APPLICATIONS OF SEQUENTIAL HYPOTHESIS TESTING TO THE DEVELOPMENT OF NON-INVASIVE BRAIN-COMPUTER INTERFACES

BY

JAMES J. S. NORTON

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Associate Professor Timothy Bretl

ABSTRACT

Sequential hypothesis testing is applied in this thesis to make two contributions to the classification of electroencephalography (EEG) data for use in non-invasive brain computer interfaces (BCIs).

The first contribution is a variable window-length classification method for use in a steady-state visual evoked potential (SSVEP)-based BCI. Instead of relying on a fixed window-length strategy—where a pre-specified amount of data is collected before classification is attempted—a sequential probability ratio test is used and data is collected until a confidence threshold is met. This variable window-length strategy was tested using a simple experiment where one of five visual stimuli were presented one at a time to three participants. An analysis of the data collected during this simple experiment show that the information transfer rate was improved by 43% when using the variable window-length strategy as compared to the fixed window-length strategy.

The second contribution is an analysis showing that it is possible to classify expected versus unexpected endings to strongly constrained sentences at better than chance accuracies using single trials of EEG. This better than chance classification accuracy is demonstrated for features based on two different event-related potentials (ERPs) elicited in response to the neural processing of meaning-related information—the N400 and the frontal positivity (FP). Using an existing dataset, classification accuracies were computed for features based on each of the two brain signals for three different classifiers (Naïve Bayes [NB]; linear discriminant analysis [LDA]; and support vector machines [SVM]), three different electrode groupings, and three different ways of analyzing the individual trials (single trial classification; classification after averaging multiple trials; and the sequential classification of trials using a sequential probability ratio test [SPRT]). Using single trials of EEG, features based on the N400, and all 26 EEG electrodes, classification accuracy with an LDA classifier was 59.96%. In analyses with features based on both the N400 and FP, classification accuracies were higher (59.25%) when three trials were averaged together before classification than they were with single trials. The initial tests with the SPRT classifier were mixed. Classification accuracies were higher for SPRT (when using the same features) than for NB (but not for LDA or SVM) when single trials or the average of multiple trials were used for classification. The analyses of classification accuracies using features based on the N400 and FP, development of the ERP Classification GUI, and the SPRT classifier represent significant steps toward the development of a new BCI paradigm based on the processing of meaning-related information.

Stephanie, thank you for sharing your life with me, the dreams we've had, the challenges we've faced, and the adventures that are yet to come.

To my family (especially my mom, dad, Tom, and Cassandra), friends, and mentors, thank you for your love and support.

To The Godfather (and of course The Godmother), thank you for giving me that little extra push.

ACKNOWLEDGMENTS

Thank you to all of the faculty, staff, and students at the University of Illinois at Urbana-Champaign who contributed to this thesis by helping me grow into both a neuroscientist and an engineer.

Special thanks to Tim, Scott, Doug, Kara, Sam, Jonathan, Rama, Gary, Minh, Diane, Rakesh, Erik, Dave, Cybelle, Danielle, and Ryan.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 SEQUENTIAL SELECTION OF WINDOW-LENGTH FOR IMPROVED SSVEP-BASED BCI CLASSIFICATION 5

2.1 Abstract 5

2.2 Introduction 5

2.3 Methodology 7

2.4 Experimental Setup 10

2.5 Results 12

2.6 Discussion 12

2.7 Conclusion and Future Work 14

CHAPTER 3 TOWARD A BRAIN-COMPUTER INTERFACE BASED ON THE PROCESSING OF MEANING-RELATED INFORMATION 15

3.1 Abstract 15

3.2 Introduction 16

3.3 ERP Classification GUI 18

3.4 Method 24

3.5 Results 32

3.6 Discussion 40

CHAPTER 4 CONCLUSION 46

REFERENCES 47

CHAPTER 1

INTRODUCTION

Brain-computer interfaces (BCIs) are devices that enable people or animals to communicate information to artificial systems through brain activity. The first BCI—as well as the term *brain-computer interface* itself—was described in 1973 by Jacques J. Vidal [1] based on his work (and that of his students, see Glassman [2] and Schwartzmann [3]) at the University of California, Los Angeles. Since their invention, the majority of research on the development of BCIs has the goal of enabling people who have severe motor disabilities—such as late stage amyotrophic lateral sclerosis (ALS) [4, 5, 6]—to communicate. In these individuals, ordinary methods of communication (e.g., speaking, sign language, eye movements, etc.) are no longer possible. Since BCIs do not rely on muscle activity, they may provide the only means for these individuals to communicate with the world around them.

Most BCIs (including the BCI described in Vidal's [1] original experimental setup), measure brain activity using electroencephalography (EEG). EEG measures the electrical activity, most likely from excitatory and inhibitory post synaptic potentials, naturally generated in the brain [7]. This electrical activity, when synchronous, causes small changes in voltage that can be measured from the surface of the scalp using electrodes. These signals are subsequently filtered (generally between 0.1 and 100 Hz), amplified, and digitized. The resulting digital signals are then translated into commands and these commands are used to control artificial systems. This translation process—the mapping of EEG signals into computer commands—is referred to as classification and represents a critical component in the design of BCIs.

The challenges inherent to the design of a classification system for BCIs have been apparent since Vidal's original experiments [1]. EEG signals have a low signal-to-noise ration (SNR); there is only a limited amount of data that can be collected during a laboratory experiment, and there are large differences in the signals elicited from different paradigms and individuals.

As such, the overall design of the classification system depends on a set of design choices, including: the BCI paradigm, preprocessing of the EEG signals, feature extraction, and the choice of a classifier.

The design of any classification system for use in a BCI depend on what brain signal is acquired and which protocol is used to acquire it. There are many different brain signals that can be used to control a BCI, such as imagined movements (motor imagery BCIs), responses evoked by unexpected stimuli (P300), or frequency entrained responses to repetitively flashing lights (steady-state visual evoked potentials [SSVEPs]). Each of these signals has distinct advantages and disadvantages. For example, the signals used to control a motor imagery BCI can be generated endogenously by users [8], whereas P300 and SSVEP-based BCIs require external stimuli to evoke brain activity. The experimental protocol used to acquire a signal from a user is also important. A BCI that elicits a P300 using a picture of a face has better performance than one that uses simple flashes of light [9]. In SSVEP-based BCIs, the amplitude of the evoked response is dependent on the frequency of the stimulus [10]. Although it is beyond the scope of this thesis, the specific ways in which a user interacts with the BCI has a major impact on classification performance; Akce, Norton, and Bretl [11] demonstrated the effect of the design of the user interface on the performance of SSVEP-based BCIs. Finally, the majority of research on BCIs has concentrated on the use of motor imagery, the P300, or SSVEPs, but there are a number of other brain signals that can be used to control a BCI. These include signals related to visual spatial attention [12], spatially distinct sounds [13], and the processing of meaning-related information [14] (discussed in Chapter 3).

After the signal has been acquired, it can be manipulated to eliminate artifacts or reduce noise. Steps taken to accomplish these goals are collectively known as preprocessing. There are many different ways to preprocess EEG data, but three of the most common are re-referencing, filtering, and artifact rejection. All EEG signals reflect voltage differences between two electrodes, the specific electrode being discussed and a reference electrode. The process of changing the electrode used as a reference after data acquisition is known as re-referencing and is common when working with EEG. A number of noise sources can be eliminated through filtering. A low-pass filter, for instance, may be used to remove high-frequency noise caused by muscle artifacts or a notch filter may be used to remove power line noise [15]. Slightly more complicated than standard filtering is the removal of artifacts caused by eye movements. Fascinatingly, Vidal [1] called the removal of eyeblinks "solved to a large extent". Yet, researchers are still investigating methods to remove eyeblink artifacts from EEG recordings. These methods include regression [16] and independent component analysis (ICA) [17].

After the data is preprocessed, indicators of specific types of brain activity or "features" are extracted. The purpose of these features is to enable the definition of numerical differences between different types of brain activity. It is possible to define these numerical differences using the raw data, but this can degrade classification performance due to *overfitting*. One advantage of working with EEG data is that decades of experimental analyses have identified potential features for use in classification. The P300, for example, was first described by Sutton et al. in 1965 [18] and has been under investigation ever since. In the case of SSVEP-based BCIs, features based on the Fourier transform and canonical correlation analysis (CCA) [19] have proven to be effective for use in BCIs. After identifying these potential features, they must be evaluated. This evaluation process can be done in many different ways and represents an active area of research [20].

The final step in the design of a classification system is the choice of classifier. The simplest form of a classifier is a threshold. If the value of a feature or probability of some event exceeds a certain level then a selection is made. More advanced classifiers, such as naïve Bayes (NB), linear discriminant analysis (LDA), step-wise linear discriminant analysis, or support vector machines (SVM) exploit numerical differences between the features to discriminate between different types of brain activity (for a review see [21]). In addition to those listed above, here we consider a type of classifier known as a sequential classifier. Instead of making a decision immediately after data comes in, sequential classifiers may decide to collect more data before making a decision. In some ways they are similar to placing a threshold on top of another classification system.

In this thesis, we consider two classification problems for the development of BCIs. In Chapter 2, we consider the application of a sequential hypothesis test to the classification of visual targets used in an SSVEP-based BCI. Then, in Chapter 3, we present initial results from classification analyses—including preprocessing, feature extraction, and the accuracy of different classifiers—of EEG data collected during the reading of sentences. In this experiment, two types of sentences were presented. These two types of sentences differed in how they ended. These different endings are known to elicit specific brain signals believed to be related to the processing of meaning [22]. In addition, we describe a graphical user interface to enable the analysis of any properly formatted ERP dataset and report initial results on the use of the sequential probability ratio test (SPRT) to classify this data. Finally, in Chapter 4, we provide a brief conclusion.

CHAPTER 2

SEQUENTIAL SELECTION OF WINDOW-LENGTH FOR IMPROVED SSVEP-BASED BCI CLASSIFICATION

2.1 Abstract

Brain-computer interfaces (BCI) utilizing steady-state visual evoked potentials (SSVEPs) recorded by electroencephalography (EEG) have exciting potential to enable new systems for disabled individuals and novel controls for robotic and computer systems. To interact with SSVEP-based BCIs, users attend to visual stimuli modulated at predetermined frequencies. A key problem for SSVEP-based BCIs is to classify which modulation frequency the user is attending, for which there is an inherent trade-off between speed and accuracy. As SSVEP signals vary with time and stimulation frequency, a fixed-length data window does not necessarily optimize this trade-off. We propose a strategy, developed from sequential analysis, to vary the window-length used for classification. Our proposed technique adapts to the data, continuing to collect data until it is confident enough to make a classification decision. Our strategy was compared to a fixed window-length method using a simple experiment involving five frequencies presented individually to three participants. Using a canonical correlation analysis classifier to compare the proposed variable-length scheme to a standard fixed-length scheme, the variable-length approach improved the classifier information transfer rate by an average of 43%.

2.2 Introduction

The direct classification of neural signals for the benefit of those with motor impairments or as an alternate input modality has intrigued researchers

This work has been previously published as [23] and is co-authored by E. Johnson, D. Jones, D. Jun, and T. Bretl; ©2015 IEEE.

since first proposed more than 40 years ago [1]. Using electroencephalography (EEG), researchers have continued to develop brain computer interfaces (BCI) for communication and control, effectively translating electrical brain activity into artificial commands. One class of BCI systems, based on steady-state visual evoked potentials (SSVEP), relies on the brain's response to repetitive visual stimuli in the user's environment [24]. These stimuli, such as a flashing LED or computer screen, cause an entrainment between populations of neurons and the stimuli that can be selectively modulated through the allocation of attention [25]. Based on the neural signatures measured by EEG, this allocation of attention can be classified, effectively allowing the user to control the input of a computer system without the need for motor interaction. The BCI devices based on this paradigm have enabled text communication for disabled individuals [26], been demonstrated for use in robotic navigation tasks [27], and have been used as inputs for computer games [28].

Despite the promise demonstrated using these techniques their utility remains limited for several reasons, including: their reliability [29], ease of use, and overall system performance [30]. Considering the third of these three limitations, BCI systems are commonly compared based on their information transfer rate (ITR), measured in bits/second [31, 32]. Since ITR is a function of accuracy, latency, and the number of classes, various schemes for improving bitrate can be imagined. For instance, an easy way to improve overall information throughput is to increase the number of classes available to the user. Even if classification is relatively slow, a high ITR can be obtained from a system with 48 classes [33]. Another approach is to improve the speed and accuracy of classification.

Several classification methods currently dominate SSVEP-based BCI systems, including: power spectral density analysis (PSDA), minimum energy combination (MEC) [34], and canonical correlation analysis (CCA) [19]. In order to classify which of several frequencies a user is attending to, these classifiers wait for a fixed length of input EEG data before making a decision. This "window-length" is chosen *a priori* by the system designer and represents a trade-off between classification speed and accuracy.

However, there is no basis for assuming the window length must be fixed. Sequential analysis [35] provides a framework that makes selecting the stopping time (i.e., choosing the window length) a part of the classification task. This methodology is commonly used in a wide range of applications including medical diagnostics and quality assurance in manufacturing. Straightforward application of the sequential probability ratio test (SPRT) [36] (a standard sequential procedure), to SSVEP classifiers is not immediately obvious due to the lack of appropriate signal models.

In this chapter we propose a sequential test which is performed directly in the classifier-feature space. This test accounts for the classification rule, does not require modeling assumptions, and can handle nonlinear feature mappings. We develop our variable-length window method for the CCA classifier presented in [19], although the methodology outlined in this chapter may be extended to other existing classification methods. We demonstrate, based on a comparative study with a traditional CCA algorithm, that our variable-length window method allows for classification on a short window-length when signal quality is high, and automatically waits for more data when signal quality is low. This sequential approach is shown to uniformly outperform a fixed window-length method in terms of ITR and classification accuracy.

2.3 Methodology

In this section, we first introduce notation for the CCA classifier [19], which is then used to develop the proposed variable-length window method.

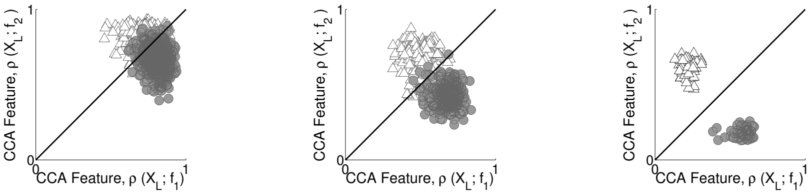
2.3.1 Stimulus-Frequency Classification Using CCA Features

The goal of the classification algorithm proposed in [19] is to infer the input frequency from multi-channel EEG data:

1. Assume there are K possible stimulus frequencies and N EEG channels.

2. The window length is assumed to be L samples long, which is L/F_s seconds, where F_s is the EEG sampling rate.

3. For a given stimulus frequency, CCA coefficients are computed using L samples from each of the N channels. The data are represented by a matrix X of dimension $L \times N$.



(a) $L = 16$ (0.125 s)

(b) $L = 64$ (0.5 s)

(c) $L = 256$ (2 s)

Figure 2.1: Plots of the CCA classifier decision rule for data of increasing window lengths. Test data with $f_1 = 6$ Hz and $f_2 = 8.57$ Hz were divided into three different lengths. As window length increases, the discriminability of the two classes increases considerably. ©2015 IEEE.

4. The largest CCA coefficient, defined to be $\rho(X; f)$ [37], is used as the feature for classification:

$$f^* = \arg\max_f \ \rho(X; f), \quad f = f_1, \ldots, f_K \tag{2.1}$$

where f^* is the classified frequency.

2.3.2 A Sequential Approach to a Variable-Length Window

The proposed sequential test essentially uses all of the data collected thus far to decide whether or not to continue collecting more data. To avoid requiring any additional assumptions about the data model, we develop a sequential test directly on the CCA feature space. This allows for direct comparison between the variable-length window and standard CCA classification.

The decision to continue collecting data depends on the current level of confidence as to which class the data belongs. The classification strategy in Equation (2.1) for two classes f_1 and f_2 (i.e., $K = 2$) is shown graphically as the solid line in each subplot of Figure 2.1. Each subplot corresponds to a different window length L, and each sample corresponds to the features extracted from an $L \times N$ block of raw EEG data. All features lying above this line are classified as f_2, and all features below it are classified as f_1. Thus, the classification boundary can be expressed as a hyperplane, h. Given the features of a data block (ρ_1 and ρ_2), the distance from the classification boundary, which expresses one's confidence, is given as the distance from the hyperplane:

$$d(\rho_1, \rho_2) = \left| h^T \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} \right| = \left| [1, -1] \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} \right| = |\rho_1 - \rho_2| \tag{2.2}$$

As shown in the sequence of scatter plots, when the window-length L increases, the confidence for every point also increases. This behavior has been quantified in the related problem of detecting a sinusoidal signal in Gaussian noise, where the Chernoff distance between the null and alternative densities (or more generally the deflection coefficient) increases by a factor of $\sqrt{L}$ [35].

The key idea of our approach is that even when L is small, there are *some* samples that could be classified correctly with high confidence. Identifying these samples and classifying them early may help to reduce the *effective* window length, when averaged over time.

Given a current window length L and data X (of dimension $L \times N$), we propose the following sequential test:

Require: confidence threshold, τ
Output: classification result, f^*

```
procedure VaryWindowLength($L$, $X$)
    if $d(\rho(X; f_1), \rho(X; f_2)) < \tau$ then
        $L \leftarrow L + B$                          ▷ increase window length
        $X \leftarrow [X; X_B]$                       ▷ augment new data
        return VaryWindowLength($L, X$)               ▷ repeat
    else
        return $f^*$ from (2.1)                       ▷ classification task
```

Here, B is defined to be a step-size, which is a basic unit of growth, and X_B is the data matrix that corresponds to the new B samples for each of the N channels. The classification task is carried out only when the minimum distance from the boundary is satisfied; until then, the window length is incrementally increased.

The threshold τ controls the trade-off between classification performance and average speed. In practice, choosing the threshold should be done *a* *priori*, and is exactly analogous to how window-length is chosen in a fixed-length strategy. Although the procedures are similar, the impact is quite different, as our variable-length strategy will be able to make classifications early, with minimal effect on performance. In fact, a fixed-length strategy is actually a special case of our proposed algorithm, with $\tau = 0$, and the initial L chosen to be the fixed window-length.

Finally, note that we focus on binary classification for the remainder of the chapter. The generalization ($K > 2$) is a straightforward extension using the geometric hyperplane interpretation.

2.4 Experimental Setup

2.4.1 Subjects

Experiments were conducted at the University of Illinois BCI lab with the authors as subjects. A James Long 128-channel EEG amplifier was used in conjunction with a National Instruments DAQ to digitize EEG signals at 128 Hz. The data were passband filtered from 1 – 30 Hz by the amplifier. EEG data were monitored during experimentation and logged by BCI2000 [38]. Participants were seated in a comfortable chair at 65 cm from a 24-inch BenQ XL2420T computer monitor. Scalp recording impedances were kept under 10 kΩ from sites (PO7, PO3, PO4, PO8, O1, OZ, O2) based on the 10-5 international system [39].

2.4.2 Stimuli and Procedure

Stimuli were implemented as a script in MATLAB in conjunction with the Psychophysics Toolbox [40]. The experiment consisted of six blocks of stimuli (6 Hz, 6.67 Hz, 7.5 Hz, 8.57 Hz, 10 Hz, and Null). During a block, a single stimulus of a given frequency was presented to the participant. The stimuli within each block were not randomized for this study as the emphasis was on a direct comparison between the two algorithms. Each stimulus was a square of identical size, subtending an angle 3.5° from fixation in each direction. Each block was composed of 20 trials each 15 seconds in length. At the beginning of each block, the participant was instructed to focus attention on the center of the flickering stimulus for the entire trial. There was a three-second interval between each trial. Stimulus onset was captured with a photodiode linked directly into the DAQ.

2.4.3 Analysis Techniques

All analyses were conducted offline following each experiment in the MATLAB environment. In order to quantify the difference in performance between the fixed-length window and our proposed variable-length window, the strategies were tested using a set of two-class classification problems. To form the two-class problems, each stimulus frequency was compared, one at a time, against all other frequencies. This gave a total of 10 comparisons for each of the three subjects. For each two-class problem, 20 trials of each frequency formed the testing dataset. Each two-class problem, therefore, had 10 minutes of testing data. A two-class CCA classifier was applied to the testing dataset, using both a fixed-length and variable-length window strategy.

Table 2.1: Performance of 3 subjects for fixed-length and variable-length window (©2015 IEEE).

	CCA (fixed-length window)		
Participant	Accuracy (%)	AWL (s)	Max ITR (bits/s)
A	94%	0.75	0.92
B	90%	0.58	0.96
C	88%	0.44	1.09
Average	91%	0.59	0.99

	CCA (variable-length window)			
Participant	Accuracy (%)	AWL (s)	Max ITR (bits/s)	% Max ITR Improvement
A	96%	0.62	1.24	35.3%
B	94%	0.48	1.44	49.8%
C	95%	0.48	1.59	45.2%
Average	95%	0.53	1.42	43.4%

2.4.4 Parameter Selection

Performance for the fixed-length window strategy was tuned by modifying the window length. The window length varied from 1/8 seconds to 1 second with 1/16 second steps. Changing the length of the fixed-length window trades off between decision speed and decision accuracy. For the variable-length strategy, the minimum block length was set at 1/8 second. To tune performance of the variable-length strategy, the threshold τ was varied from 0 to 0.3 in steps of 0.01. Varying the threshold trades off between classification speed and accuracy.

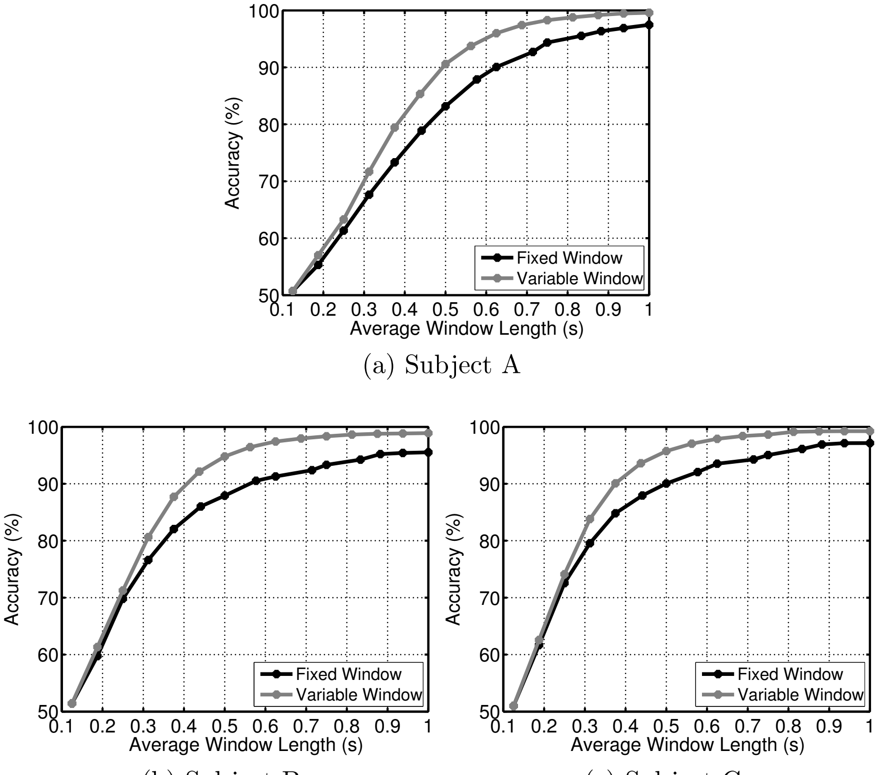
2.5 Results

For both fixed-length and variable-length strategies, the percent accuracy, average window-length (AWL), and ITR was calculated and averaged over all 10 two-class comparisons. The maximum ITR for each subject is reported in Table 2.1 for both fixed and variable-length strategies. The variable-length approach increases ITR by an average of 43% over all three subjects. For our subjects, using a variable-length strategy is an effective way to improve the performance of a CCA classifer.

In addition to the parameter configuration that maximizes ITR, it is possible to trace out the performance trade-off between classification accuracy and average window length for the two strategies. The results are summarized in Figure 2.2. For all three subjects, the classification performance curves for the variable-length window exceeds the curve for the fixed-length window. As the threshold for the variable-length window is lowered, it approaches the performance for the fixed-length window.

2.6 Discussion

The maximum ITR achieved by each of the three subjects in the variable-length window exceeded the performance of the fixed-length window. This is very encouraging, and shows uniform improvement of our variable-length approach. Performance numbers, however, are derived from idealized comparisons of ITR and are not directly comparable to performance numbers



(a) Subject A

(b) Subject B

(c) Subject C

Figure 2.2: Classification accuracy vs average window length for all three subjects averaged over 10 two-class comparison cases. For the fixed-length window, the performance curve is generated by changing the length of the fixed window. For the variable-length case, the performance curve is generated by altering the decision threshold. These curves were then linearly interpolated and averaged together across all 10 comparisons. For all three subjects, the variable-length window performs uniformly better, and approaches the performance of the fixed-window only for low average window lengths. ©2015 IEEE.

from real-time BCI systems. The relative improvement of the variable-length approach does suggest that incorporating this strategy will improve performance.

Using this simple experimental data and CCA classification, the comparisons of fixed-length and variable-length windows validate the intuition for applying sequential analysis. Because the quality of the data varies with time, a variable-length window can exceed the performance of a fixed-length window. We hypothesize that the variable-length strategy can be applied to other SSVEP classifiers, such as PSDA and MEC [34], provided the performance of the classifier improves as a function of the length of data used to classify.

Although the results in this study only consider the two-class case, the variable-length strategy can be extended to the multiple-class case. For CCA, this would involve finding the CCA correlation for each frequency of interest. These features would form an n-dimensional hypercube, with hyperplane decision boundaries.

Finally, this study does not apply any channel selection or denoising techniques. Again, this is because these results demonstrate the relative improvement of applying a variable-length window in place of a fixed-window, not an absolute performance metric.

2.7 Conclusion and Future Work

Since SSVEP signals vary with time, conditions, and stimulation frequencies, fixed-length windows are not necessarily optimal. This work proposed a variable-length window method for classification using CCA for SSVEP-based BCI. Our intuition about performance varying over time is consistent with the obtained results. In particular, a variable window-length strategy is shown to be uniformly better than a fixed window-length strategy, resulting in an average ITR improvement of 43%. As demonstrated, our proposed approach does not require any additional assumptions or signal models relative to existing CCA-based classifiers.

One implication of the achieved performance improvement is that faster, more accurate classification may improve the overall usability of SSVEP-based BCI systems. This may be particularly important for long-term applications, where attention and signal quality is expected to vary greatly due to effects such as fatigue and variable recording conditions.

As our approach naturally lends itself to extensions, further studies could consider multiple classes and other classification algorithms. Although current results demonstrate the relative improvement of the variable-length strategy over the fixed-length strategy, they do not yet demonstrate the performance of a real-time BCI system; future work could also explore the efficacy of this approach for online BCI tasks.

CHAPTER 3

TOWARD A BRAIN-COMPUTER INTERFACE BASED ON THE PROCESSING OF MEANING-RELATED INFORMATION

3.1 Abstract

In this chapter we show that it is possible to classify expected versus unexpected endings to strongly constrained sentences using single trials of EEG data. Furthermore, we demonstrate this for features based on two different brain signals elicited by the processing of meaning-related information—the N400 and the frontal positivity (FP). Using an existing dataset, we assessed the classification accuracy of features based on these two brain signals for: three different classifiers, three different electrode groupings, and three different analyses of the individuals trials (single-trial classification; classification after averaging multiple trials; and classification using the sequential probability ratio test [SPRT]). All of these analyses were done using the *Event-related potential (ERP) Classification GUI*, a MATLAB interface we developed for classification analyses on event-related potential datasets. When using 26 features—one from each electroencephalography (EEG) electrode—it was possible to classify expected versus unexpected endings (N400 - 58.49%; FP - 57.00%) at above chance accuracies. Averaging multiple trials together before classification improved accuracy from 55.81% with single trials to 59.25% after averaging three trials. The results obtained using SPRT were mixed (compared with the single trials or the average of multiple trials), classification accuracy was higher than Naïve Bayes, but not support vector machines or linear discriminant analysis. The analyses of classification accuracy using features based on the N400 and FP, development of the ERP Classification GUI, and the SPRT classifier represent significant steps toward a new brain-computer interface paradigm based on the neural processing of

This work was done in collaboration with C. Smith, R. Hubbard, E. Johnson, D. Dickson, K. Federmeier, and T. Bretl.

meaning-related information.

3.2 Introduction

Non-invasive brain-computer interfaces (BCIs) measure and interpret brain activity in near real-time. The majority of research on the development of BCIs concentrates on the use of a small number of paradigms; where each paradigm relies on a particular brain signal (motor imagery, the P300 event related potential [ERP], and steady-state visual evoked potentials [SSVEPs]). There are, however, a number of other brain signals that could probably be used as the basis of a BCI. For example, there are multiple neural signals known to occur in response to unexpected endings in strongly constrained sentences. These signals could be used in the development of BCIs based on the processing of meaning-related information.

For the past 40 years, the neural mechanisms involved in the processing of meaning-related information have been studied using ERPs [41]. By having individuals silently read sentences—presented visually, one word at time—while their brain activity was recorded, Kutas and Hillyard discovered that semantically inappropriate words cause a negative deflection in electroencephalography (EEG) activity. This negative deflection starts approximately 250 ms and peaks 400 ms after the onset of a semantically inappropriate word [41] and is thus now known as the N400.

Further investigation of the N400 over the intervening years has shown that the amplitude of the N400 is inversely related to *cloze* probability. In the sentence reading protocol of Kutas and Hillyard [41], cloze probability is the empirically determined probability that a specific word ends that sentence. In other words, as the likelihood of a specific ending word in a sentence increases, the greater the reduction in the N400 if that expected word is observed.

This relationship between cloze and the N400 was further explored in a study by Federmeier et al. [14]. In this study, two sets of sentences were shown to participants, a set of *strongly constrained* sentences and a set of *weakly constrained* sentences. In the strongly constrained sentences, a specific word completed the sentence more than 67% of the time. In the weakly constrained sentences, on the other hand, a specific word completed the sentence less than 42% of the time. Within these two sets, there were different potential endings, an *expected ending* and an *unexpected ending.* The expected ending was the word, determined through a survey, that was most likely to complete the sentence. The unexpected ending was a syntactically correct (and plausible) word, but unlikely to end that particular sentence. Critically, the cloze probability of the unexpected endings in both the strongly constrained sentences and weakly constrained sentences were controlled to be equal. This enabled the dissociation of the effects of constraint from the effects of cloze.

Federmeier et al. [14] replicated previous work showing that the reduction in the N400 is graded by cloze. The N400 was smallest for expected endings in strongly constrained sentences and second smallest for expected words in weakly constrained sentences. There did not appear, however, to be any difference in the amplitude of the N400 elicited in response to unexpected endings in the strongly constrained sentences versus unexpected endings in the weakly constrained sentences. Since the unexpected endings in these two conditions had equal cloze—but different levels of constraint—it was inferred that the N400 is not sensitive to the constraint of the sentence.

This study also revealed a second ERP component related to the processing of meaning-related information. This component was sensitive to constraint; it was only present in the EEG responses to strongly constrained sentences with unexpected endings. Labeled the *frontal positivity* (FP) for its location and positive voltage deflection, it has a frontal scalp distribution and appears 500–900 ms after the onset of the ending word. Federmeier et al. [14] cautiously theorized that it might reflect a cognitive processing cost of making strong predictions due to context.

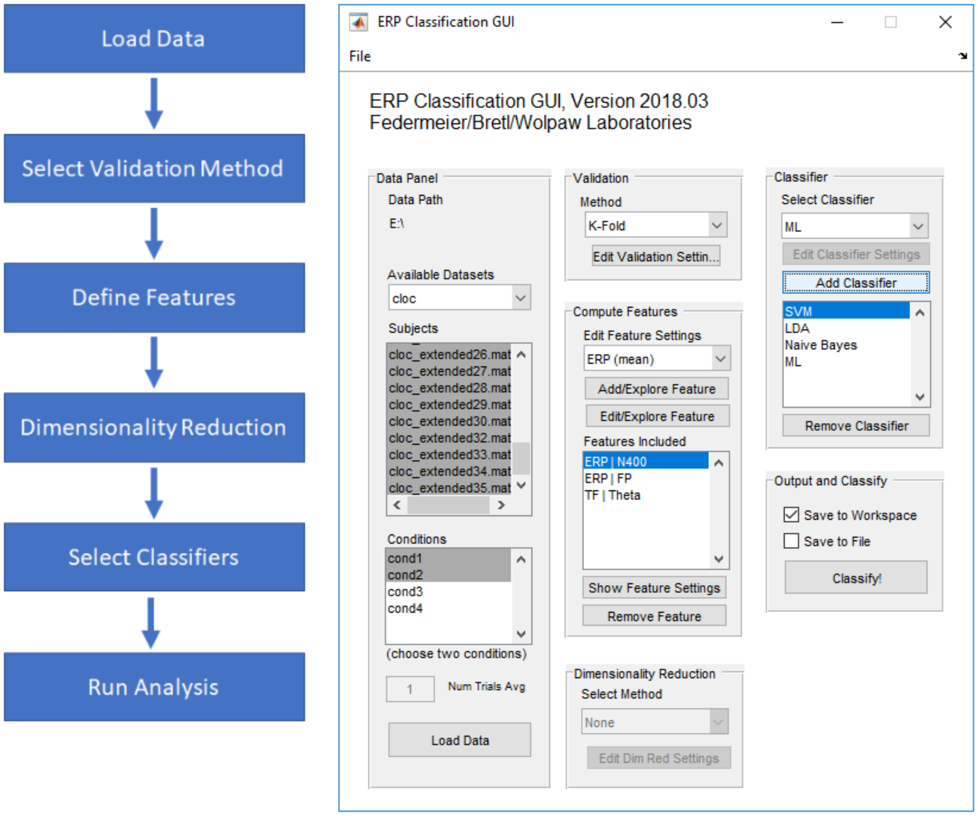
While EEG has been instrumental in the investigation of the neural mechanisms underlying the processing of meaning-related information, there have been few attempts [42, 43, 44] to build a BCI based on these signals. Furthermore, there has been no work (that we are aware of) that has explored the use of the multiple evoked (N400 and FP) EEG responses that occur in response to the processing of meaning-related information in a BCI. Such a BCI may have applications for healthy individuals, such as for recognition of implicit, but semantically important information in human-computer interfaces [45].

Here, we took a first step toward the development of a BCI based on the neural processing of meaning-related information. We investigated whether multiple neural responses to expected versus unexpected endings to highly constrained sentences could be classified at above chance levels. In other words, we analyzed whether it is possible to build a model of EEG responses to each of these two endings and subsequently use that model to predict whether someone had observed an expected or unexpected sentence ending in a new response. For this analysis, we used the data from the study of Federmeier et al. [14]. We based our classification *features*—numerical representation(s) of the EEG data that enable responses to expected versus unexpected sentence endings to be determined on known EEG responses to meaning-related information, namely the N400 and the FP. (See [46] for an in-depth discussion of machine learning topics such as features.) We also considered whether the combination of features from both of these brain signals improved classification accuracy over each brain signal in isolation. In addition, we compared several classifiers that use a fixed number of trials before attempting to classify the user's neural signals with a classifier that utilizes the sequential probability ratio test (SPRT) [36].

The results of our analyses show that strongly constrained sentences with expected endings can be separated from strongly constrained sentences with unexpected endings with better than chance accuracies. When using a single feature from each of the 26 EEG electrodes and the timing/frequency range of the N400, classification accuracy with a naïve Bayes (NB) classifier was above 60%.

3.3 ERP Classification GUI

An ongoing review we are conducting of the BCI literature shows that of the more than 1500 publications with the phrase "brain-computer interface" or "brain-machine interface" in the title or abstract (published before 2015), the main contribution of nearly 400 of these papers were improvement to classification systems (unpublished data from Norton). Furthermore, nearly 100 of these papers (and perhaps more) used datasets from one of the BCI competitions or another online data repository. Yet, the majority of researchers conducting these analyses devised a set of custom analysis scripts and treat each project as stand-alone and independent of one another.



ERP Classification GUI, Version 2018.03
Federmeier/Bretl/Wolpaw Laboratories

Data Path: E:\
Available Datasets: cloc
Subjects: cloc_extended26.mat, cloc_extended27.mat, cloc_extended28.mat, cloc_extended29.mat, cloc_extended30.mat, cloc_extended32.mat, cloc_extended33.mat, cloc_extended34.mat, cloc_extended35.mat
Conditions: cond1, cond2, cond3, cond4
(choose two conditions)
1 Num Trials Avg

Method: K-Fold
Edit Feature Settings: ERP (mean)
Features Included: ERP | N400, ERP | FP, TF | Theta
Select Method: None

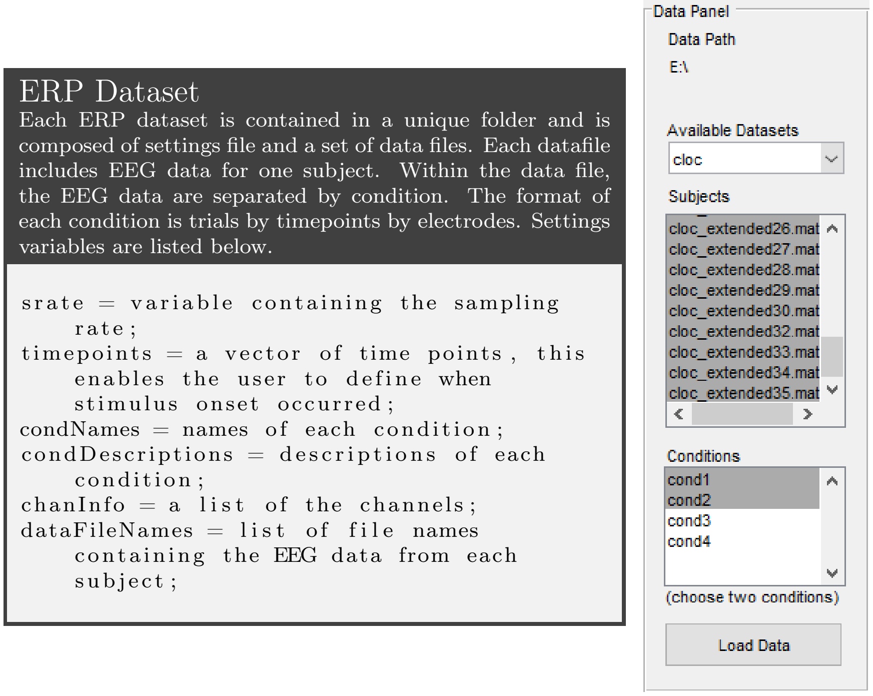
Select Classifier: ML
SVM, LDA, Naive Bayes, ML
☑ Save to Workspace
☐ Save to File

Figure 3.1: (left) The processing steps involved in most classification analyses, including those reported here. (right) The ERP Classification GUI.

Here, we create a system that allows many ERP datasets and classification system to be analyzed. At a high level, all ERP classification problems involve the same set of steps (Figure 3.1 [left]). Thus, instead of treating all ERP problems as being unique and requiring a custom classification system, we have created a graphical user interface (GUI) for handling ERP classification problems in general. This GUI enables a classification analysis to be performed on any ERP dataset of an appropriate structure. Once the data are loaded, the user is able to make selections that define the exact settings of the analysis to be performed. These settings include defining the subjects, conditions to be compared, cross-validation method, features, dimensionality reduction method, and classifier. Two design choice were made to reduce complexity. First, only binary classifications are available. If the dataset contains more than two distinct classes of data, the user must select which two types of data they would like to compare during the analysis. Second, we simplified the data format by assuming that artifact rejection is complete and that the individual trials have already been extracted. Existing EEG analysis toolboxes, such as FieldTrip [47] or EEGLab [48], can be useful for both of these pre-processing steps.

We provide additional details on this system—the ERP Classification GUI (Figure 3.1 [right])—here and then use it to perform the classification analyses on the data from Federmeier et al. [14].

3.3.1 Loading Data



ERP Dataset

Each ERP dataset is contained in a unique folder and is composed of settings file and a set of data files. Each datafile includes EEG data for one subject. Within the data file, the EEG data are separated by condition. The format of each condition is trials by timepoints by electrodes. Settings variables are listed below.

```
srate = variable containing the sampling
    rate;
timepoints = a vector of time points, this
    enables the user to define when
    stimulus onset occurred;
condNames = names of each condition;
condDescriptions = descriptions of each
    condition;
chanInfo = a list of the channels;
dataFileNames = list of file names
    containing the EEG data from each
    subject;
```

Data Panel

Data Path

E:\

Available Datasets: cloc

Subjects

cloc_extended26.mat
cloc_extended27.mat
cloc_extended28.mat
cloc_extended29.mat
cloc_extended30.mat
cloc_extended32.mat
cloc_extended33.mat
cloc_extended34.mat
cloc_extended35.mat

Conditions

cond1
cond2
cond3
cond4

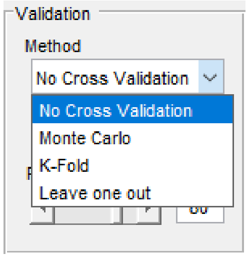
(choose two conditions)

Figure 3.2: (left) A description of the variables and data files contained within each ERP Classification GUI dataset. (right) The load menu enables the user to select which dataset (from a list of all of the available datasets), the individual data files that they wish to load, and the two conditions they would like to perform their classification analysis on. Multi-class classification is not currently possible.

The GUI is designed to handle multiple datasets with a common data structure. To load a dataset into the GUI for use, a data folder (containing all of the datasets) must first be defined. Within this data folder, each dataset must be contained within a subfolder and this subfolder must contain individual data files for each participant in the study and a single settings file.

After the data folder location is passed to the GUI, it is possible to load all of the datasets within that folder. The structure of each ERP Classification GUI dataset and the load data menu are shown in Figure 3.2.

3.3.2 Cross-Validation



Validation

Method

No Cross Validation

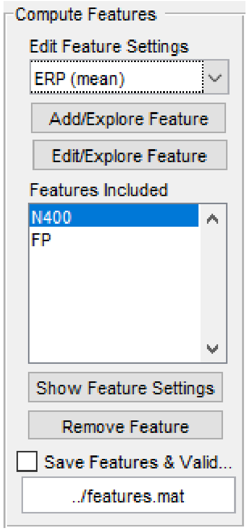
Figure 3.3: Cross-validation menu in the ERP Classification GUI.

When performing a classification analysis, the data must be separated into a training set (used to train the model) and a testing set (used to validate the model). Variations in the trials selected for each of these two datasets cause differences in the classification accuracy. To help with this problem, several different types of *cross-validation* methods are available (Figure 3.3). During cross-validation, the classification process is repeated multiple times. Each time, different subsets of the trials are placed into either the training or testing set. This produces a more stable assessment of the classification accuracy.

In our system, the user may select: no cross-validation, k-fold cross-validation, leave-one-out cross-validation, and Monte Carlo based cross-validation. As its name suggests, no cross-validation performs the requested analysis a single time. This is not as accurate, but can be useful for testing. In k-fold cross-validation, the data are divided into k chunks. All but one of these chunks is used for training the data and then the last chunk is used for testing the data. The system then iterates through the chunks, so that every chunk is used as the testing set at least once. Leave-one-out cross-validation is equivalent to k-fold cross-validation where k is equal to the number of trials in the data. In other words, all but one of the trials are used in the training set. k-fold cross-validation and leave-one-out cross-validation are discussed further in Lemm et al. [49]. The final cross-validation method available is

Monte Carlo based cross-validation. In this method, a user determined portion of the data to be placed in the training set and the rest of the data are used in the testing set. Once this proportion is defined, individual trials of data are randomly selected to be placed in the training or testing set and the classification analysis is performed. This process is completed many times and the classification accuracy is defined as the average across all of those repetitions.

3.3.3 Feature Selection



Edit Feature Settings
ERP (mean)

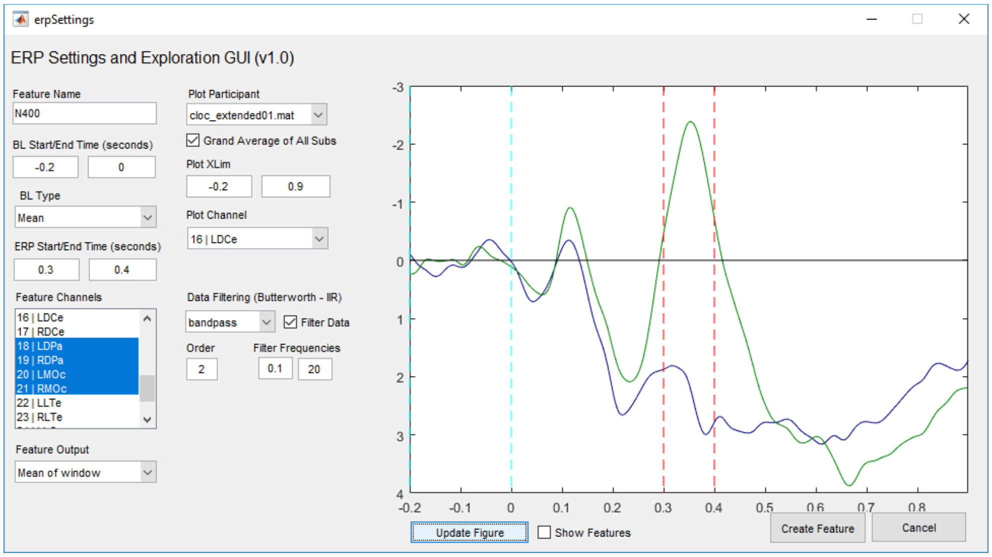
Features Included
N400
FP

□ Save Features & Valid...
../features.mat

Figure 3.4: Feature selection and analysis menu in the ERP Classification GUI.

The most important component of this GUI is its tool set for feature selection (Figure 3.4). Currently, two kinds of features are available for use in the classification analysis: time-domain features and frequency-domain features.

When we refer to time-domain features, we are considering traditional methods of ERP analysis [15]. These includes features such as peak amplitude and peak onset latency. The main GUI has a time-domain feature sub-system that allows users to analyze ERPs and determine which of these features could be useful for classification (Figure 3.5). There is a similar sub-system for analyzing frequency-domain features that is beyond the scope of


erpSettings

ERP Settings and Exploration GUI (v1.0)

Feature Name: N400

BL Start/End Time (seconds): -0.2 0

BL Type: Mean

ERP Start/End Time (seconds): 0.3 0.4

Feature Channels: 16 | LDCe, 17 | RDCe, 18 | LDPa, 19 | RDPa, 20 | LMOc, 21 | RMOc, 22 | LLTe, 23 | RLTe

Feature Output: Mean of window

Plot Participant: cloc_extended01.mat

☑ Grand Average of All Subs

Plot XLim: -0.2 0.9

Plot Channel: 16 | LDCe

Data Filtering (Butterworth - IIR): bandpass ☑ Filter Data

Order: 2

Filter Frequencies: 0.1 20

Figure 3.5: ERP Classification GUI time-domain feature analysis window.

this chapter. In ERP analysis, frequency-domain features are of interest to analyze changes that are not phase-locked to the stimulus.

3.3.4 Dimensionality Reduction

If there are too many features, the classifier may *overfit* the data. If overfitting occurs, the classifier will work well with the data used to create the model, but will not work well with new data. One way to avoid this is to choose fewer features. Another way to solve the problem is to try to represent a larger number of features in a lower-dimensional space. This process, known as dimensionality reduction, can improve classifier accuracy under the right conditions. In the GUI, the only dimensionality reduction method that has been implemented is principal component analysis (PCA). A discussion of three different techniques for dimensionality reduction, including PCA, can be found in [50].

3.3.5 Classifier

The final step before initiating the classification analysis is selecting the classifier and parameters of the classifier. There are multiple classifiers programmed into the GUI (Figure 3.6), they include: naïve Bayes (NB), lin-

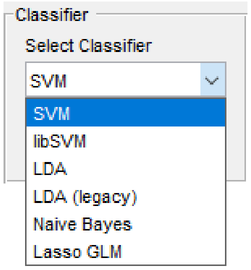


Figure 3.6: Classifier selection menu in the ERP Classification GUI.

ear discriminant analysis (LDA), support vector machines (SVM), maximum likelihood (ML), and the sequential probability ratio test (SPRT) [36]. Each of these classifiers can be used to separate the classes of EEG data, but each accomplishes that task in a slightly different way. SVM and LDA are discussed by Lotte et al. [21] in their review on classifiers for BCI, Myung [51] provides an excellent tutorial on ML (our version of ML assumes the data are normally distributed), and a discussion of NB can be found in [52].

3.3.6 Output

The results are reported to the user as a confusion matrix and an overall accuracy. This data can also be saved by the user for further analysis.

3.4 Method

The methods and materials used in this chapter have previously been described in detail by Federmeier et al. [14].

3.4.1 Participants

Thirty-two college aged subjects participated in this study (16 women and 16 men). None of the subjects had any prior history of neurological illness. Due to technical issues with the data, one subject was excluded from the classification analysis.

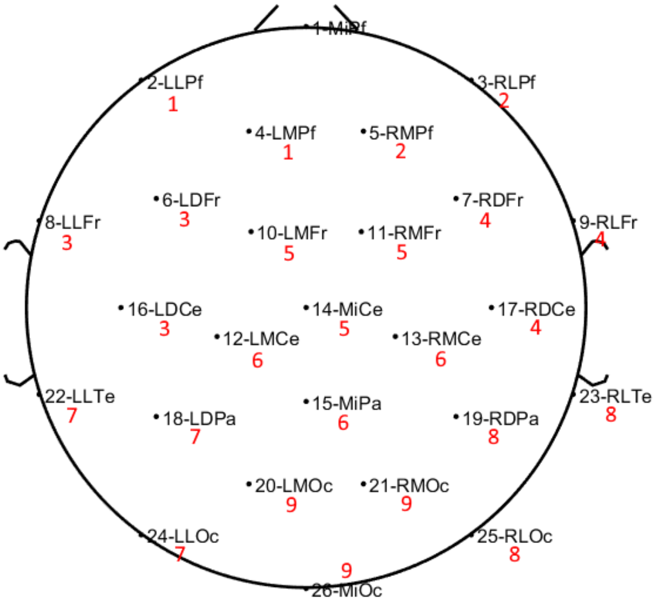


Figure 3.7: Map of equally spaced EEG electrode locations [14] including both electrode number and name. In addition, each electrode was assigned to a cluster of 2–3 electrodes. The cluster that the electrode was assigned to is denoted below each electrode name (in red). The figure was produced using the *topoplot* function in EEGLab [48].

3.4.2 Signal Acquisition

A Grass amplifier was used to record EEG activity from 27 electrodes (Figure 3.7; 26 equally spaced electrodes [sometimes referred to as a geodesic montage [14]] and an electrode on the right mastoid) at a sampling rate of 250 Hz. During acquisition, the data were recorded in reference to the left mastoid at impedances of less than 5 kΩ and analog filtered (using a bandpass filter [0.01–100 Hz] and a notch filter [60 Hz]). In addition to the EEG electrodes, two channels of electrooculogram (EOG) data were also recorded. These were used to detect horizontal eye movements and eyeblinks.

3.4.3 Procedure

Each of the participants read 282 sentences (from four conditions) while their EEG was recorded. The four conditions were defined by two variables: constraint and ending. Sentences could either be strongly constrained or weakly constrained. The constraint of a sentence was determined by that sentence's cloze. Full details on how cloze was determined can be found in Federmeier et al. [14]. For the present analysis, we considered the 141 strongly constrained sentences used in the study of Federmeier et al. [14]. As previously discussed, these strongly constrained sentences could have expected or unexpected ending words. Each strongly constrained sentence with an expected ending had a cloze probability of greater than 67%. The strongly constrained sentences with an unexpected ending had an overall cloze probability of 3.1%. An example sentence—taken from Table 2 in Federmeier et al. [14]—is shown in Table 3.1. Of the 141 sentences, 71 ended in an expected word, while the other 70 ended in an unexpected (but plausible) word.

Table 3.1: Example of a strongly constrained sentence with both an expected and an unexpected (but plausible) ending. This example is take from Table 2 of Federmeier et al. [14].

Sentence	Expected	Unexpected
He bought her a pearl necklace for her	birthday	collection

During the original experiments, each of the sentences were presented to the participants on a computer screen, one word at a time. Each sentence was preceded by a fixation (consisting of plus signs) that lasted for 500 ms. After the fixation, an initial blank screen of variable length (500–1200 ms) appeared. After the initial blank screen, each words was presented for 200 ms and followed by a blank screen for 300 ms. After the final word, there was a 3000 ms pause between sentences. We refer to the EEG recorded during the reading of the ending word of a single sentence as a *trial*.

3.4.4 Data Processing

Re-Referencing. The data were originally referenced to an electrode placed on the left mastoid. During data processing, all of the data was re-referenced to the average of two electrodes, the electrode on the left mastoid and another electrode on the right mastoid.

Bandpass Filtering. After the data were re-referenced, they were digitally filtered using a sixth-order (zero-phase) IIR-bandpass filter with a passband between 0.5 and 100 Hz. This filter helped to remove high-frequency noise due to muscle activity or harmonics of 60 Hz as well as removed low-frequency components due to wire movement.

Artifact Rejection and Filtering. Before the data was imported into the ERP Classifier GUI, it was analyzed for artifacts. The sources of these artifacts included amplifier blocking, horizontal eye movements, and eyeblinks.

Artifacts were labeled by an experienced investigator. Trials with horizontal eye movements and amplifier blocking were removed from the dataset. Eyeblinks were handled differently depending on the percent of trials with blinks. If less than 15 percent of the trials contained eyeblinks, then eyeblinks were simply rejected. If more than 15 percent of the trials contained eyeblinks, then independent component analysis (ICA) was used to filter eyeblinks from the data. ICA components were calculated using the EEGLab toolbox in MATLAB [48]. Components related to eyeblinks were chosen by the investigators.

3.4.5 Data Analysis

The goal of this analysis was to assess how well two different sentence endings (expected and unexpected) from strongly constrained sentences could be classified using EEG data. All data analysis was performed in MATLAB using the ERP Classification GUI described in Section 3.3. All classification analyses used the Monte Carlo method for cross-validation (100 repetitions). These analyses were performed on the data from each of the 31 subjects included in this study. All results are reported for the average across subjects except where noted otherwise.

We conducted analyses to assess the classification accuracy of features based on two brain signals (and their combination) related to the processing of meaning-related information. These analyses considered three different electrode groupings (single electrode, cluster of neighboring electrodes, and all electrodes); three different classifiers (NB, SVM, LDA); and three analyses of the individual trials (single trial classification, classification after averaging multiple trials, and the sequential classification of trials using SPRT).

Features. We analyzed the classification accuracy of features based on two different brain signals known to be elicited by unexpected endings to highly constrained sentences, the N400 [41] and FP [14].

The N400 is a negative centro-parietal deflection in EEG voltages in response to the integration of meaningful information [53]. For the purpose of this classification analysis, we define a feature based on the N400 as average the voltage occurring 320–420 ms after the onset of the ending word. Similarly to the analyses performed by Federmeier et al. [14], we further filtered our data using an IIR bandpass filter from 0.001–20 Hz before feature extraction. The grand average N400 for *MiPa* is shown in Figure 3.8. Before measurements were taken, all of the individual trials were baselined to the average of -200–0 ms before stimulus onset.

Table 3.2: For our analyses, three different groupings of EEG electrodes were considered. One of these grouping was based on clusters of 2–3 neighboring electrodes. The clusters (and electrodes included in each cluster) are listed.

Cluster	Electrodes
1	2,4
2	3,5
3	6,8,16
4	7,9,17
5	10,11,14
6	12,13,15
7	18,22,24
8	19,23,25
9	20,21,26

In addition to the well-established N400, Federmeier et al. [14] reported a second ERP—referred to as FP—elicited in response to unexpected endings in strongly constrained sentences. Our definition of features based on FP was the average voltage 620–720 ms after the onset of the ending word. We used the same baseline for FP as we used for the N400. Since Federmeier et al. [14] noted that this frontal positivity was slow, we low-pass filtered the data at 5 Hz before extraction. The grand average frontal positivity for *RMPf* is shown in Figure 3.9.

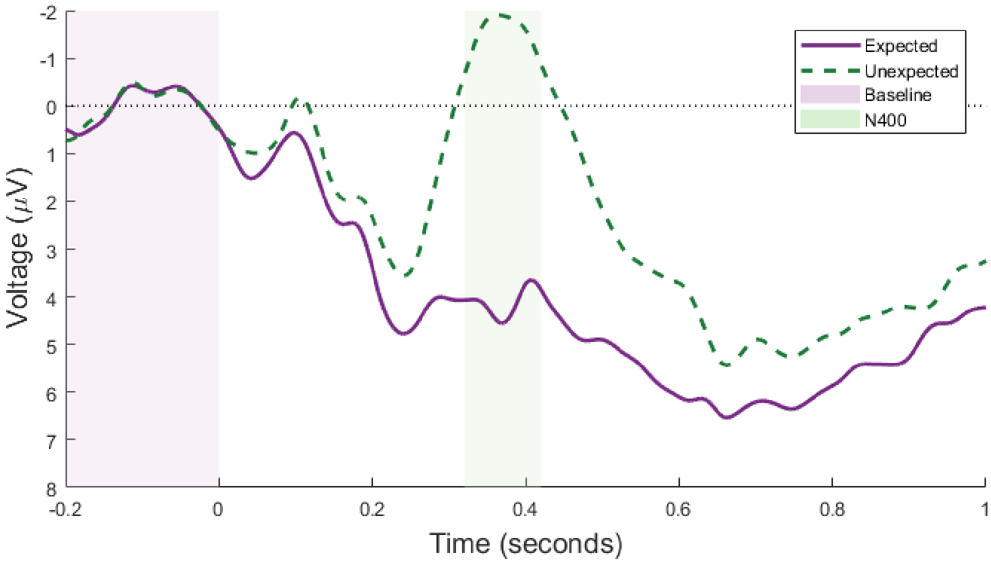


Figure 3.8: Grand average ERP for strongly constrained sentences at *MiPa*. The ERP elicited by expected versus unexpected endings are shown in dark purple and dark green respectively. The time window of the baseline is shaded in light purple. The time window of the N400 is shaded in light green.

3.4.6 Electrode Grouping

For the electrode analyses we exclude *a priori* knowledge of the scalp distribution of each of the three features and instead chose to analyze differences in classification accuracy using three different electrode grouping methods. The first method considered the classification accuracy of the features when using individual electrodes. The second method of grouping used *clusters* of two to three electrodes (Figure 3.7, Table 3.2). The analyses on the clusters considered each electrode in the cluster individually (i.e., not the average of the electrodes). Thus, each *observation*—features extracted from one trial—in this analysis had two to three features (as opposed to one feature in the single electrode analyses). The final used features based on all 26 EEG electrodes (i.e., there were 26 features).

3.4.7 Classifiers

During the aforementioned analyses we compared the classification accuracy of our features and electrode groupings using three different classifiers: SVM, LDA, and NB.

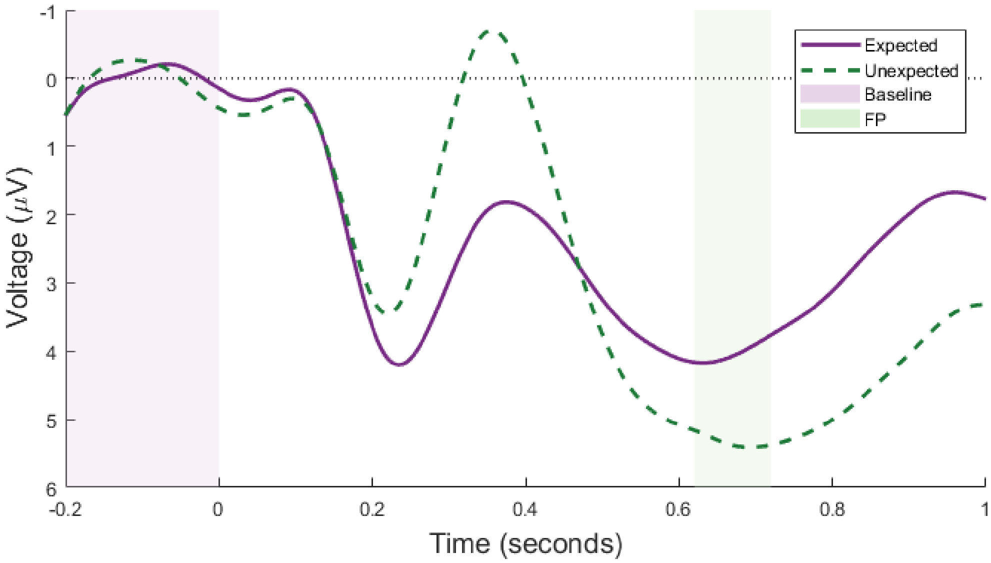


Figure 3.9: Grand average ERP for strongly constrained sentences at *RMPf*. The ERP elicited by expected versus unexpected endings are shown in dark purple and dark green respectively. The time window of the baseline is shaded in light purple. The time window of the frontal positivity is shaded in light green. The data in this figure was low-pass filtered at 5 Hz before averaging and plotting.

3.4.8 Different Analyses of the Individual Trials

Finally, we compared three different analyses of the individual trials. The first two analyses are referred to as *fixed* trial analyses, they use a predetermined number of trials every single time, these are: a single trial analysis and an analysis of classification accuracy when multiple (two or three) trials are averaged together before classification. These fixed trial analyses used the same classifiers as our other analyses (SVM, LDA, and NB). Finally, we also analyzed the data using a *sequential* classifier (SPRT). Instead of making a decision immediately after a trial is obtained, SPRT may decide to collect more data before classifying.

For these analyses, we considered the classification accuracy using features based on both the N400 and FP. The electrodes selected for each brain signal were based on *a priori* knowledge of the spatial distributions of the of the N400 (electrodes: 12, 13, 14, 15) and FP (electrodes: 1, 2, 3, 4, 5) [14, 54].

SPRT. SPRT [36]—following the notation and equations found in [55]—allows us to test the hypothesis that observations (x; a vector of features) of a random variable (X; all of the trials available for a condition) are drawn from one of two probability density functions ($f_i(x)$ for $i = 0, 1$). The null hypothesis is:

$$H_0 : X \sim f_0(x)$$

and the alternative hypothesis is:

$$H_a : X \sim f_1(x)$$

These hypotheses are tested using a likelihood ratio. For the first observation (x_1), the likelihood ratio is defined as:

$$\Lambda_1 = \frac{f_0(x_1)}{f_1(x_1)}$$

Here, we compute the log likelihood ratio, because its simplifies the calculations, as:

$$\log(\Lambda_1) = \log(f_0(x_1)) - \log(f_1(x_1))$$

and test this against two decision boundaries A and B. A and B are defined as:

$$A = \frac{\beta}{(1-\alpha)} \text{ and } B = \frac{(1-\beta)}{\alpha} \tag{3.1}$$

where α and β are the false positive and false negative error rates respectively. If:

$$\log(\Lambda_1) < \log(A) \tag{3.2}$$

then we conclude that the null hypothesis H_0 is true. Else If:

$$\log(\Lambda_1) > \log(B) \tag{3.3}$$

then we conclude that the alternative hypothesis H_a is true. Else:

$$\log(A) \leq \log(\Lambda_1) \leq \log(B) \tag{3.4}$$

then we conclude that we are not confident enough to make a decision and collect more data. In this case, we take another trial vector of features (x_2)

and repeat the process. The only difference is for trial $j = 2...n$:

$$\log(\Lambda_n) = \sum_{j=1}^{n} \log(f_0(x_j)) - \log(f_1(x_j)) \tag{3.5}$$

In our specific case, we assume that X, $f_1(x)$, and $f_2(x)$ are well modeled as multivariate normal distributions. Thus, we first need to estimate the parameters (μ_i and Σ_i) for $f_0(x)$ and $f_1(x)$. This can be done using the training data. μ_i is vector of composed of the mean values of each feature from the training data for class i and Σ_i is the covariance matrix of the training data for class i.

Once we have the parameters, we calculate the log likelihood (according to the equation given in [56]) of a new data sample under each of the models as:

$$\log(f_i(x_n)) = -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma_i| - \frac{1}{2}\sum_{j=1}^{n}(x_j - \mu_i)^T\Sigma_i^{-1}(x_j - \mu_i)$$

where n represents the number of data samples being tested and r is the dimensionality of the multivariate normal distribution. In our case, since we are considering the data one sample at a time, our equation simplifies to:

$$\log(f_i(x_n)) = -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x_n - \mu_i)^T\Sigma_i^{-1}(x_n - \mu_i) \tag{3.6}$$

For our analyses, we assume that $\alpha = \beta = \tau$ and test the accuracy of SPRT for $\tau = 0.01, 0.02, ..., 0.40$.

3.5 Results

The results are reported for classification using features based on each of the two brain signals (by electrode grouping and classifier) and then for the three different analyses of the individual trials. For each classification analysis, significant increases in classification accuracy (from chance) were determined using a random permutation test (Test 1 in [57]). For this random permutation test, we used the same exact classification system as in our other analyses, except: (1) we randomized the class labels and (2) we performed 500 repetitions of each classification analysis. No corrections for multiple comparisons were performed.

3.5.1 Brain Signal and Electrode Grouping

Single Electrode. Classification accuracies when using a single a feature from a single electrode were not significantly better than chance for any electrode, classifier, or feature.

Cluster of Neighboring Electrodes. When using clusters of neighboring electrodes, there were clusters where the classification accuracy was slightly (but significantly) better than chance for features based on both the N400 and FP.

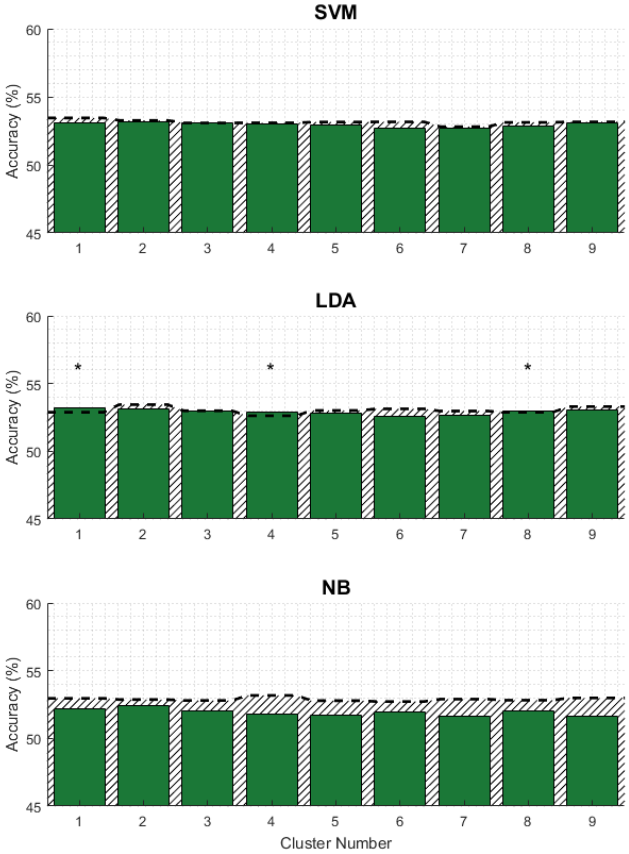


Figure 3.10: Classification accuracy for the analyses using features based on the N400 and clusters of neighboring electrodes (by cluster and classifier). Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance ($p < 0.05$) is denoted with a *.

For the N400 based features (Figure 3.10) with the LDA classifier, there were three clusters where the classification accuracy was better than chance: Cluster 1 (53.23%), Cluster 4 - (52.88%), and Cluster 8 - (52.94%).

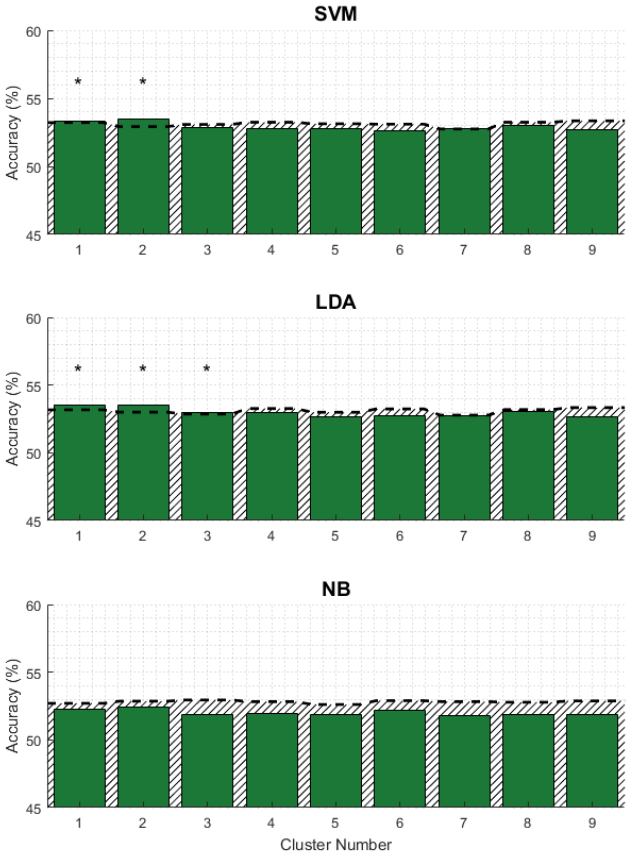


Figure 3.11: Classification accuracy for the analyses using features based on FP and clusters of neighboring electrodes (by cluster and classifier). Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance ($p < 0.05$) is denoted with a *.

For features based on FP (Figure 3.11), Clusters 1 and Cluster 2 had classification accuracy that was significantly better than chance for both SVM (Cluster 1 - 53.35%; Cluster 2 - 53.47%) and LDA (Cluster 1 - 53.55%; Cluster 2 - 53.53%). In addition, when using LDA, Cluster 3 had significantly better than chance classification accuracy (52.99%)

Since classification was better for the clusters of neighboring electrodes than for the individual electrodes (there were clusters of neighboring electrodes [but no individual electrodes] with better than chance classification accuracy) we report the classification accuracies of the individual subjects averaged across clusters in Figures 3.12 and 3.13. Visualizations of classification accuracy by subject and cluster of neighboring electrodes are shown in Figures 3.14 and 3.15.

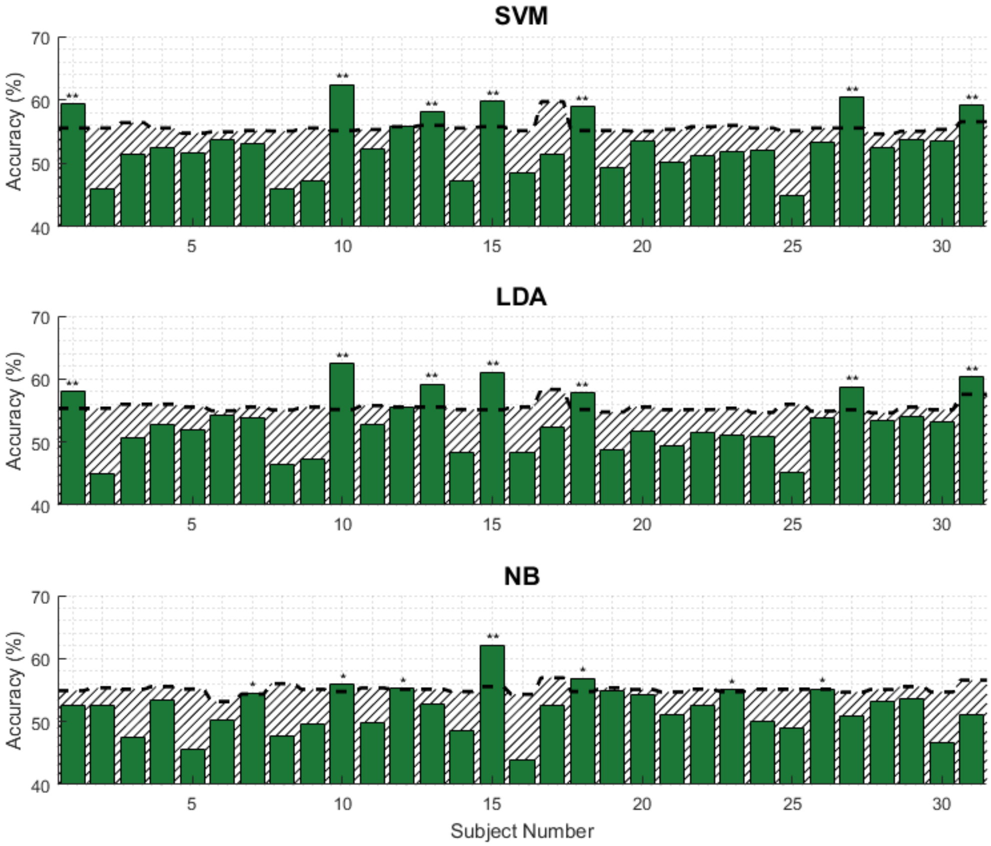


Figure 3.12: Classification accuracy by subject and classifier for features based on the N400 (averaged across clusters of neighboring electrodes). Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance is denoted as: $p < 0.05$ with * and $p < 0.01$ with **.

Classification accuracies varied considerably between participants. Considering features based on the N400, subjects 1, 10, 13, 15, 18, 27, and 31 all had classification accuracies that were better than chance ($p < 0.01$) when using either SVM or LDA. Subjects 10 and 15 even had classification accuracies above 60%. The results for NB were slightly different. Data from subjects 7, 12, 23, and 26 classified at better than chance accuracies ($p < 0.05$) with NB, but not with the other two classifiers.

The classification accuracies when using FP were significantly better than chance for subjects 3, 7, 11, 12, 14, 18, 20, 21, 22, 25, and 28 for at least one

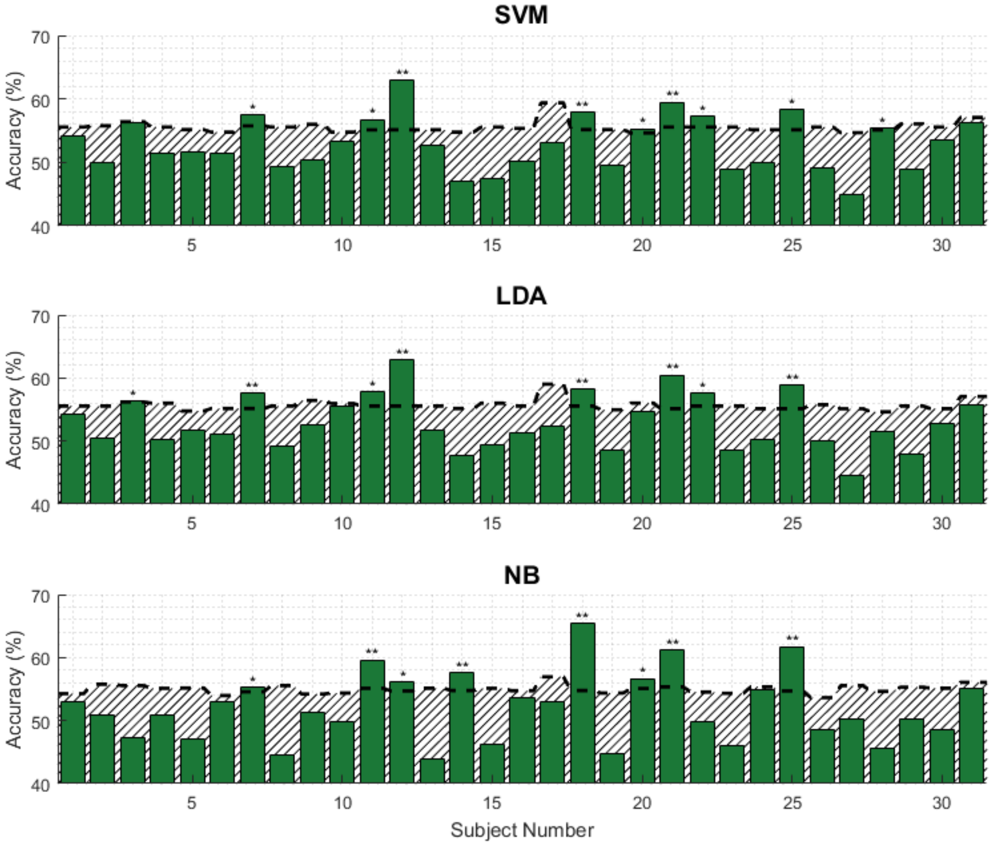


Figure 3.13: Classification accuracy by subject and classifier for features based on FP (averaged across clusters of neighboring electrodes). Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance is denoted as: $p < 0.05$ with * and $p < 0.01$ with **.

of the classifiers.

Since multiple participants had above chance classification accuracies for each of the brain signals, we analyzed the correlation between classification accuracy when using features based on the N400 versus features based on FP. No significant correlations ($p > 0.5$) were found for any of the three classifiers.

The variance of classification accuracy between subjects appears to be larger than the variance between clusters of neighboring electrodes within subjects for features based on both the N400 and FP (visualized in Figures 3.14 [N400] and 3.15 [FP]). When considering the data from the LDA classifier, the mean variance between subjects was 22.82% for features based on the N400 and 19.88% for features based on FP; the within subject variance

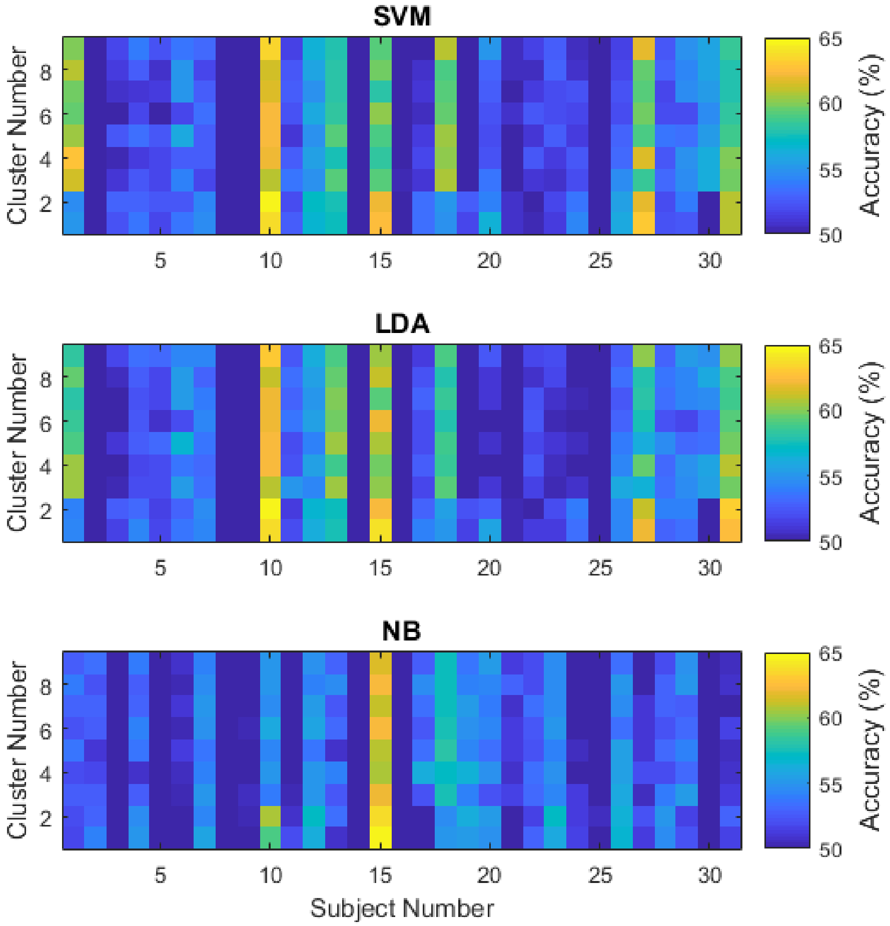


Figure 3.14: Classification accuracy by subject, cluster of neighboring electrodes, and classifier for the features based on the N400. Higher classification accuracies are represented in lighter yellows; lower classification accuracies are represented in darker blues.

was 1.87% for the clusters of neighboring electrodes using features based on the N400 and 1.91% for FP.

All Electrodes. Classification accuracy for all three classifiers (SVM, LDA, and NB) and features based on both brain signals was better than chance when using all 26 EEG electrodes (Figure 3.16). Overall, the average classification accuracy—across subjects and classifiers—of the N400-based features (58.49%) was slightly higher than the average classification accuracy of the FP-based features (57.00%). For the N400-based features, SVM (57.38%) and LDA (59.96%) had similar classification accuracy, while NB had an average classification accuracy of higher than 60% (60.13%). For the features based on FP, SVM (57.67%) and LDA (58.40%) were similar; classification

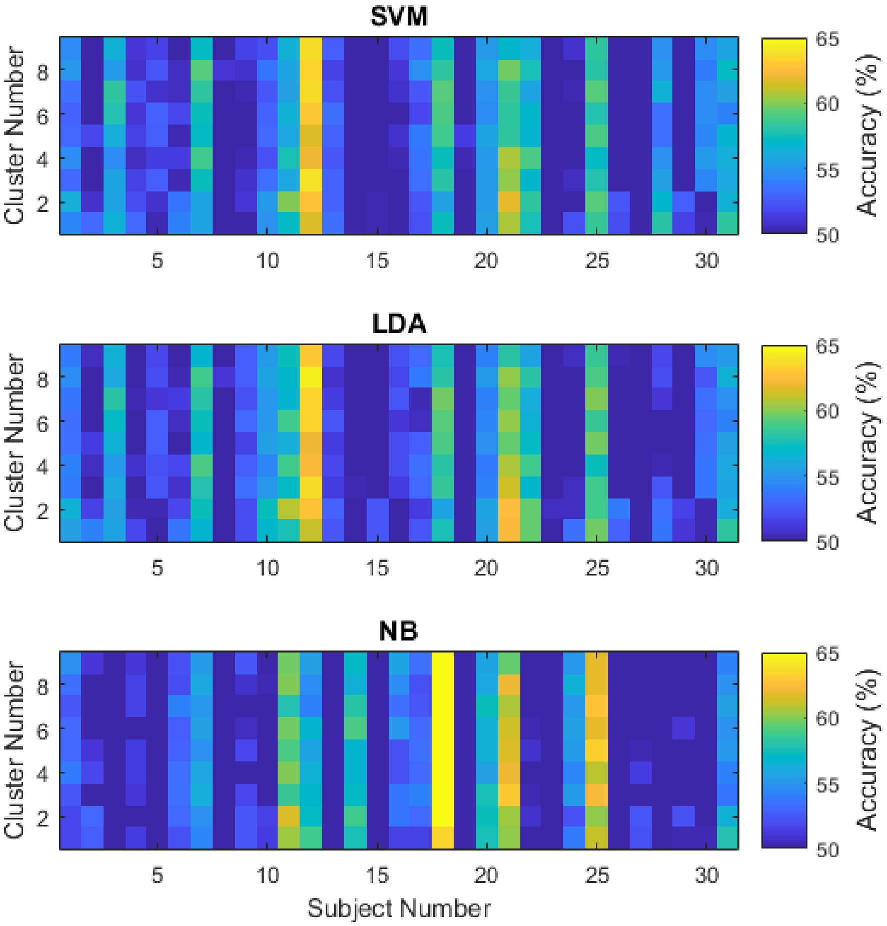


Figure 3.15: Classification accuracy by subject, cluster of neighboring electrodes, and classifier for the features based on FP. Higher classification accuracies are represented in lighter yellows; lower classification accuracies are represented in darker blues.

accuracy using NB was 54.94%.

3.5.2 Different Ways of Analyzing the Individual Trials

Here, we report the results from three different analyses of the individual trials. The classification analyses in this section used features based on both the N400 and FP. When averaging across subjects, classification accuracy was above chance for all three classifiers when using single-trials, the average of two trials before classification, and the average of three trials before classification. Figure 3.17 (left bar in each subplot) shows the average classification accuracy for the single trial analysis (for each classifier). All three

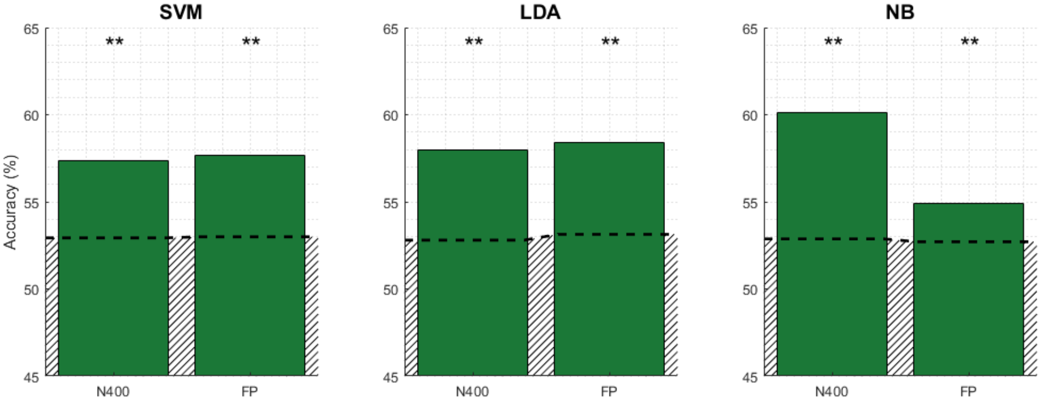


Figure 3.16: Classification accuracy by brain signal and classifier averaged across subjects with a feature from each of the 26 EEG electrodes. Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance is denoted as: $p < 0.05$ with * and $p < 0.01$ with **.

classifiers were able to classify the single trials of EEG at better than chance accuracies using features based on both the N400 and FP. SVM (57.29%) and LDA (57.30%) each had similar classification accuracy. Averaging multiple trials together before classification improved overall accuracy (Figure 3.17 [right two bars in each subplot]). For SVM (60.97%) and LDA (61.30%), average classification accuracy when averaging three trials together before classification was higher than 60%. NB classification accuracy was worse for this combined feature set when using single trials (53.07%), averaging two trials together before classification (54.75%), or averaging three trials together before classification (55.49%).

Initial results of the sequential classification of trials using SPRT resulted in slightly lower classification accuracy (on average) than the fixed trial analyses. Figure 3.17 shows the results of the SPRT classifier (for different values of τ) overlaid on the results from the fixed trial analyses. Figure 3.18 shows the classification accuracy (left) and number of trials per classification attempt (right) as a function of the τ. As τ decreased, SPRT appeared to have higher classification accuracy than Naïve Bayes, but have similar or worse classification accuracy than LDA or SVM. The maximum accuracy for SPRT was 59.23% for a $\tau = 0.04$ which required 4.5 trials per classification attempt.

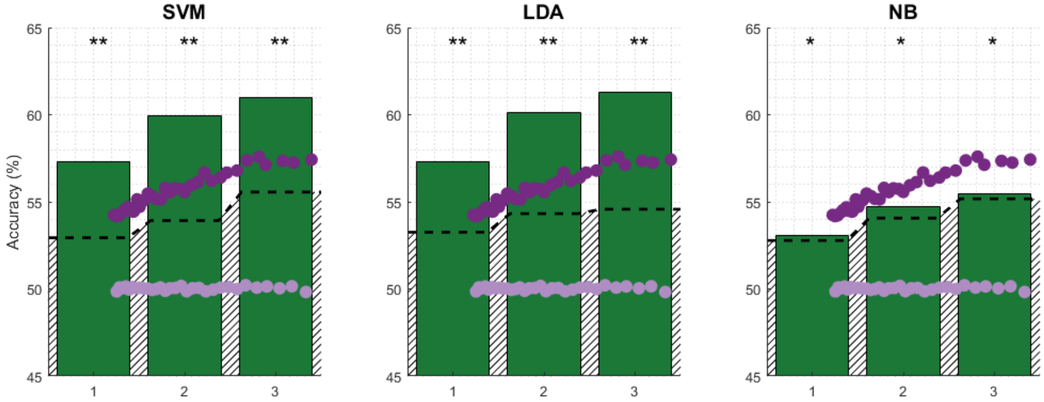


Figure 3.17: Comparison of single trial classification accuracy, classification accuracy when averaging multiple trials together before classification, and the sequential classification of trials using (SPRT). Each dark purple dot represents the classification accuracy of SPRT for a different value of τ. Each light purple dot represents the classification accuracy of SPRT for a different value of τ when the labels were randomized. Chance classification accuracy ($p > 0.05$) is indicated by the shaded area. Classification accuracy that is significantly better than chance is denoted as: $p < 0.05$ with * and $p < 0.01$ with **.

3.6 Discussion

The results of this study show that expected versus unexpected endings to highly constrained sentences can be classified at better than chance accuracy using single trials of EEG. They also show that this is true using features based on either the N400 and/or the FP. Furthermore, an interface for performing classification analyses on ERP datasets in MATLAB and a sequential classifier based on SPRT are presented. In this discussion, we consider our results, ways in which our classification analyses could be improved, and potential directions for future work.

Of all of the single-trial analyses, classification accuracy was highest when using features from all 26 EEG electrodes (for features based on the N400 or FP; Figure 3.16). Despite the limited number of trials, classification accuracy was higher with more features (26, one for each electrode) than with fewer features (1 for the individual electrode analyses or 2–3 for the analyses using the clusters of neighboring electrodes). It is likely that the appropriate electrodes to include for the features based on the N400 or FP is somewhere between the number of electrodes included in the clusters and all of the electrodes. Future work on the classification GUI should include more advanced

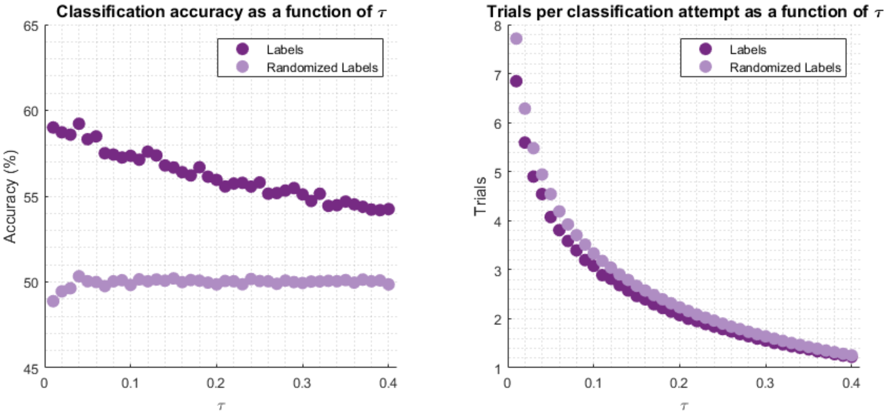
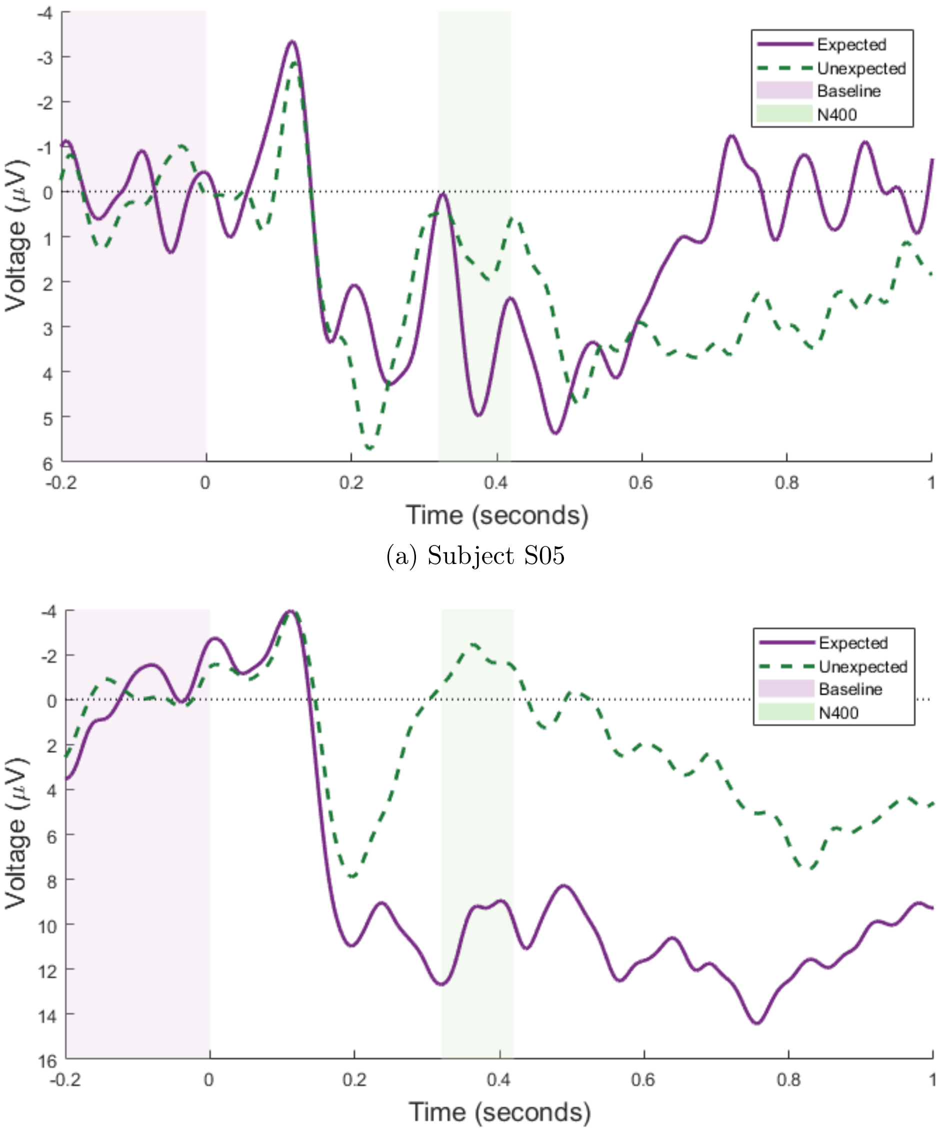


Figure 3.18: (left) Classification accuracy as a function of τ for the SPRT classifier. (right) Number of trials per classification attempt as a function of τ for the SPRT classifier. Classification using the appropriate labels is shown in dark purple and with randomized labels in light purple.

methods for selecting (or otherwise weighting) the electrodes to include in the classifier, such as in [58] or those reviewed in [59].

Classification accuracy varied considerably between the subjects. When averaging across the clusters of neighboring electrodes, the classification accuracy of 11 subjects was above chance for features based on the N400 and FP for at least one of the classifiers. Figures 3.19a and 3.19b provide further evidence of the differences between individuals. Classification accuracy was significantly higher than chance for S18, but not for S05. Likewise, there is a visually apparent N400 for S18 (Figure 3.19b), but there do not appear to be any differences in EEG responses to expected versus unexpected sentences endings 320–420 ms after the onset of the final word for S05 (Figure 3.19a). There are several potential explanations for these individual differences. The simplest explanation is that the classification features (timing, filtering, etc.) were not optimized to the individuals. It is possible that if we made different design choices, there would be less variability across individuals. This explanation, however, does not explain Figures 3.19a and 3.19b. It is also possible that the individual differences are related to the task. Consider, the data used in our analyses were originally collected for a psychological study of the processing of meaning [14]. The task in this study was to sit in a room and read sentences (one word at a time) for an hour. Some of the individual subjects may have engaged well with the task, while others may not have participated in the task at all. Given previous investigations on the role of



(a) Subject S05

(b) Subject S18

Figure 3.19: Average ERP for strongly constrained sentences at *MiPa* demonstrating the individual differences between subjects for (top) S05 and (bottom) S18. The ERP elicited by expected versus unexpected endings are shown in dark purple and dark green respectively. The time window of the baseline is shaded in light purple. The time window of the N400 is shaded in light green.

engagement on classification performance when using a BCI [60], that a task that is more engaging would reduce inter-subject variability. A better explanation of these individual differences would enable researchers to understand who might be able to use a BCI based on the processing of meaning-related information.

For the analyses of the clusters of neighboring electrodes, the between subject differences in classification accuracy appear to be larger than the within subject differences. The variance of classification accuracy between individuals is higher (∼20%) than the variance in classification accuracy between the clusters of neighboring electrodes (∼2%). This is visualized in Figures 3.14 and 3.15; classification accuracy visually appears to be vertically striped. A subject with high classification accuracy for Cluster 1 generally has high classification accuracy for Cluster 2. Whereas, there are large differences in classification accuracy for specific clusters across subjects. It is possible that this is the result of our relatively naïve electrode selection methods. Brain signals (such as the N400 or FP) vary across individuals in terms of timing, spatial distribution, and frequency. Methods that allow individual optimization of the features may improve overall classification accuracy. It may again suggest (see above and Figures 3.19a and 3.19b) that research on the development of a BCI based on the processing of meaning-related information want to improve the method in which the N400 and FP are elicited. A task that elicits these signals more consistently across individuals may be critical to the development of this new BCI paradigm.

High classification accuracy with features based on the N400 was not predictive of high classification accuracy with features based on FP. An analysis did not find a significant correlation between the classification accuracy of features based on these two brain signals across subjects. Only three of the 19 subjects who had better than chance classification accuracy for either feature based on the N400 or FP (for one of the classifiers) had better than chance classification accuracy for both. Therefore, it might be expected that including features based on both brain signals would improve the number of individuals with above average classification accuracy. Our preliminary analysis using a combination of features based on both the N400 and FP did not result in classification accuracies that were better than those based on all EEG electrodes for features based on one of the brain signals. Further research on how the features based on each of these signals could be combined may prove valuable.

Following well-established methods of ERP analysis [15], averaging multiple trials together improved overall classification accuracy. We also compared the results from this averaging procedure with a sequential classification method (SPRT) that allows the classifier to decide to wait for more data if a certain confidence threshold is not met. Comparing the fixed strategy with the sequential strategy, they both had higher classification accuracies when more data was considered. The fixed strategy had a higher classification accuracy (under the tested conditions) for two of the three classifiers (SVM and LDA) than SPRT. Neither the fixed nor the sequential strategy, however, uniformly outperformed the other approach. For our instantiation of SPRT, we assumed that the data from each of the features was normally distributed. It is possible that this assumption was significantly violated by the data. A different choice of model may lead to higher classification accuracies. It is also possible that SPRT is the wrong choice of sequential classification method for this particular kind of classification problem. Further analysis of the current data and refinement of the classifier is necessary to answer these questions.

In this chapter, we described an EEG Classifier GUI for conducting classification analyses. While this GUI made it possible to test multiple classifiers with a single click of the button and to reload specific parameters of analysis for later use, there are many aspects of the system that could be improved. For example, including automated feature selection methods would each enhance the usability of the system. In addition, extending the data visualization interface to the time-frequency based feature analysis may help users to understand their data.

The analyses described in this chapter are preliminary, but there are many design choices (in addition to those covered earlier in this discussion) that could affect the results. These include the following:

- Preprocessing

  Here, we chose to use a linked mastoid reference and a specific bandpass filters. Changing the choice of reference electrode(s) or using a different filter could eliminate noise and improve overall classification accuracy.

- Choice of metric

  Our N400 and FP-based features were computed as the mean voltage within a time window. The mean may be the wrong metric to use. The EEG Classifier GUI also enables the use of other metrics, such as the raw time points or the median of the time window. Any of these metrics may be more predictive of expected versus unexpected endings in highly constrained sentences and improve overall classification accuracy.

- Different features
  In these analyses, we used time domain-based features. Frequency-domain or spatial features, however, may also be useful for classification. For example, BCIs based on changes in the sensory motor rhythm depend on differences in the spatial distribution of signals recorded from the scalp using EEG. Within the EEG Classification GUI, features that are based on spatial information (such as common spatial patterns [CSP] [58]) could be implemented and used to improve classification accuracy.

- Addition brain signals
  We tested the classification accuracy of features based on two brain signals related to the processing of meaning-related information in our analyses. There are other brain signals, however, that could also have been included in our analyses. For example, in a re-analysis of the data from Federmeier et al. [14], Rommers et al. [54] found additional *induced* (non-time locked) changes in EEG activity in response the processing of meaning-related information. Specifically, Rommers et al. [54] used a time-frequency analysis to show a stronger increase in theta power (3–7 Hz) in response to unexpected words in highly constrained sentences. This stronger theta increase was broadly distributed and occurred 300–700 ms after the onset of the unexpected word. In addition to the theta changes, Rommers et al. [54] also noted alpha and beta band changes in the EEG after unexpected endings to strongly constrained sentences. Thus, including features related to changes in theta, alpha, or beta power may improve overall classification accuracy.

All of these enhancements to the current classification system may prove beneficial and represent potential directions for future work.

CHAPTER 4

CONCLUSION

In this thesis, we presented work on the application of sequential hypothesis testing to the classification of electroencephalography (EEG) data for use in brain-computer interfaces (BCIs). In Chapter 2 we demonstrated a sequential strategy for the classification of steady-state visual evoked potentials. Under the conditions that we tested, this sequential strategy performed uniformly better than a fixed strategy. In Chapter 3, we presented work toward the development of a BCI based on the processing of meaningful information. Specifically, we demonstrated that it is possible to separate highly constrained sentences with expected endings from highly constrained sentences with unexpected endings using single trials of EEG data. We also presented the ERP Classification GUI, an interface for performing classification analyses on event-related potential (ERP) data. Finally, we obtained initial results from a sequential classification strategy for ERP data using a sequential probability ratio test. The work in each of these two chapters improves our understanding of the development of classification systems for EEG-based BCIs and may improve the overall performance of EEG-based BCIs.

REFERENCES

[1] J. Vidal, “Toward direct brain-computer communication,” *Annual Review of Biophysics and Bioengineering*, vol. 2, no. 1, pp. 157–180, June 1973.

[2] J. A. Glassman, “An empirical study into the efficient processing of electroencephalographic data,” Ph.D. dissertation, University of California, Los Angeles, 1973.

[3] D. H. G. Schwartzmann, “Black box system identification via the topological dimensionality approach with applications to neurophysiological problems,” Ph.D. dissertation, University of California, Los Angeles, 1972.

[4] L. A. Miner, D. J. McFarland, and J. R. Wolpaw, “Answering questions with an electroencephalogram-based brain-computer interface,” *Archives of Physical Medicine and Rehabilitation*, vol. 79, no. 9, pp. 1029–1033, September 1998.

[5] E. W. Sellers and E. Donchin, “A P300-based brain–computer interface: Initial tests by ALS patients,” *Clinical Neurophysiology*, vol. 117, no. 3, pp. 538–548, March 2006.

[6] T. Vaughan, D. Mcfarland, G. Schalk, W. Sarnacki, D. Krusienski, E. Sellers, and J. Wolpaw, “The Wadsworth BCI research and development program: At home with BCI,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 229–233, June 2006.

[7] P. Olejniczak, “Neurophysiologic basis of EEG,” *Journal of Clinical Neurophysiology*, vol. 23, no. 3, pp. 186–189, June 2006.

[8] J. R. Wolpaw and D. J. McFarland, “Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 51, pp. 17 849–17 854, December 2004.

[9] J. Jin, B. Z. Allison, T. Kaufmann, A. KÃijbler, Y. Zhang, X. Wang, and A. Cichocki, "The changing face of p300 BCIs: A comparison of stimulus changes in a p300 BCI involving faces, emotion, and movement," *PLoS ONE*, vol. 7, no. 11, p. e49688, November 2012.

[10] R. Kuś, A. Duszyk, P. Milanowski, M. Łabęcki, M. Bierzyńska, Z. Radzikowska, M. Michalska, J. Żygierewicz, P. Suffczyński, and P. J. Durka, "On the quantification of SSVEP frequency responses in human EEG in realistic BCI conditions," *PLoS ONE*, vol. 8, no. 10, p. e77536, October 2013.

[11] A. Akce, J. J. S. Norton, and T. Bretl, "An SSVEP-based brain computer interface for text spelling with adaptive queries that maximize information gain rates," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 5, pp. 857–866, September 2015.

[12] H. Awni, J. J. Norton, S. Umunna, K. D. Federmeier, and T. Bretl, "Towards a brain computer interface based on the N2pc event-related potential," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on.* IEEE, November 2013, pp. 1021–1024.

[13] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "A novel 9-class auditory ERP paradigm driving a predictive text entry system," *Frontiers in Neuroscience*, vol. 5, p. 99, 2011.

[14] K. D. Federmeier, E. W. Wlotko, E. D. Ochoa-Dewald, and M. Kutas, "Multiple effects of sentential constraint on word processing," *Brain Research*, vol. 1146, pp. 75–84, May 2007.

[15] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press, 2014.

[16] G. Gratton, M. G. Coles, and E. Donchin, "A new method for off-line removal of ocular artifact," *Electroencephalography and Clinical Neurophysiology*, vol. 55, no. 4, pp. 468–484, April 1983.

[17] T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Extended ICA removes artifacts from electroencephalographic recordings," in *Advances in Neural Information Processing Systems*, 1998, pp. 894–900.

[18] S. Sutton, M. Braren, J. Zubin, and E. R. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, November 1965.

[19] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 6, pp. 1172–1176, June 2007.

[20] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[21] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, “A review of classification algorithms for EEG-based brain–computer interfaces,” *Journal of Neural Engineering*, vol. 4, no. 2, pp. R1–R13, January 2007.

[22] M. Kutas and K. D. Federmeier, “Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (ERP),” *Annual Review of Psychology*, vol. 62, no. 1, pp. 621–647, January 2011.

[23] E. C. Johnson, J. J. S. Norton, D. Jun, T. Bretl, and D. L. Jones, “Sequential selection of window length for improved SSVEP-based BCI classification,” in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.* Institute of Electrical & Electronics Engineers (IEEE), July 2013, pp. 7060–7063.

[24] D. Regan, “Steady-state evoked potentials,” *Journal of the Optical Society of America*, vol. 67, no. 11, pp. 1475–1489, November 1977.

[25] S. T. Morgan, J. C. Hansen, and S. A. Hillyard, “Selective attention to stimulus location modulates the steady-state visual evoked potential,” *Proceedings of the National Academy of Sciences*, vol. 93, no. 10, pp. 4770–4774, May 1996.

[26] I. Volosyak, H. Cecotti, D. Valbuena, and A. Graser, “Evaluation of the Bremen SSVEP based BCI in real world conditions,” in *2009 IEEE International Conference on Rehabilitation Robotics*, Institute of Electrical & Electronics Engineers (IEEE). Institute of Electrical & Electronics Engineers (IEEE), 2009, pp. 322–331.

[27] A. Akce, J. Norton, and T. Bretl, “A brain-machine interface to navigate mobile robots along human-like paths amidst obstacles,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems.* Institute of Electrical & Electronics Engineers (IEEE), October 2012.

[28] E. C. Lalor, S. P. Kelly, C. Finucane, R. Burke, R. Smith, R. B. Reilly, and G. Mcdarby, “Steady-state VEP-based brain-computer interface control in an immersive 3D gaming environment,” *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 3156–3164, 2005.

[29] B. Allison, E. Wolpaw, and J. Wolpaw, “Brain–computer interface systems: Progress and prospects,” *Expert Review of Medical Devices*, vol. 4, no. 4, pp. 463–474, July 2007.

[30] F. Popescu, B. Blankertz, and K. Müller, "Computational challenges for noninvasive brain computer interfaces," *IEEE Intelligent Systems*, 2008.

[31] J. Wolpaw, N. Birbaumer, W. Heetderks, D. McFarland, P. Peckham, G. Schalk, E. Donchin, L. Quatrano, C. Robinson, and T. Vaughan, "Brain-computer interface technology: A review of the first international meeting," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 164–173, June 2000.

[32] G. Schalk, J. Wolpaw, D. McFarland, and G. Pfurtscheller, "EEG-based communication: Presence of an error potential," *Clinical Neurophysiology*, vol. 111, no. 12, pp. 2138–2144, December 2000.

[33] X. Gao, D. Xu, M. Cheng, and S. Gao, "A BCI-based environmental controller for the motion-disabled," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 137–140, June 2003.

[34] O. Friman, I. Volosyak, and A. Gräser, "Multiple channel detection of steady–state visual evoked potentials for brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 4, pp. 742–750, April 2007.

[35] B. C. Levy, *Principles of Signal Detection and Parameter Estimation.* New York, NY: Springer US, 2008.

[36] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, June 1945.

[37] H. Hotelling, "Relations between two sets of variables," *Biometrika*, vol. 28, no. 3, pp. 321–377, December 1936.

[38] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, June 2004.

[39] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, February 2007.

[40] D. H. Brainard, "The psychophysics toolbox," *Spatial Vision*, vol. 10, no. 4, pp. 433–436, January 1997.

[41] M. Kutas and S. Hillyard, "Reading senseless sentences: Brain potentials reflect semantic incongruity," *Science*, vol. 207, no. 4427, pp. 203–205, January 1980.

[42] J. Geuze, M. A. J. van Gerven, J. Farquhar, and P. Desain, “Detecting semantic priming at the single-trial level,” *PLoS ONE*, vol. 8, no. 4, p. e60377, April 2013.

[43] M. van Vliet, C. MÃijhl, B. Reuderink, and M. Poel, “Guessing what’s on your mind: Using the n400 in brain computer interfaces,” in *Brain Informatics.* Springer Berlin Heidelberg, 2010, pp. 180–191.

[44] M. A. Wenzel, M. Bogojeski, and B. Blankertz, “Real-time inference of word relevance from electroencephalogram and eye gaze,” *Journal of Neural Engineering*, vol. 14, no. 5, p. 056007, August 2017.

[45] B. Blankertz, L. Acqualagna, S. Dähne, S. Haufe, M. Schultze-Kraft, I. Sturm, M. Ušćumlic, M. A. Wenzel, G. Curio, and K.-R. MÃijller, “The Berlin brain-computer interface: Progress beyond communication and control,” *Frontiers in Neuroscience*, vol. 10, November 2016.

[46] C. Bishop, *Pattern Recognition and Machine Learning.* Secaucus, NJ: Springer-Verlag New York, Inc., 2006.

[47] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, “FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data,” *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–9, 2011.

[48] A. Delorme and S. Makeig, “EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, March 2004.

[49] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. MÃijller, “Introduction to machine learning for brain imaging,” *NeuroImage*, vol. 56, no. 2, pp. 387–399, May 2011.

[50] L. Cao, K. Chua, W. Chong, H. Lee, and Q. Gu, “A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine,” *Neurocomputing*, vol. 55, no. 1-2, pp. 321–336, September 2003.

[51] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90–100, February 2003.

[52] D. D. Lewis, “Naive Bayes at forty: The independence assumption in information retrieval,” in *Machine Learning: ECML-98.* Springer Berlin Heidelberg, 1998, pp. 4–15.

[53] M. Kutas and K. D. Federmeier, “Electrophysiology reveals semantic memory use in language comprehension,” *Trends in Cognitive Sciences*, vol. 4, no. 12, pp. 463–470, December 2000.

[54] J. Rommers, D. S. Dickson, J. J. S. Norton, E. W. Wlotko, and K. D. Federmeier, “Alpha and theta band dynamics related to sentential constraint and word expectancy,” *Language, Cognition and Neuroscience*, vol. 32, no. 5, pp. 576–589, May 2016.

[55] W. W. Piegorsch and W. J. Padgett, “Sequential probability ratio test,” in *International Encyclopedia of Statistical Science.* Springer Berlin Heidelberg, 2011, pp. 1305–1308.

[56] A. J. Izenman, *Modern Multivariate Statistical Techniques.* Springer New York, 2008.

[57] M. Ojala and G. C. Garriga, “Permutation tests for studying classifier performance,” *Journal of Machine Learning Research*, vol. 11, no. June, pp. 1833–1863, 2010.

[58] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, “Designing optimal spatial filters for single-trial eeg classification in a movement task,” *Clinical Neurophysiology*, vol. 110, no. 5, pp. 787–798, May 1999.

[59] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, August 2007.

[60] J. L. Mullins, “SSVEP-based BCI performance in children,” M.S. thesis, University of Illinois at Urbana-Champaign, 2015.