends support to this hypothesis.

The ratio-2 group generally performs better than the correlation group in the first 2.5 seconds in the reader-independent condition (oral & silent), but otherwise the two groups have comparable performance. Both the ratio-1 and ratio-2 groups have similar performance across all conditions. The mean power group performs better in the silent, reader-independent condition than in the reader-specific condition, again lending support to a long-running observation in this thesis that more training data helps.

It is very interesting that the ratio-2 group can achieve accuracies between 70% and 75% for four times using information from just the first two seconds of a trial, in the oral, reader-independent condition. It is also very interesting that accuracies of 65% and 70% can be achieved just by using the first 0.5 second of a trial.

One consistent trend is that in the silent, reader-specific condition, the mean power of theta and the beta family bands are consistently the best performing features.

Children

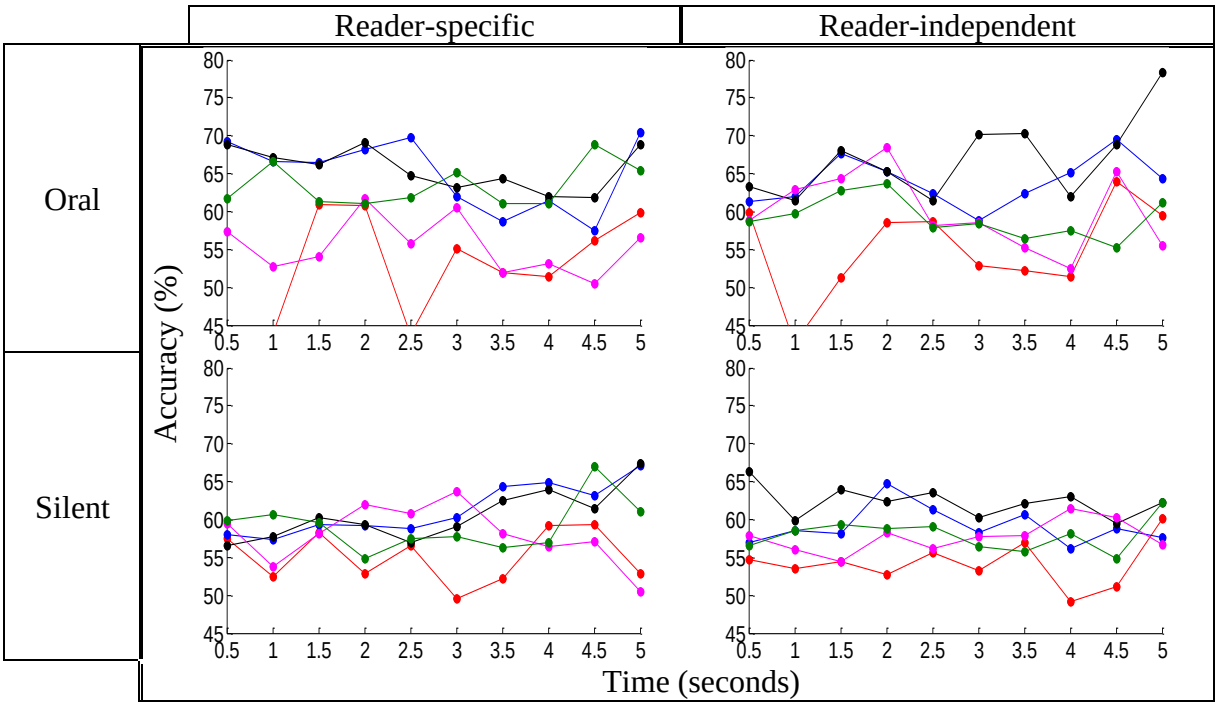


Figure 5.3: Classification results for truncated sentence trials, in children.

In children, we observe that the mean EEG feature group has generally low performance as well as in adults. In fact, there are trends that are common to adults and children. For example, accuracies as high as 70% can be achieved using only the first 0.5 or 1 second worth of information. One main difference from adults is that more between-subject training data does not help here, which again suggests that there is less consistency across children subjects, as shown earlier in section 5.1.2.

5.2.3 Summary

Although there are some distinct patterns across the unsegmented and segmented cases, the reasons behind other results are not immediately clear. We see that some features are the best predictive features in the unsegmented case while also being the best predictive features in certain segments. The other results that could not be explained immediately could be due to random variations and noise in the data set.

It is also not conclusive that relative segmentation improves the classification accuracy for any combination of subject group and reading condition. In both adults and children together, the highest accuracy achieved in the oral, reader-independent condition, is 65% which is higher than

the 60% accuracy achieved in the unsegmented case. Also, in the segmentation case for children, the ratio-1 of beta-midbeta has a surprisingly low *p*-value of 3.7e-5. This highlights the potential of segmentation. However, on the other hand, segmentation does not seem to work as well in adults. The highest accuracies achieved in the segmented and unsegmented cases are comparable in any reading condition. This mixed performance of segmentation could be attributed to the feature extraction methods; specifically, the features might not be compatible with segmentation.

It is also not entirely conclusive that truncation improves classification performance (by discarding potentially uninformative data beyond a threshold), or increases our understanding of the temporal structure of a sentence trial. One key point is that the mean EEG feature group generally has lower performance than other groups. Arguably, the high accuracies of 70–75% achieved by the ratio-1 and ratio-2 groups could be due to the fact that there are many features in each, and some of the features happened to fit the data set better than features from other groups did. Nonetheless, it is our hope that we were able to find a feature that performs consistently well. One consistent trend found in adults is that in the silent, reader-specific condition, the mean power of theta and the beta family bands are consistently the best performing features.

Chapter 6

Conclusion

This chapter discusses the main results, highlights the contributions and provides avenues and insights for future work.

6.1 Discussion of results

In Chapter 4, we learned that the 4-way classification task was only able to produce a modest classification accuracy of 31% for adults, 35% for both adults and children together, and 24% for children. We also learned that some pairs of word types are more easily distinguishable than others, and that some pairs require information from multiple windows to be distinguished while others require only one. The easy/hard pair could be distinguished with an accuracy of 70% for adults and 64% for children, which bodes well for detecting when the user has reading difficulty.

In Chapter 5, there are interesting and unexpected results in the feature analyses which warrant further investigation in the future. In the analyses of the correlation feature, we observe that in silent reading, the delta-theta, delta-alpha and theta-alpha correlation coefficients are higher for hard sentence trials than for easy ones. The correlation coefficients for the oral reading condition are also higher than that for the silent condition. The classification task shows the ratio-2 of theta-beta/lowbeta to be stable, even in the presence of muscle artifacts, across the silent and oral reading conditions. In the truncation analyses, one consistent trend found in adults is that in the silent, reader-specific condition, the mean power of theta and the beta family bands are consistently the best performing features. Referring to Table 5.2 in subsection 5.1.2, we see that the best features for adults in the silent reading condition (specific & independent) are ratios-1 and -2 involving the theta and beta family bands as well. This pattern is certainly interesting and worth further investigation.

The modest results achieved (58–69% across all classification tasks) could be hampered by the small amount of data we collected in the pilot study as well as the simple approaches we used. It is less likely due to the quality of the single, dry electrode Mindset device, as studies discussed in the related work section (and many more) have shown that the device is capable of collecting quality information. We are still in the process of collecting more data and we are optimistic that with more data and advanced methods (as discussed in "Future work" below), we would be able to achieve higher accuracy in the future.

6.2 Contributions

This thesis has two main contributions. First, it demonstrates the potential of exploiting the temporal structure of word trials in detecting the "easy" and "hard" mental states. Second, it shows which EEG and frequency band features are more (or less) predictive of the mental states. Overall, this work is a step closer to automating tutorial decisions in intelligent tutors.

This thesis also highlights the unique challenge faced in the application of EEG in a reading tutor and asynchronous BCI systems. Trials are of widely varying durations due to different reading abilities and the fact that the user controls when the next word or sentence is shown. To deal with the varying durations, we proposed and evaluated the technique of relative segmentation, in particular, 3-segmentation in which each sentence trial is divided into 3 equal parts that can be operationalized as the start, middle and end of the sentence. Too simple to be novel, segmentation nevertheless serves as a convenient and straightforward method to deal with trials of varying durations. Although it is not conclusive that this method works significantly better than just averaging over the entire trial, it is hoped that more data would be able to shed more light on its actual performance.

6.3 Future work

This section discusses several avenues for future work and insights to the problem of detecting mental states. Some of the directions suggested here also seek to answer the main questions raised by the results in this work:

1. Why is the reader-independent classification accuracy sometimes, but not always higher than the reader-specific accuracy?
2. Why do adults and children have differing results in the ERP analyses and the classification tasks?
3. Why do the correlation coefficients for delta-theta, delta-alpha and theta-alpha appear to be larger for hard sentence trials than for easy ones?
4. Why do the correlation coefficients for frequency bands appear to be larger for oral trials than for silent ones?

6.3.1 Artifacts

It would be interesting to study what artifacts arise during oral reading of sentences and how such artifacts would affect the EEG signal. It might even be possible to exploit such artifacts to improve the classification accuracy.

It would also be interesting to study how exactly the behavior of the children and adults when they are using the reading tutor contributes to artifacts and the quality of the signal.

Another direction is to devise a reliable method for detection and removal of ocular artifacts, without the use of EOG.

6.3.2 Adults vs. children

The ERP analyses and classification results for adults and children differ in many places. It would be interesting to further investigate the many differences highlighted by the results in this thesis.

6.3.3 Per-subject study

In many of the classification tasks, although not shown in this work, some subjects achieved much lower accuracies than others. Thus, it might be worthwhile to investigate the per-subject differences to validate the need to have subject-specific models.

6.3.4 Information transfer between subjects

Sometimes having more data to train on improves classification accuracy, as shown in section 4.4.1, where the reader-independent classification accuracy for adults is significantly greater than that of reader-specific. Having more training data reduces the model variance of the classifier, that is, it stabilizes the generalization ability of the classifier. More importantly, it is hypothesized that subjects' data are transferable—in the sense that subjects share similar characteristics in their EEG responses to the same stimuli. Although this hypothesis has already been proven true in ERP studies, it is unclear this hypothesis applies to mental states, specifically whether the user is having difficulty or not. A future work could test this hypothesis by carrying out two classification tasks. In the first task, the testing and training data are from the same subject. In the second one, the testing data is also from the same subject but the training data is from other subjects. It is important to always maintain the same size for the training data in both tasks, so that the only factor that is different is the source of the training data. Also preferably there is an abundance of data for the results to be reliable.

6.3.5 Lexical properties

This work largely ignored the lexical properties of words and sentences. In the pilot study, Mostow et al. showed that there are significant correlations between lexical properties, such as age of acquisition (AOA), and mean power of frequency bands. Future work could investigate the relationship between lexical properties and other EEG and frequency power features. It might also be possible to represent class labels such as easy and hard with lexical properties. For example a threshold value could divide the range of values for a lexical property into two intervals, one that represents easy and the other hard; a classifier could be trained to predict from EEG and/or frequency band features the value of the lexical property which could then be translated into a class label, based on the threshold value. This indirect method of mapping EEG features to lexical properties would work if lexical properties are more predictive of class labels. In fact, Sudre et al. [2012] employed this method to predict from magnetoencephalography (MEG) features which noun word a subject is thinking. Their work differs from this suggestion in that they used MEG and *semantic* properties instead. Also, the ability to predict which word the user is thinking of is not as useful since the Reading Tutor knows *a priori* what word it is showing to the user. Despite the differences, this could still be a direction worth exploring.

6.3.6 Labeling of trials (semantic property)

The labeling of the stimuli used is based solely on lexical properties, but it is unclear how lexical properties are exactly related to the "easy" and "hard" mental states we are trying to detect; in other words, how lexical properties are related to detecting when the user is having difficulty or not. The assumption in the pilot study and in this work is that words with lexical properties

indicative of high difficulty can trigger the "hard" mental state, and any ERP elicited indicates the mental state. Future work could investigate this assumption further. Though lexical properties, such as the orthographic neighborhood size of pseudo-words, can elicit ERPs as shown in Laszlo et al.'s work, again, the ERPs might not be entirely indicative of the easy or hard mental state. A hard word can actually be easy for the user and vice versa, all depending on whether the user has learned that word before or not; in other words, a word's difficulty is neither universal nor transferable between subjects (although it could actually be so, for example, for 90% of the population). This means that a hard word might still be able to elicit an ERP simply by virtue of its lexical properties, even though the user finds the word easy; the ERP that arises could be attributed to other brain processes, not necessarily the "hard" mental state. Perhaps a more accurate way is to allow the user to label a given stimulus as being easy or hard after he or she is exposed to it; a scale of 1–5 could also be used. This form of labeling can be considered as a semantic property, a specific meaning given by the subject's brain, which could be more accurate than just using lexical properties. Sudre et al. used semantic properties based on common nouns which have universal concepts associated with them. For example, the noun "hammer", has the universal concept of a tool. Future work could also investigate whether the ERPs that arise from reading hard words are due to lexical properties or the "hard" mental state.

6.3.7 Modeling sentence trials

The ERP analyses show the potential of using the brain responses to words in the EEG signal to distinguish between easy and hard words. It could be worthwhile to do the same for sentences, by capturing the brain responses to each individual word in the sentence. But modeling a sentence trial is challenging, especially for the silent reading condition, since it is difficult to know which word the user is reading at a time. This difficulty could be overcome with the use of eye movement tracking, which also allows the EEG signal to be time-locked to each word. Another way is to use a probabilistic model that could segment the EEG signal based on the *expected* starting and ending points of each word, with reference to the average duration required to read a word silently as well as the entire duration of the sentence trial. This could be possible since the reading tutor knows *a priori* the sentence that is presented to the subject. Though this probabilistic model might not be very accurate, it is worthwhile to investigate the level of performance it can achieve. A machine learning algorithm could be devised to learn the time needed to read a word silently, taking note that this could vary between subjects.

Modeling of oral sentence trials could be done with the aid of recorded speech from the user. Speech recognition could identify the time at which each word is spoken, leading to an accurate segmentation of the EEG signal.

Another approach is to train a dynamic model such as the hidden Markov model (HMM) on the EEG signal for a sentence trial or on the sequence of brain responses to words, treating the input as a time series. Chiappa and Bengio [2004] trained input-output HMM (IOHMM) and HMM on EEG signal to distinguish between mental tasks, achieving an accuracy of about 66.7%. They suggested that performance could be improved by better capturing the EEG *rhythmic changes*,

and the exact starting point of the cognitive task in the signal. It is likely that different sentence stimuli could produce different rhythmic changes; this issue is made worse by the different reading abilities of the subjects. Indeed, extensive preliminary analyses (not shown in this work) that directly applied the HMM to the EEG signal did not work out well, partly because the analyses did not consider that EEG rhythmic changes could vary from subject to subject, stimulus to stimulus.

6.3.8 Effect of duration

Arguably, features extracted from longer trials would be more stable, that is, having a lower variance, since such trials have more EEG and power samples. The effect of duration might be even more pronounced for features extracted from the power spectrum since the sampling rate is only 8 Hz. Hard sentence trials have longer durations. If features extracted from hard sentence trials have less variance, it might be easier to predict a given hard trial correctly than it is to predict an easy one, for example, as shown in section 5.2.1 where hard trials have a higher recall. Future work could investigate how the duration of a trial could affect the variance and/or quality of the features extracted. If duration does affect the quality of an extracted feature, leading to an improvement in classification accuracy, then feature extraction itself would be exploiting duration, albeit indirectly.

Bibliography

ABDI, H. 2007. Bonferroni and Šidák corrections for multiple comparisons. In: N.J. SALKIND (Editor), Encyclopedia of Measurement and Statistics. Sage Publications, Thousand Oaks, CA, USA.

BAKER, R., D'MELLO, S., RODRIGO, M.M., GRAESSER, A. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies 68*(4), 223-241.

BAUMEISTER, J., BARTHEL, T., GEISS, K.R. and WEISS, M. 2008. Influence of phosphatidylserine on cognitive performance and cortical activity after induced stress. *Nutritional Neuroscience 11*(3), 103-110.

BENJAMINI, Y. and HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society 57*(1), 289-300.

BERKA, C., LEVENDOWSKI, D.J., LUMICAO, M.N., YAU, A., DAVIS, G., ZIVKOVIC, V.T., OLMSTEAD, R.E., TREMOULET, P.D., CRAVEN, P.L. 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine 78*(5, Suppl.), B231-B244.

BIZAS, E., SIMOS, P.G., STAM, C.J., ARVANITIS, S., TERZAKIS, D., MICHELOYANNIS, S. 1999. EEG Correlates of Cerebral Engagement in Reading Tasks. *Brain Topography 12*(2), 99-105.

CASHERO, Z. 2011. *COMPARISON OF EEG PREPROCESSING METHODS TO IMPROVE THE CLASSIFICATION OF P300 TRIALS*, Colorado State University, Fort Collins, Colorado, USA.

CHAWLA, N.V., JAPKOWICZ, N. and KOTCZ, A. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter 6*(1), 1-6.

CHI, Y.M., JUNG, T.-P. and CAUWENBERGHS, G. 2010. Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review. *IEEE Reviews in Biomedical Engineering 3*, 106-119.

CHIAPPA, S. and BENGIO, S. 2004. HMM and IOHMM Modeling of EEG Rhythms for Asynchronous BCI Systems. In *European Symposium on Artificial Neural Networks*, Bruges (Belgium), 2004, 199-204.

COLTHEART, M. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology 33A*, 497-505.

CROWLEY, K., SLINEY, A. and MURPHY, I.P.D. 2010. Evaluating a Brain-Computer Interface to Categorise Human Emotional Response. In *Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies*, 5-7 July, 2010, 276-278.

DAMBACHER, M., KLIEGL, R., HOFMANN, M. and JACOBS, A.M. 2006. Frequency and predictability effects on event-related potentials during reading. *Brain Research 1084*(1), 89-103.

DONOGHUE, J.P., NURMIKKO, A., BLACK, M. and HOCHBERG, L.R. 2007. Assistive technology and robotic control using motor cortex ensemble-based neural interface systems in humans with tetraplegia. *The Journal of Physiology*, 603–611.

FARWELL, L.A. and DONCHIN, E. 1988. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology 70*(6), 510–523.

FINGELKURTS, A.A., FINGELKURTS, A.A. and KÄHKÖNEN, R.K.T.A.S.B.V.P.O.J.S. 2006. Increased local and decreased remote functional connectivity at EEG alpha and beta frequency bands in opioid-dependent patients. *PSYCHOPHARMACOLOGY 188*(1), 42-52.

FISCH, B.J. 1991. Artifacts. In *Spehlmann's EEG Primer*, B.J. FISCH, Ed. Elsevier Publishing Company, Amsterdam, The Netherlands, 108–124.

FISHER, R.A. 1915. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika 10*(4), 507-521.

FISHER, R.A. 1921. On the `probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*, 3-32.

FRIEDERICI, A.D., PFEIFER, E. and HAHNE, A. 1993. Event-related brain potentials during natural speech processing: effects of semantic, morphological and syntactic violations. *Cognitive Brain Research 1*, 183-192.

FRISTON, K.J. and BÜCHEL, C. 2006. Functional Connectivity: Eigenimages and multivariate analyses. In *Statistical parametric mapping: the analysis of functional brain images*, Academic Press, 492-507.

HAAPALAINEN, E., KIM, S., FORLIZZI, J.F. and DEY, A.K. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 2010, 301-310.

HINZ, A. 1989. The Tower of Hanoi. *Enseignement Mathematique 35*, 239-321.

HUETTEL, S.A., SONG, A.W. and MCCARTHY, G. 2008. *Functional Magnetic Resonance Imaging*. Sinauer Associates.

JASPER, H.H. 1958. The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology 10*, 370-375.

JOYCE, C.A., GORODNITSKY, I.F. and KUTAS, M. 2004. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology 41*.

JUNG, T.-P., HUMPHRIES, C., LEE, T.-W., MAKEIG, S., MCKEOWN, M.J., IRAGUI, V. and SEJNOWSKI, T.J. 1998. Removing Electroencephalographic Artifacts: Comparison between ICA and PCA. In *Proceedings of the 1998 IEEE Signal Processing Society Workshop*, 1998, 63-72.

KAY, S.M. 1988. *Modern Spectral Estimation: Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ. 453-455 pp.

KUTAS, M. and FEDERMEIER, K.D. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology 62*, 621-647.

KUTAS, M. and HILLYARD, S.A. 1983. Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition 11*(5), 539-550.

LASZLO, S. and FEDERMEIER, K.D. 2011. The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology 48*(2), 176-186.

LEVENE, H. 1960. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press.

LIANG, M., ZHOU, Y., JIANG, T., LIU, Z., TIAN, L., LIU, H. and HAO, Y. 2006. Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport 17*(2), 209-213.

LUCK, S.J. 2005. *An Introduction to the Event-Related Potential Technique*. MIT Press.

LUO, A. and SULLIVAN, T.J. 2010. A user-friendly SSVEP-based brain–computer interface using a time-domain classifier. *Journal of Neural Engineering 7*.

LUTSYUK, N., ÉISMONT, E., PAVLENKO, V. 2006. Correlation of the characteristics of EEG potentials with the indices of attention in 12-to 13-year-old children. *Neurophysiology 38*(3), 209-216.

MAROSI, E., BAZÁN, O., YAÑEZ, G., BERNAL, J., FERNÁNDEZ, T., RODRÍGUEZ, M., SILVA, J., REYES, A. 2002. Narrow-band spectral measurements of EEG during emotional tasks. *International Journal of Neuroscience 112*(7), 871-891.

MITCHELL, T.M., HUTCHINSON, R., NICULESCU, R.S., PEREIRA, F., WANG, X., JUST, M. and NEWMAN, S. 2004. Learning to decode cognitive states from brain images. *Machine Learning 57*, 145-175.

MOSTOW, J. and BECK, J. 2007. When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In *Scale-Up in Education*, B. SCHNEIDER and S.-K. MCDONALD, Eds. Rowman & Littlefield Publishers, Lanham, MD, 183--200.

MOSTOW, J. and BECK, J.E. 2009. Why, What, and How to Log? Lessons from LISTEN. In *Proceedings of the Second International Conference on Educational Data Mining*, Córdoba, Spain, July 1-3, 2009, 269-278.

MOSTOW, J., CHANG, K.-M. and NELSON, J. 2011. Toward Exploiting EEG Input in a Reading Tutor. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, Auckland, NZ, 2011, 230-237.

NEUROSKY 2009a. Brain Wave Signal (EEG) of NeuroSky, Inc.

NEUROSKY 2009b. NeuroSky's eSense™ Meters and Detection of Mental State.

NEUROSKY 2012a. How to convert raw values to voltage?, http://support.neurosky.com/kb/technology/how-to-convert-raw-values-to-voltage.

NEUROSKY 2012b. NeuroSky - Brainwave Technology, http://www.neurosky.com/AboutUs/BrainwaveTechnology.aspx.

NG, A.Y. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first International conference on Machine learning*, 2004, 78.

NIST 2012. NIST/SEMATECH e-Handbook of Statistical Methods. In: J. PRINS (Editor), http://www.itl.nist.gov/div898/handbook/prc/section4/prc43.htm.

NUNEZ, P.L. and SRINIVASAN, R. 2005. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, USA, New York.

RABINER, L.R. and GOLD, B. 1975. *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

RANGASWAMY, M., PORJESZ, B., CHORLIAN, D.B., WANG, K., JONES, K.A., BAUER, L.O., ROHRBAUGH, J., O'CONNOR, S.J., KUPERMAN, S., REICH, T. and BEGLEITER, H. 2002. Beta power in the EEG of alcoholics. *Biological psychiatry 52*(8), 831-842.

SHARBROUGH, F., CHATRIAN, C., LESSER, R., LUDERS, H., NUWER, M. and PICTON, T. 1991. American Electroencephalographic Society guidelines for standard electrode position nomenclature. *Clinical Neurophysiology 8*, 200–202.

SINGER, W. 1999. Neuronal synchrony: A versatile code for the definition of relations? *Neuron 24*, 49–65.

SMIT, D.J.A., STAM, C.J., POSTHUMA, D., BOOMSMA, D.I. and GEUS, E.J.C.D. 2008. Heritability of "small-world" networks in the brain: A graph theoretical analysis of resting-state EEG functional connectivity. *Human Brain Mapping 29*(12), 1368-1378.

SNEDECOR, G.W. and COCHRAN, W.G. 1989. *Statistical Methods*. Iowa State University Press.

SRINIVASAN, R., WINTER, W.R., DING, J. and NUNEZ, P.L. 2007. EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of Neuroscience Methods 166*(1), 41-52.

STAM, C.J., HAAN, W.D., DAFFERTSHOFER, A., JONES, B.F., MANSHANDEN, I., WALSUM, A.M.V.C.V., MONTEZ, T., VERBUNT, J.P.A., MUNCK, J.C.D., DIJK, B.W.V., BERENDSE, H.W. and SCHELTENS, P. 2009. Graph theoretical analysis of MEG functional connectivity in Alzheimer's disease. *Brain 132*(1), 213-224.

STOLINE, M.R. 1981. The Status of Multiple Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs. *The American Statistician 35*(3), 134-141.

STROOP, J.R. 1935. Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology 18*(6), 643-662.

SUDRE, G., POMERLEAU, D., PALATUCCI, M., WEHBE, L., FYSHE, A., SALMELIN, R. and MITCHELL, T. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage 62*(1), 451-463.

SUPEKAR, K., MENON, V., RUBIN, D., MUSEN, M. and GREICIUS, M.D. 2008. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. *PLOS Computational Biology 4*(6).

WALLSTROM, G.L., KASS, R.E., MILLER, A., COHN, J.F. and FOX, N.A. 2004. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. *International Journal of Psychophysiology 53*(2), 105-119.

WEISS, S. and RAPPELSBERGER, P. 2000. Long-range EEG synchronization during word encoding correlates with successful memory performance. *Cognitive Brain Research 9*(3), 299–312.

WELCH, P.D. 1967. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics 15*(2), 70-73.

WHITHAM, E.M., LEWIS, T., POPE, K.J., FITZGIBBON, S.P., CLARK, C.R., LOVELESS, S., DELOSANGELES, D., WALLACE, A.K., BROBERG, M. and WILLOUGHBY, J.O. 2008. Thinking activates EMG in scalp electrical recordings. *Clinical Neurophysiology 119*(5), 1166-1175.

WHITHAM, E.M., POPE, K.J., FITZGIBBON, S.P., LEWIS, T., CLARK, C.R., LOVELESS, S., BROBERG, M., WALLACE, A., DELOSANGELES, D., LILLIE, P., HARDY, A., FRONSKO, R., PULBROOK, A. and WILLOUGHBY, J.O. 2007. Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clinical Neurophysiology 118*(8), 1877–1888.