Automated Machine Learning Framework for EEG/ERP Analysis: Viable Improvement on Traditional Approaches?

AUTOMATED MACHINE LEARNING FRAMEWORK FOR EEG/ERP ANALYSIS: VIABLE IMPROVEMENT ON TRADITIONAL APPROACHES?

BY

ROBER BOSHRA, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF NEUROSCIENCE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2016) McMaster University

(Neuroscience) Hamilton, Ontario, Canada

TITLE:	Automated Machine Learning Framework for EEG/ERP Analysis: Viable Improvement on Traditional Approaches?
AUTHOR:	Rober Boshra B.Sc., (Computer Science) Dalhousie University, Halifax, Canada
SUPERVISOR:	Dr. John F. Connolly and Dr. James P. Reilly
NUMBER OF PAGES:	xxiii, 104

Dedicated to Maro

and the family she left behind

Acknowledgements

First and foremost I would like to thank my supervisors: Dr. John Connolly and Dr. James Reilly. Over these past two years they have given countless hours trying to make a scientist out of the wide-eyed undergraduate they got. John's support and patience has been extremely valuable in adapting to graduate school and the different aspects of academic life. His enthusiasm for science and knowledge made these years a truly fun and rewarding experience. Jim has been a true pillar of support through some deep lows during these two years. He has continuously showed confidence and faith in my abilities at times when I lacked both. For that, and his always welcoming attitude to the silliest of questions, I am very grateful. I am very happy to say that I had (and will have!) the best supervisors ever!

I would also like to deeply thank Dr. Sue Becker for her time and support from the first supervisory committee meeting and to the last days leading to my defence. Her input has been invaluable at keeping me in focus whenever I lost track. Many thanks go to my undergraduate supervisor Dr. Thomas Trappenberg for helping me get this far. His words, support, and advice have been a true guidance.

Thank you to the comrades of the neuroscience department, and to Sandra Murphy for being awesome! Sandra has always made my experience at MiNDS a truly pleasant one.

I would also like to thank all current and previous fellow graduate students in the department of linguistics and languages and members of the language, memory and brain lab for always being great company: Amanda, Richard, Cassandra, Kyle, Connie, Chelsea, Dan, Diane, Edalat, Narcisse, Heather, Jitka, Zoë, Bryor, Angela, Alex, and Karen. It has always been fun being that guy from the neuroscience department. You didn't only make me feel welcome at the department; you made me feel home in Canada. Thank you.

I would also like to thank the folks back at my first home: Peter, Rami, Shedo, Dodo, John, 7ennes, Mon, Charles, Key, Taison, and Marco. They have believed in me and never ceased to be the great friends they are. You always make it worthwhile to go back, guys. Also, I cannot forget to thank Haoxu and Sunny for their awesome friendships. I am deeply grateful to Jitka, Kyle, Roksana, Amanda and Heather. They have truly made the past two years a highlight of my life. I will never forget all the memories we had together; t-rexes, sassafras, trinities, belle, and webster... all of it.

Last but, definitely, not least I would like to thank my father, mother, and sister for their never-ending love and support. My parents have sacrificed everything to get me to where I am today; and for that I am forever grateful. Thank you, there is no way I can ever repay a fraction of what you have given me. Thank you to my amazing fiancée Mira who somehow stuck with me through all the manic moments, shared with me the happy days and got me through the weeks of academic despair.

Abstract

Event Related Potential (ERP) measures derived from the electroencephalogram (EEG) have been widely used in research on language, cognition, and pathology. The high dimensionality (time x channel x condition) of a typical EEG/ERP dataset makes it a time-consuming prospect to properly analyze, explore, and validate knowledge without a particular restricted hypothesis. This study proposes an automated empirical greedy approach to the analysis process to datamine an EEG dataset for the location, robustness, and latency of ERPs, if any, present in a given dataset. We utilize Support Vector Machines (SVM), a well established machine learning model, on top of a preprocessing pipeline that focuses on detecting differences across experimental conditions. A hybrid of monte-carlo bootstrapping, cross-validation, and permutation tests is used to ensure the reproducibility of results. This framework serves to reduce researcher bias, time spent during analysis, and provide statistically sound results that are agnostic to dataset specifications including the ERPs in question. This method has been tested and validated on three different datasets with different ERPs (N100, Mismatch Negativity (MMN), N2b, Phonological Mapping Negativity (PMN), and P300). Results show statistically significant, above-chance level identification of all ERPs in their respective experimental conditions, latency, and location.

Notation and abbreviations

ANN	Artificial Neural Network
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
BCI	Brain Computer Interfacing
DNN	Deep Neural Network
EEG	Electroencephalography
EP	Evoked Potential
ERP	Event Related Potential
ICA	Independent Component Analysis
LORETA	Low Resolution Electromagnetic Tomography
LSTM	Long Short Term Memory (network)
MEG	Magnetoencephalography
MMN	Mismatch Negativity
MRMR	Maximum Relevance Minimum Redundancy

PCA	Principal Component Analysis
PMN	Phonological Mapping Negativity
RBF	Radial Basis Function
SD	Standard Deviation
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
SVM	Support Vector Machine

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cognitive processes such as attention, vigilance, memory, and language are targets of research that aims to uncover the intricacies of the most complex organ present in the human body: the brain. An avenue for tapping into the processes that occur in the brain lies in the use of electrophysiological measurement methods. Such methods, however, come with many issues due to the sensitivity of currently available recording equipment, nature of the biological signals and, consequently, noisiness of data recorded. The intent of this thesis is to formulate a framework that provides a fast, validated, and automated method of analyzing electrophysiological signals that can be used to more efficiently answer research questions in cognitive neuroscience.

This study has five main parts: first, an introduction to the required background and a brief literature review of both traditional views and novel advancements in the field; second, an introduction to the modules, algorithms and paradigms that are to be used; third, a summary of the validation criteria utilized and applied to generate the results of the modules; fourth, the framework that uses the smaller parts to generate useful information out of a generic dataset; fifth, some insight into the applicability of the framework to three

different EEG datasets by showing the process and results.

1.1 Electroencephalography and the Event Related Potential method

Electroencephalography (EEG) is a method that has been used extensively in the field of neuroscience, dating to the 1930s (Luck, 2014). Briefly, EEG provides a means by which brain signals, mostly large postsynaptic potentials, can be measured non-invasively from the scalp. Amplifiers are used to capture the small electric potentials generated by millions of neurons during brain activity. This, in turn, makes the procedure very susceptible to noise that is amplified as well. Different procedures, paradigms and recording characteristics are used in EEG labs to serve different scopes of research, and different hypotheses. However, a few important yet variable features of EEG implementation in research will be explicitly mentioned in this section; namely: referencing, electrode density, and basic preprocessing.

An EEG captures electric potentials through electrodes with metal tips that are set up with conducting gel to minimize impedance with the scalp/skin. While results are usually discussed in terms of electrode locations, recorded signals are the electrical potential difference between a measuring electrode (active) and a reference. The most important aspect of a reference electrode is that it captures the overall noise in a participant's head region, yet ideally does not capture the particular brain response in question; otherwise, its subtraction would remove the response in question from the data. Commonly used references are tip of the nose, average of the mastoids, earlobe(s), and average reference. In practice, both tip of the nose and average of mastoid references are commonly used, along with some

usage of the average reference. However, there have been arguments made against using the average, especially in low density setups (e.g., Desmedt et al. 1990). In the present thesis, only mastoid reference is utilized. An in-depth discussion of different referencing methods can be found in (Dien and Santuzzi, 2005, Desmedt and Tomberg, 1990).

Different numbers of electrodes have been used in different studies based on study-specific criteria. Some of these criteria include setup time, participant population, and planned method usage. For instance, in patient and child populations, a lower number of electrodes is used to shorten setup time and reduce participant irritation. Conversely, methods like Independent Component Analysis (ICA) (Comon, 1994), and low resolution electromagnetic tomography (LORETA) (Pascual-Marqui et al., 1994) require a large number of electrodes in order to function adequately, in terms of providing reliable and valid estimates of component identities and their significance. Typically, more electrodes are used when either multiple uncorrelated, independent, or spatially separated components are to be extracted. While EEG has low spatial resolution, relative to fMRI, some insight on localization can be provided by analysis techniques involving a high density array.

While there is not one defined pipeline that all researchers use for preprocessing, there are several steps which are typically run on raw data. These steps are outlined in the second chapter. Please note that the order of these steps is not necessarily consistent across labs, setups, or experiments. A traditional pipeline of EEG analysis includes analog filtering of data during recording, further digital filtering of recorded data during preprocessing, referencing of the data, artifact rejection, artifact correction, segmentation, and baseline correction.

Event-related potentials (ERPs) are electrophysisiological brain responses elicited to specific types of stimuli (events). These signals are typically hard to observe in continuous

EEG recordings; thus, it is common to utilize averaging windows of preprocessed EEG data (trials or segments) across individual occurrences of a type of stimulus in question, setting the stimulus as a time-locking point. ERPs are known to reflect a wide variety of brain functions including attention, cognition, and memory. Some of the early potentials are emitted or elicited due to direct sensory input (exogenous) and are mostly referred to as evoked potentials (EPs). Others, commonly later in latency, can be observed as a result of cognitive processing of the time-locked event (endogenous). Examples of ERPs and EPs used in the present thesis are discussed below.

1.1.1 N100

The N100 is a negative-going evoked potential (EP) peaking at 80-150 ms following the presentation of a stimulus. While it has usually been studied in the auditory modality, it has also been shown to arise following visual stimuli (Näätänen and Picton, 1987). It was shown that the N100 is exogenously driven in that it varies with different kinds of stimulus manipulations such as frequency (Hz), loudness (dB), and duration (ms) (Näätänen and Picton, 1987). An exogenous EP is described as arising due to a sensory signal propagating from sensory organs and reaching parts of the brain detectable by the EEG. This EP has been shown to be elicited following the presentation of a variety of auditory and/or visual stimuli including tones, speech, and animal sounds. When the transient aspects of the stimuli are controlled, the N100 responses exhibit similar characteristics on repeated presentations. Attention is not a requirement for the elicitation of the N100. On the contrary, the EP can be observed in subjects during periods of inattention, coma, and sleep. While aspects of the N100 waveform differ when a subject is presented with a sequence of varying stimuli, it has been shown not to be the underlying brain response behind pattern

matching and recognition (Näätänen and Picton, 1987, Näätänen et al., 2005).

1.1.2 Mismatch Negativity (MMN) and the N2b

Mismatch negativity is an ERP commonly associated with the brain's auditory processing system. It arises in response to deviation from an established pattern. For instance, several identical tones followed by an identifiably different (deviant) tone will elicit a negative peak 150-250 ms after the onset of the deviant stimulus. The MMN has been shown to arise due to several types of deviant stimuli (Todd et al., 2008, Näätänen, 1992). Even though the temporal aspect of the MMN is similar to that of the N100, studies have shown that the MMN is dissociated from the N100 in its function (Näätänen et al., 2005). The MMN has been argued to be a manifestation of a part of the underlying mechanism for auditory awareness (Näätänen et al., 2005). Auditory awareness should not be confused with attention, however, as the MMN has been shown to arise in an inattentive subject who is not actively identifying the types of target stimuli presented. The lack of a requirement for active participation makes the MMN a viable option for testing the degree of consciousness in individuals in comas, vegetative states, and minimal consciousness states (Morlet and Fischer, 2014). It has also been shown to be a good predictor of coma recovery (Cowan et al., 1993). Other clinical research applications of the MMN have been shown in Todd et al. (2003, 2008), where it is demonstrated that schizophrenic patients emit smaller MMN responses when compared to healthy controls. Such results signify higher auditory discriminatory thresholds in schizophrenic populations.

In paradigms similar to what have been highlighted above, adding a task requiring active attention from participants to the stimuli is known to elicit a negative peak (to deviants) after, and sometimes overlapping with, the MMN (Näätänen et al., 1982); this negativity

has been repeatedly replicated and is labeled the N2b. The ERP has been shown to originate from generator(s) anterior to those of the MMN's and to be closely associated with the appearance of the P3a. Furthermore, attenuation of the MMN has been reported in attention for lower intensity deviants (see Näätänen et al. 1993). This, however, was shown to not be the case in response to lower frequency deviants; for the MMNs detected were equal in amplitude and unaffected by all sort of attention modulation.

1.1.3 Phonological Mapping Negativity (PMN)

The PMN has been shown to be observable after a violation of phonological expectation with auditory stimuli (Connolly and Phillips, 1994). For example, in (1), the contextual information in the sentence gives rise to an expectation that the word *luck* will follow the word *bad*. This expectation can be characterized by a high cloze probability, a quantifiable measure defined as the probability of a target word given a particular sentence (Kutas and Federmeier, 2014). A violation of that expectation elicits an N400 response which is characterized by a negative peak 400 ms post-stimulus onset (Kutas and Federmeier, 2014).

(1) The gambler had a streak of bad luggage.

(2) Don caught the ball with his glove.

While that is the case for semantically violating sentence-final words, as well as low cloze probability, for words that are semantically matched (albeit with smaller amplitudes) a different ERP arises in response to the particular violation of an expected initial phoneme in the word, irrespective of the semantic matching. For example, in (2), the ending *glove* is semantically compatible, however the more highly expected word (*hand*) has a different

initial phoneme. Consequently, this mismatch generates a negative-tending peak that normally occurs 250-300 ms post stimulus (Connolly and Phillips, 1994, Newman and Connolly, 2009). Moreover, it has been shown that the component is elicited independently from the N400 and can be individually manipulated by experimental procedures, as in (2).

1.1.4 P300

The P300 has been intensively studied as a marker for attention and memory. It typically is characterized by a positive peak arising 300 ms after stimulus onset, although in practice it can be delayed (Polich, 2007). The P300 is very robust and, compared to other ERPs, it is the easiest component to detect on a single trial basis (Blankertz et al., 2011). The experimental design most associated with the P300 is the oddball paradigm. The classic version of this experimental design involves two stimuli that differ along some dimension (e.g., duration or intensity) with one stimulus (the standard) occurring more often (e.g., 80% of the time) than the other (the deviant, which occurs the remaining 20% of the time). The participant is instructed to attend to the stimuli being presented and respond (e.g., button press) to the less frequently occurring stimulus. The P300 is seen in response to a deviant stimulus that is identified as such by the participants. However, the P300 has also been shown to arise in response to stimuli during periods of inattentiveness or sleep. The ERP was also observed in cases of consciousness disorders, although it was seen as an indication of covert consciousness (Perrin et al., 2006, 1999). The P300 is not stimulus-specific as it can be elicited (or emitted) in response to auditory, visual, somatosensory, and olfactory stimuli for example. Also, the nature of the stimuli in the auditory and visual modalities can range from simple tones/light flashes to words, pictures, and familiar sounds (e.g., telephone ring). The presence of the P300 in an oddball paradigm is driven more by

frequency of occurrence of the relevant stimulus than its modality.

This ERP can be further classified into two observed components: the P3a and the P3b (Polich, 2007). The P3b is associated with tasks that involve memory processing of some sort, with or without attention. It can arise due to a violation of a certain expectancy, or presentation of a target stimulus. For example, in a paradigm shown in Lefebvre et al. (2005) when an expected sequence of digits is broken, a P3b is elicited. The oddball paradigm is another example of a P3b elicitation (Polich, 2007). The P3a has been associated with observing novel stimuli, such as a dog bark or a car honk, given that the context does not provide an expectation for such stimuli. The P3a has been proposed to arise as the brain is undergoing attention allocation (Polich, 2007). Due to the robustness of the ERP, it has been used in several brain-computer interfacing paradigms in spellers, cursor control, and concealed information tests (Krusienski et al., 2006, Wolpaw and McFarland, 2004, Meijer et al., 2007).

1.2 Conventional ERP Analysis Techniques

Time-locked analysis (e.g., averaging) is historically the default way of processing and presenting ERPs. This technique is used to study differences across conditions and experimental groups, utilizing brain responses that are elicited or emitted after differing stimulus types. Data is epoched (segmented) into trials, where each trial consists of signals recorded by the EEG and time-locked to stimulus presentation. As a convention of ERP studies, a portion of data before stimulus presentation is kept as a form of baseline in contrast to brain responses after stimulus onset. As signal-to-noise ratio in raw EEG signals is very low, trials are normally averaged for each experimental condition. Averaging serves to remove background oscillations, artifacts, and other cognitive responses that are not of interest,

while emphasizing the brain response(s) relevant to a study. Characteristics of observed ERP(s) such as latency, baseline-peak amplitude, and area under the curve are viable candidates for analysis as dependent variables in typical analyses of variance (ANOVA) across different factors pertaining to experimental conditions, population groups, and brain regions of interest (Teplan, 2002, Luck, 2014).

In typical studies, one of the above-mentioned features is used as the continuous dependent variable in an ANOVA in order to account for differences across experimental categories. This is usually done by utilizing a mixed model with between-subject factors (sex, treatment, population group) and within-subject factors (condition, brain region). While this approach is effective and has served to analyze numerous EEG/ERP studies, it has four main limitations. Firstly, an ANOVA provides limited capability in accounting for the time dimension inherent in ERP data. While this is usually solved by specifically using single points (or averages of consecutive points) along the time axis representing a particular ERP, many disadvantages arise such as researcher bias, and loss of information. Secondly, a loss of power arises resulting from the reduction of numerous trials across many subjects to grand averages that represent the general trend of responses across a group. Thirdly, the use of high density EEG recordings efficiently is not trivial when utilizing ANOVAs. Data dimensionality reduction, similar to trial-averaging, is required which in turn reduces topographic information. Lastly, an ANOVA does not provide a clear measure by which to control for continuous independent variables such as: age, and time since an important event prior to EEG recording. This is a clear disadvantage when accounting for continuous independent variables is necessary for testing a certain hypothesis. There exist some solutions to counteract these limitations. These alternatives will be discussed here briefly, although by no means is this a complete account.

When several times of interest are in question, for example when multiple ERPs are expected to arise due to experimental design choices, one option is to extract features from time windows that correspond to each ERP in question. This window location can then be added as an extra factor in the ANOVA. While this method provides a means by which to satisfy analysis needs, the window approach also introduces many parameters that need to be adjusted by the researcher such as window length and features extracted from a designated window. As researcher choices multiply, the likelihood for uncorrected multiple comparisons pertaining to sets of selected parameters increases; although there are well-established correction methods, failure to use them appropriately remains an issue. Another complication arises when comparing ERPs that overlap temporally. A clear windowed definition of different ERPs becomes difficult and would lead to further window boundary manipulations. Although not always possible, there are many examples of counteracting this issue with clever use of experimental design techniques to isolate components. An example can be seen in experiments shown by Näätänen et al. (2005) to disentangle the N100 and the MMN; another is the study by Connolly and Phillips (1994), demonstrating the dissociation of the N400 from the PMN.

The general trend of grand averaging across subjects stands in contrast to more advanced analysis methods that are finer grained for both time and evaluation of performance from the individual subject. Single-trial basis of analysis has recently become more prevalent due to advancements in signal processing, analysis techniques, and machine learning. Two main aspects of this problem need to be highlighted. The first is that, ignoring the fact that having low signal-to-noise ratio (SNR) would inherently increase the likelihood of Type II errors, using single trials directly will artificially inflate the degrees of freedom. In a normal ANOVA setup, this would not be statistically sound as, for example, an increase

in the number of trials in a particular experiment should not be equated to an increase in the number of subjects. The second concern is the loss of information that is entailed by the inherent cross-trial variability within a subject. An effect that increases through time then proceeds to diminish, due to fatigue for example, would affect an entire subject's data. Artificial points can be added to the timeline of the experiment to split trials into categories that can be analyzed separately. However, many issues, not different from ones highlighted above, may also arise with that solution. An in-depth discussion of issues and pitfalls of using ANOVAs on EEG/ERP data pertaining to this point can be found in Dien and Santuzzi (2005).

Similar to the single-trial problem, using a high number of electrodes without forms of simplification can also cause inflation of the degrees of freedom (Dien and Santuzzi, 2005). This is especially wasteful when high density EEG is used in the recording, only to be simplified to averages of electrode signal values in regional blocks. This reduces the usefulness of said EEG setups and results in a bottleneck on information inference during the analysis phase. These issues can be counteracted by more effective techniques.

Using continuous independent variables is sometimes required when dealing with specific research questions such as those of Meulman et al. (2015). In this study, age of language acquisition is tested for effects on the elicitation of syntax-related ERPs. While an analysis of covariance (ANCOVA) can be used to control for these variables, more sophisticated analysis methods are needed to test for different relations between the dependent variable and the continuous independent variables. A prime example of a novel statistical tool capable of this type of analysis is the general additive model (GAM) used by Meulman et al. in the same study.

1.3 Machine Learning Background

1.3.1 Objective

Model derivation, pattern recognition, and learning are all aspects of computation that are generally applied directly by human experts or handcrafted specifically to fulfill specific directives. Machine learning (ML) is one of the names used to describe the study of algorithms that can be executed by computers in order to reach one or more of the aforementioned objectives. The field has gained significant traction in recent years and its methods are currently used in many aspects of our modern society, including smartphones, search engines, voice recognition, and autonomous vehicles.

A main goal of machine learning is to extract knowledge from data. For example, given a dataset of images, a machine learning algorithm will try to create a model that is capable of identifying the class of a certain image. Possible example classes include: portrait, nature landscape, and text. The model created is mainly dependent on the data supplied to the learner and requires relatively little human-expert interaction.

Within machine learning, there are roughly three major divisions that tackle learning differently: supervised, unsupervised, and reinforcement learning. In supervised learning, a large, labeled dataset is provided to the learner. The learner uses generalized learning techniques to produce a model that is capable of predicting what the labels are. While supervised learning is well-suited to classification tasks, it is also possible to use such machines to generate regressor models that are able to predict a value given some input data. For instance, a regressor model could predict the price of a house given features such as neighborhood, how old the building is, and condition of the estate. In unsupervised learning methods, the algorithm is given a set of unlabeled data. The learner is then tasked

with finding patterns, groupings, and other forms of information that describe the data. This process is logically similar to a child grouping types of cars together as a sort of class without being explicitly provided a label each time he/she sees a different car. Lastly, reinforcement learning follows the Psychology construct of the same name in which an algorithm is given an indication of whether or not it is doing well a few steps after it has taken an action. For example, a checkers machine-learned model player has been developed by giving the learner an indication of reward when it won, or punishment, when it lost. The learner is responsible for retracing its steps and recognizing what turns could be improved upon to win in later games. Even though machine learning methods have also been developed to form algorithms that are an amalgamation of two or more of these discrete types, the basic ideas behind them revolve around these three categories. In this thesis, only methods of supervised machine learning have been used. For a more in-depth review of machine learning methods, refer to Michalski et al. (2013).

1.3.2 Common Problems in Machine Learning

The main goal of machine learning is to automatically generate a model that generalizes to new data, which has not been seen during the learning process. That, however, is not normally an easy task since learning algorithms often fall into problems of overfitting, local minima, and poor validation.

Since a learner's task is to use features of a particular dataset to infer underlying structure, it is common for the learner to draw very definite conclusions (to the point of observation memorization) on the given input. Normally, this results in a model that can attain almost perfect classification accuracy on the data it was trained on, while achieving very

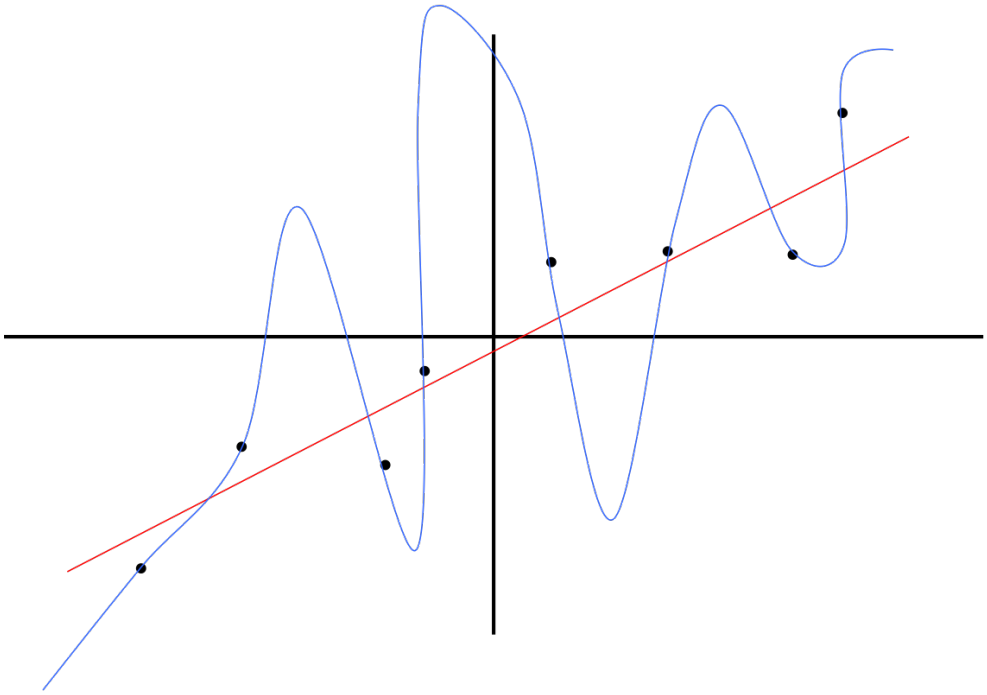

Figure 1.1: An example illustrating the problem of overfitting in machine learning. The blue curve highlights a possible learned model on the datapoints (black dots) with high overfitting, and the red line represents a better model that is likely to generalize better. (Buduma, 2014)

poor results on unseen data. For example, given data y (y-axis) sampled from a linear function in terms of feature x (x-axis) plus some additive noise ε, a likely overfitting model can be seen in figure 1.1. The appropriate model (red) is not achieved due to the ML algorithm trying to attain perfect accuracy (blue) on the training data (black). Instances similar to this example critically reduce machine-learned model accuracy, and are categorized under overfitting. This is the reason why a fundamental aspect of machine learning is to divide the dataset into at least two different sets: training and testing. The resultant accuracies on the testing set are almost always the reported one for a given model. This, however, is a manner by which to simply check generalization and not to actually improve it. Many methods have been adopted to partly solve overfitting problems, some of which are used in this thesis. In-depth discussion of the advanced methods are not within the scope of this thesis, but can be explored further in Browne (2000), Srivastava et al. (2014), Larochelle et al. (2009).

Stemming from the overfitting problem arises the issue of local minima. Many machine learning algorithms are built on top of a minimization optimization problem on the error of the model resulting from an arbitrary point on the parameter hyperplane compared to the truth values held in the labeled data. Concretely, the machine learning algorithm starts with a function (model formed by randomly-set parameters) that simulates how a dependent variable (the target of prediction) varies as the independent variables (features) differ. Initially, that function is incorrect where the difference between the model-predicted value and the true value recorded in the data is called the error. Using optimization algorithms, that error is then minimized iteratively by changing the model parameters. This optimization problem is convex (there is a unique minimum with the lowest error possible) for several algorithms, most prominently in linear Support Vector Machines (SVMs). That, however,

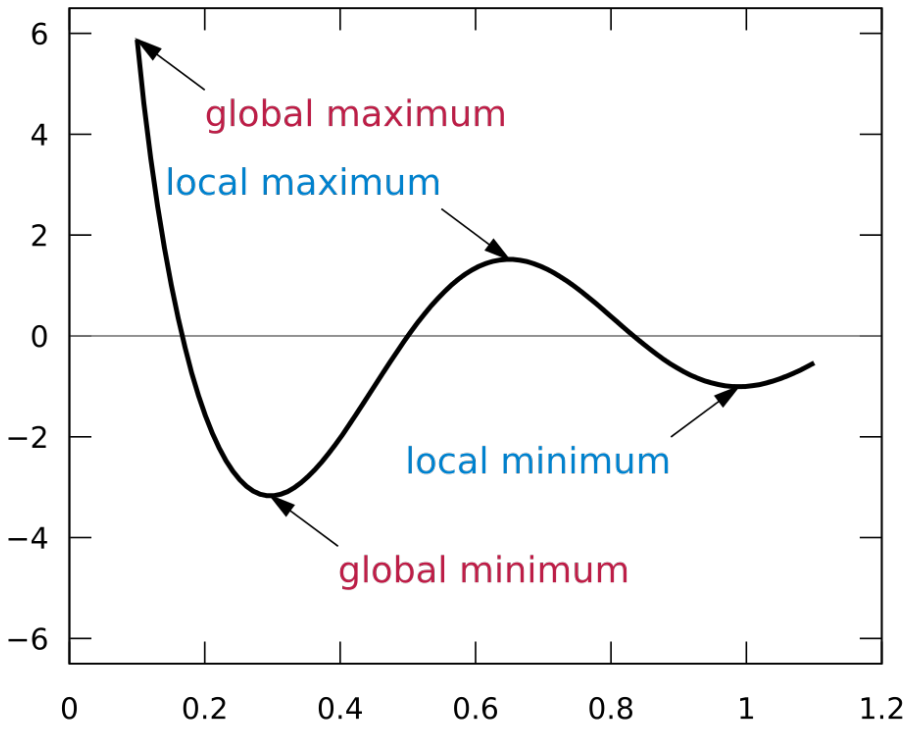

Figure 1.2: An example of local and global minima in optimization problems. (Commons, 2013)

is not always the case with more complex learning methods such as deep neural networks. In those cases, for specific initialization parameters, the learner can get stuck in a minimum that is low in its local position on the manifold but is not the optimal solution globally (see figure 1.2). Often, this is counteracted by running the learner several times with different initialization parameters and choosing the one that generalizes most and provides the best results. There are also several methods that improve or transform the search space to one solvable as a convex-like problem, but the details of those are outside the scope of this thesis. Several examples of these methods can be found in Rasmussen (2006), Hofmann et al. (2008).

Lastly, the issue of poor validation is a substantial one when it comes to generating models that are practical and provide comparable results on unseen data. This issue, while theoretically quite different, is similar in nature to statistical analysis results not being reproducible given identical experiments, data, and researchers. While machine learning fundamentally uses the separation of training and testing data, there are other systematic biases that can be caused by improper use of validation. For instance, given a considerably small dataset, it is possible to run the machine learning algorithm repeatedly with different hyperparameters minimizing the errors for specific training and testing sets. However, this introduces the possibility of generating a model that does not necessarily generalize well to unseen data, but to only this specific test set given the provided training one. This issue is usually counteracted by using K-fold cross validation, which involves splitting the data into K mutually exclusive folds. K-1 folds are used as the training set, with the remaining fold as the testing set. The learner is then run on all possible combinations. The mean resulting accuracy on all folds is then considered to be the generalization accuracy. The issue of multiple runs of modified learners on a dataset can be associated with multiple

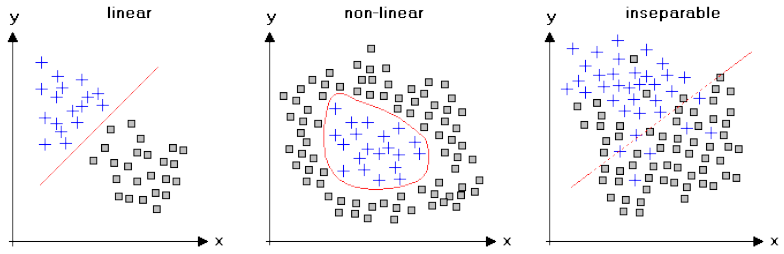

Figure 1.3: Examples of linear separability, non-linear separability, and inseparability (Lohninger, 1999).

comparisons in statistics where the p-value is adjusted to compensate for random chance effects. Other commonly used solutions include: using a cross validation set to optimize hyperparameters, permutations on each run of the learner, Monte Carlo, Bootstrapping, and comparisons to random data. Several of these methods have been implemented within the proposed framework and will be discussed in more detail in the next chapter.

1.3.3 Classifiers

Support Vector Machines (SVMs) have been mathematically defined by Cortes and Vapnik (1995). In classification, they are characterized by separation with a hyperplane that provides the maximal margin between classes. In the linear case, for example, given two linearly separable classes, the logistic regression algorithm would yield a result that divides the two classes. However, logistic regression would not provide a unique answer, since there are infinite possible lines that can split provided data. The SVM improves on that ambiguity by providing the solution that maximally separates the two classes. This, while not affecting the result in some examples, achieves two main points: 1) It provides a better basis for generalization; and, 2) generates a unique convex problem solution. While examples are normally shown for two dimensions, SVMs can easily be applied to feature

spaces of more. In practice, however, data is sometime not linearly separable (see figure 1.3). In these cases, an SVM can be extended using the kernel method to transform a nonlinear problem into a domain where it can be solved linearly. This allows data that are better classified using polynomial or radial hyperplanes to be solved as a convex problem. In cases where the classes are not separable, the SVM has been derived to also yield a unique solution which, although it does not split the data perfectly, aims to counteract outliers and generalize to unseen data.

While SVMs are generally robust to distractor features, as the ratio of features to data points (observations) increases, the models generated are less capable of capturing underlying patterns; an issue that SVMs share with other machine learning methods. There are three main possible solutions to that problem: collection of more data, use of a feature selector, and/or utilization of a human expert to select relevant features. The first solution is possible in cases when data collection is not costly, which is not always possible in research and with specific types of datasets. The second solution is capable of improving performance, however, since a feature selector's accuracy is usually sub-optimal, there is a chance that a useful feature can be seen as a distractor and vice versa. The last solution is often not realistic given the intricate and time-consuming nature of processing some data. For the sake of this thesis project, the second solution was taken when dealing with SVMs in cases where a brute force approach of removing distractor features was not computationally feasible.

There are many alternative algorithms that are widely used in ML classification applications that have not been used in the framework discussed in this thesis. That decision was made primarily due to time-constraints, as more sophisticated ML techniques require considerably more customization to work for specific types of data, especially time-series.

The two main approaches that were put into consideration and briefly tested, although no results are provided here due to insufficient testing, are deep neural networks and Gaussian processes (Deng et al., 2014; Rasmussen, 2006). SVMs were instead used modularly, allowing for later replacement with better-performing counterparts if proven beneficial to the application.

1.3.4 Machine Learning on EEG/ERP

Machine learning's capability of automatically extracting knowledge from data has made it a primary tool for online analysis of EEG data in recent years. The main field affected by, and heavily dependent on, ML is brain-computer interfacing (Krusienski et al., 2006, Wolpaw and McFarland, 2004, Blankertz et al., 2004). Brain-computer interfacing (BCI) revolves around forming a dependable, accurate, and precise link between a computer and a human user, utilizing brain generated signals detectable by a form of brain imaging without depending on muscle activity. As previously mentioned, EEG's high temporal resolution, in addition to its mobility and low price, have made it a prime candidate for use in BCI. However, the field is still hindered by many issues. These include low information transfer rates, low accuracy, and user discomfort. Nonetheless, the field is continuously improving as processing power increases, BCI-specific techniques are refined, and new machine learning methods are developed.

On the other hand, ML has also been considered for use on ERP data as an automated decoder. By definition, running ML on ERP data is usually done post-recording and is less concerned with overcoming transient aspects which are more prevalent in realtime processing. As there is no time constraint, an observation used for learning can be a single trial, a single subject's average, or an averaged subset of trials depending on the hypothesis in

question. For example, single subject approaches have been used to create a model capable of identifying individuals with schizophrenia that have a high chance of responding to Clozapine therapy (Ravan et al., 2015). Additionally, Ramkumar et al. (2013) demonstrated that exogenous responses to visual stimuli can be accurately detected in data captured by magnetoencephalography (MEG). Importantly, the study showed a case where brain responses that match the expected biological behavior were unobservable using traditional analysis approaches, but were detected using ML.

Single-subject classification of an MMN dataset utilizing SVMs was performed by Parvar et al. (2015). Results showed high discrimination between the two conditions, averaged for each subject. These tools were utilized in the classification of several other ERPs, showing high classfication accuracies for all when averaging across single-subjects (Sculthorpe-Petley et al., 2015).

One of the most common forms of BCI utilizes a user's P300 response to a stimulus representing the user's choice; other stimuli representing unwanted choices don't elicit a P300. The accuracy of detecting such ERPs with as few repetitions as possible has been a prime research direction in BCI that is concerned with single-trial classification (Blankertz et al., 2011). These methods, however, do not normally apply directly to other ERPs, since the P300 is very robust in comparison. On the other hand, Dou et al. (2007) have proposed a framework to datamine generic ERP datasets in order to achieve goals similar to the ones highlighted in this thesis. The main differences lie in the amount of data expected and needed by the two frameworks; this thesis is concerned with single-study data, in contrast with Dou's motivation to compile multiple-study data to generate rules that correspond to single ERPs.

1.4 Summary of Research and Question

Due to the variable nature of EEG/ERP trials, it is a nontrivial problem to classify single trials with considerable accuracy. While in some EEG research an ERP is robust enough to be detected on an almost trial-by-trial basis, exploring other ERPs remains stuck in the domain of averaging and statistically analyzing differences between grand averages. The main focus of this thesis is to develop a framework of analysis and argue for its ability to report when an ERP is elicited, where it originates in terms of electrode locations, and how consistent it is across trials. There are advantages to this approach; firstly, when validated correctly, machine learning is capable of forming models that are independent of researcher bias, since they are validated on unseen data. Secondly, while a hypothesis with proper experimental design remains a requirement, this approach would also be well-suited to exploratory ERP research. For example, given a dataset with known experimental condition differences but unknown ERP responses, this framework would be able to analyze how conditions differ in terms of brain responses empirically. Lastly, machine learning is mainly an automation utility. This lends itself to accelerating the process of analysis where numerous researcher hours can be delegated to a predefined analysis procedure that mainly uses computer time.

It is important to note that while this method aims to facilitate research in the field, good experimental design is key to resolving research questions. While utilizing the methods presented here can uncover differences in cognitive responses, the implication of such findings can only be useful when applied in conjunction with well-designed experiments and properly controlled variables. A major drawback of using machine learning methods, however, is the need for large datasets. Single subject case studies with low numbers of trials per condition, aside from very robust ERP experiments, are unlikely to produce any

significant results. On the other hand, it can be argued that small datasets would be hard to deal with even using conventional methods.

The hypothesis of this thesis is that the proposed machine learning based approach is capable of automatically highlighting differences across conditions. Such differences signify underlying ERP components in a generalized and statistically validated manner that rivals, in some cases surpasses, conventional methods. Application to three different datasets, collected as parts of different independent studies, has been shown to produce robust results that conform to both the literature of the studied datasets' paradigms and the results achieved using traditional ERP analysis.

Chapter 2

Methods

In the present chapter, datasets to be used as a demonstration of the capability of the framework are discussed followed by a detailed explanation of the the framework's structure. Section order reflects the sequence of analysis starting at acquiring the dataset, and ending with the validation of the results produced by the sub-modules contained and run by the framework.

2.1 Datasets

Three datasets are used to test and demonstrate the capabilities of the proposed analysis framework; they are referred to as follows: coarticulation violation, mismatch negativity, and attentional P300. The first dataset was collected as part of an original study discussed in the dissertation by Kramer (2014). The other two were part of a post-concussion effects study in adolescents, in addition to healthy controls (unpublished data).

All datasets employed the same electrophysiological recording methods. Continuous EEG was recorded using the Biosemi ActiveTwo 64 Ag/AgCl electrode system (see figure

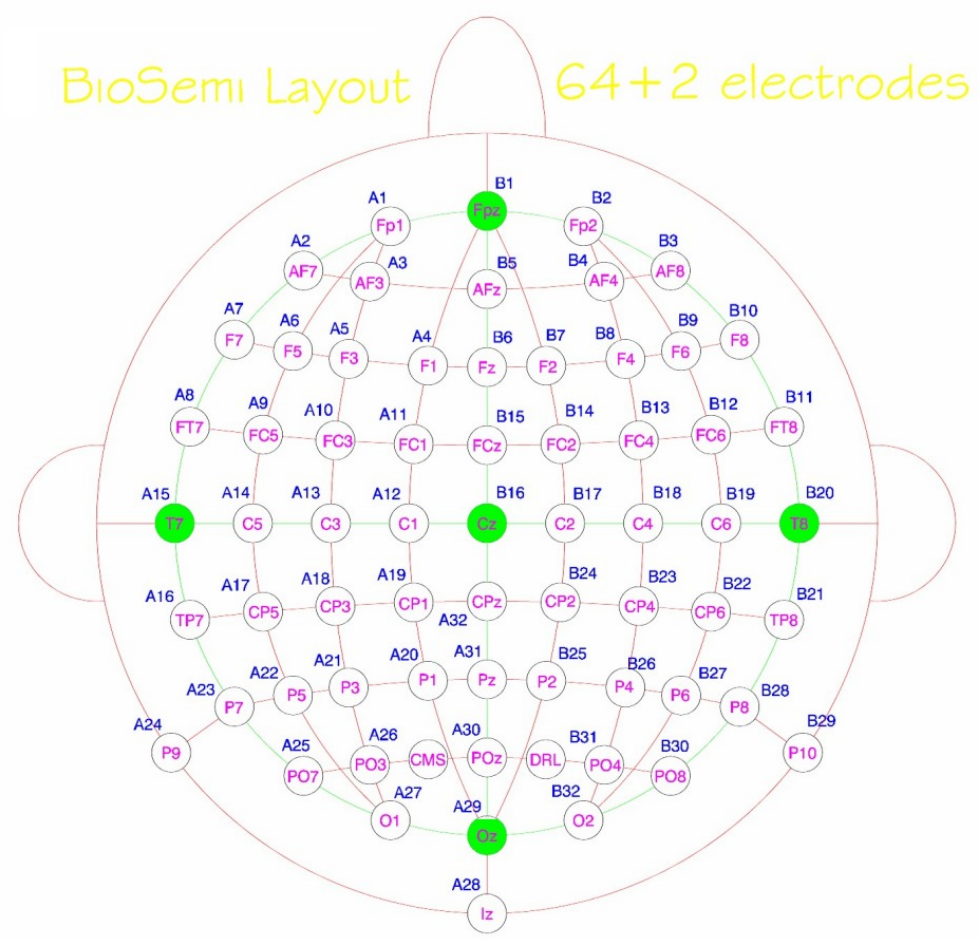

Figure 2.4: The standard 64-electrode Biosemi layout used by all three datasets discussed in the present thesis.

2.4). Electrodes were arranged on the head using a cap allowing for accurate placement according to the 10-20 labeling system. A sampling rate of 512 Hz was used during acquisition with hardware bandpass filtering between 0.01 Hz and 100 Hz. Online recordings were referenced to the driven ground electrode circuit and rereferenced offline to the average of the mastoids. Electrooculographic (EOG) activity was recorded from electrodes placed above and over the outer canthus of the left eye. Data was collected and stored to be analyzed offline.

2.1.1 Coarticulation violation

This experiment was conducted to study the effect of coarticulatory violations on cognitive processing in the brain using EEG/ERP techniques. Coarticulation is characterized by a single phoneme having different sounds depending on neighbouring speech sounds. For example, the sound /b/ in *bat* would have a different sound than the one present in *beet* due to the different vowels following the consonant. Coarticulatory cues have been shown to contain information that facilitate word recognition (McQueen et al., 1999, Gow and McMurray, 2007, Archibald and Joanisse, 2011). This study aimed to analyze the brain response to contextually primed words having various types of coarticulatory anomalies (see below). The hypothesis was that a violation would elicit an early cognitive response manifesting as a negative peak between 250-350 ms after stimulus onset corresponding to the PMN. Three ERP components were analyzed as part of the original study: the N100, P300, and PMN. Further details of this experiment can be found in the original study by Kramer (2014).

2.1.1.1 Participants

Data were collected from twenty-two native English speakers (14 female) participating through the linguistics department course credit system. No history of neurological, auditory, or visual problem was reported by any of the participants. The McMaster University linguistics undergraduate research pool was used to recruit participants for data collection. This research was approved by the local research ethics board and all participants provided informed consent.

2.1.1.2 Stimuli and experimental conditions

The stimuli consisted of 76 different 3 letter monosyllabic English words where the middle letter was a vowel surrounded by two consonants. Each of the words had one of the four corner vowels of English: /i, u, æ, ɑ/. A vowel was given an onset with an anterior stop /p, t, b, d, m, n/ and an oral stop /p, t, k, b, d, g/ with the restriction that the combination forms an English word.

Stimuli were recorded from five female native speakers of English where a participant heard a randomized 20% subset of the stimuli produced by each speaker. Praat sound software (Boersma, 2002) was used to splice onsets from one word to another in custom pairs. A trial started with the presentation of a written word, followed by an audio recording according to the trial's condition, and ending with the participant's response to whether the two matched. There were three experimental conditions in total: congruent, incongruent, and unrelated. In the congruent condition, an audio recording of the word presented visually was spliced at the consonant-vowel juncture with another production of the same word. The incongruent condition was the auditory presentation of the word presented visually, but spliced with another word that had an identical consonant structure but different vowel.

Lastly, in the unrelated condition, the spoken word was spliced as in the congruent condition but was presented after a lexically-incongruent written word. During analysis, only the congruent and incongruent trials were compared; the unrelated condition was only included as an obvious mismatch, and was not within the scope of the coarticulation-specific analysis.

2.1.1.3 Procedure

Testing took place at the Language, Memory and Brain Lab at McMaster University. The procedure lasted approximately 2 hours including setup time. Auditory stimuli were delivered binaurally using earphones (Etymotic Research) and an amplifier (ARTcessories HeadAmp4). Both visual and auditory stimuli were presented using Presentation software (NeuroBehavoiuralSystems Presentation 14.7).

A testing session session consisted of two runs through the experiment file each including a total of 315 presented tokens (630 total) with 184 congruent, 396 incongruent, and 56 unrelated stimulus tokens. Each trial was initiated with a fixation cross for 1250 ms followed by the presentation of a written word on the screen. Following the written word priming, the spoken word was played to the participant. Participants were asked to respond to the spoken word by indicating whether or not the previously shown word was identical to the word they had just heard by using two mouse buttons (left click for same word, right click for different word).

2.1.2 Mismatch negativity

This dataset was captured as a part of a post-concussion effects study that aimed to explore the ERP differences between a healthy participant's brain signals and those of a person

that has recently experienced a minor traumatic brain injury (mTBI)/concussion. The data that has been utilized in this thesis corresponds to the 20 healthy controls of that study. Data collected in this study has been analyzed using conventional methods to assess the differences between the two populations (unpublished data at the date of writing).

The three different deviants were assessed individually in comparison with the standard tones. The ERPs expected to arise as a result of stimuli presented in this study are the N100 (all tones), and the MMN (deviants only) which correspond to a negative peak at the 100 ms point, and a negative peak at the 200 ms post stimulus onset points respectively.

2.1.2.1 Participants

Data were collected from twenty undergraduate participants (16 female) for course credit. All participants were native English speakers with no history of neurological, auditory, or visual problems. Recruitment was done through the McMaster University linguistics undergraduate research pool. This research was approved by the local research ethics board and all participants provided informed consent.

2.1.2.2 Stimuli and experimental conditions

Stimuli used in the acquisition of this dataset were a replication of the experimental paradigm by Todd et al. (2008). The paradigm elicited the MMN in response to 3 different deviants: frequency (1200 Hz), duration (125 ms), and intensity (90 dB sound pressure level [SPL]). Standard tones were 50 ms long at 1000 Hz and 80 dB SPL. The deviants form 18% of the stimuli presented with the remaining being standard tones. The different deviants are equally mixed (6% each) and randomized for each session. In total, there are 1968 standard tone trials, and 144 trials of each deviant.

Triggers are placed at the initial point of each presented tone marking its type: standard, frequency deviant, duration deviant, or intensity deviant. The interstimulus interval (ISI) was 500 ms. All stimuli were presented using Presentation software (NeuroBehavoiural-Systems Presentation 14.7).

2.1.2.3 Procedure

Stimuli were presented binaurally, using earphones (Etymotic Research) and an amplifier (ARTcessories HeadAmp4), with no response required from the participants. A video was played with no sound during the experiment as a distractor and participants were asked to not pay attention to auditory cues. The experiment lasted 23 minutes in addition to setup time; no break was explicitly included in the paradigm.

Although measures were taken to eliminate participants' attention to presented tones during the experiment, it was possible that a participant might still focus on the tones. Thus, four participants' averages had positivities around the 300 ms regions signifying elicitation of P300s instead of the expected MMNs. No measures were taken during preprocessing to filter those participants out to conform to the consistency criterion of the method in cases where some confounding variables cannot be avoided. It is important to note that the number of trials for a given participant that show P300-like components can not be determined; however, if we make a conservative assumption that a quarter of the deviant trials were incorrectly labeled as generating an MMN, then:

$$Percentage\,of\,incorrect\,deviants = \frac{0.25 \times Participants\,with\,P300}{Partcipants} \times 100 \approx 5\%\,of\,deviants$$

Note that a ratio assumption above a quarter (of deviants eliciting P300s per subject) is

less conservative; for it will give the proposed method a larger margin of explainable error. Moreover, a lower ratio was postulated to produce subject averages with less pronounced positivities; however, it can still be argued that the larger, more robust P300s can affect the averages with fewer elicitations. Hence, results reported in the present thesis are not modified for incorrectly-labeled MMN deviants.

2.1.3 Attentional P300

This dataset was also part of the concussion study discussed in the last subsection. Only healthy participants' data were used in the analysis presented in this thesis as ERPs from participants with histories of concussion could be affected in terms of latency variability from trial to trial, elicitation latency , and amplitude. ERPs expected in this study are: the N100, and the P300. The N100s follow the justification presented for the previous dataset. The P300 should be elicited due to participants being instructed to attend to the stimuli, in order to actively complete the task, that a rare deviant tone is different from the more frequent standard tones. The P300 was expected to manifest as a positive peak at 300 ms post-deviant-stimulus onset. Moreover, the addition of attention does not directly inhibit the elicitation of the MMN; however, attention was expected to modulate the ERP. Attention was also expected to elicit the negative component called the N2b.

2.1.3.1 Participants

Data were collected from twenty undergraduate participants (16 female) for course credit. All participants were native English speakers with no history of neurological, auditory, or visual problems. Recruitment was done through the McMaster University linguistics undergraduate research pool. This research was approved by the local research ethics board

and all participants provided informed consent.

2.1.3.2 Stimuli and experimental conditions

The MMN paradigm from Todd et al. (2008) was adapted to elicit the P300 by adding a task of identifying standard tones from deviant tones. The number of stimuli was reduced by a factor of four (600 total) resulting in 492(36) trials of standard tones (each deviant). Triggers were kept for the tones, with ones added for participant responses. A response trigger indicated when a response took place in addition to whether it was one of the following: true standard, true deviant, false standard, and false deviant; true indicated a participant's correct identification of the stimulus type. To accommodate for response time delays, the original paradigm's ISI was extended to 1000 ms. All stimuli were presented using Presentation software (NeuroBehavoiuralSystems Presentation 14.7).

2.1.3.3 Procedure

Participants were seated 1m away from a monitor in the Language, Brain, and Memory Lab. Stimuli were presented binaurally using earphones (Etymotic Research) and an amplifier (ARTcessories HeadAmp4). A fixation cross at the center of the monitor was present at all times on which the participants were asked to fixate through the duration of the experiment. Participants were instructed to identify each tone heard as a standard (commonly occurring) or deviant (one of the three less commonly occurring) by pressing one button for the standard and another for the deviant. Participants were not instructed to memorize any of the deviants as specific targets.

A break was presented halfway through the experiment where the participant was instructed to switch fingers (counterbalancing within participant). The initial button setup

was also counterbalanced across participants. The procedure lasted 10 minutes in addition to the break and setup time.

2.2 Offline Preprocessing

Data was passed through a finite impulse response bandpass filter (order 50) with cutoff frequencies at 0.5 - 15 Hz and using a Hanning window. Using the MATLAB function *filtfilt*, the filter had a zero-phase effect enabling ERPs to be traced in time without the need for time-shifting. This had the effect of doubling the filter order.

The used frequency boundaries were chosen based on apriori knowledge from the literature indicating that most known ERPs lie in that range. More specific filters that constrain the data further to highlight specific responses could be used in adapting the algorithm to individual ERPs, however that is outside the scope of this thesis.

Continuous data was viewed using Fieldtrip's semi-automated artifact rejection software where a trial was rejected if its statistical features exceeded a defined threshold. Features included: variance (< 4000), kurtosis (< 10), and Z values (< 15). Trials containing values that were outliers (observed visually) within the dataset were selected for removal. Since this approach is very subjective to the user, a very conservative approach was taken when removing artifactual trials yielding slightly noisier data but being less susceptible to researcher bias.

Independent Component Analysis (ICA) was run on individual participant data to estimate components that corresponded to ocular artifacts. The InfoICA algorithm from the Fieldtrip package, which in turn utilizes the one implemented by EEGLAB, was used to generate and display components and their topographical maps. The algorithm was run on segmented trials corresponding to all analyzed experimental conditions in a given dataset.

Components that showed fronto-central mappings and had high correlations with the vertical/horizontal EOG electrodes were removed. Electrode data was recalculated using the remaining components to generate artifact-corrected signals. The two components most commonly removed corresponded to blinks and horizontal eye movements, although there were instances when only blinks were correctly detected by the blind source separation algorithm. That can be attributed to some participants closely following instructions to maintain their gaze at the fixation cross during stimulus presentation.

DC detrending was considered to correct for amplifier drift. However, the final framework did not utilize any form of detrending as the effects were negligible; an effect partly attributable to the cutoff for the highpass filter (0.5 Hz).

Two different referencing schemes were considered for data analyzed in the present thesis: mastoids and tip-of-nose. The differences between grand averages for the two methods were analyzed. Preliminary observations showed no significant difference between methods, as grand averages were almost identical. The choice was made to use the average of the mastoids as the referencing scheme for this study, partly due to some participants showing noisy readings from the tip-of-nose electrode.

Using appropriate markers from each of the experiments, epochs are extracted and cataloged according to their conditions. An epoch was 1200 ms long extending from 200 ms before stimulus onset to 1000 ms after. All epochs from an individual subject were baseline corrected by subtracting the average of all prestimulus portions from the individual epochs across all conditions.

After all preprocessing steps were complete, the data were downsampled. Due to the large number of analysis steps and repetitions in the framework, data were decimated to 128 Hz to reduce processing power needed for analysis.

All individual trials collected from all subjects were aggregated and treated as one since single-subject analysis was not within the scope of this study. After a dataset was passed through this pipeline, the output was M tensors of T x C x S dimensions where M was the number of conditions, T was trials, C was channel, and S was sample. In this study, C was 64 and S was (128 x 1.2). Data were saved to disk for later analysis in a format that is readable by MATLAB (.mat).

2.3 Framework outputs

The framework discussed in the present thesis serves to uncover three aspects of an ERP component present in a generic EEG dataset: latency, topography, and strength/consistency. These criteria were mined from data collected from all participants in a particular study. This analysis method does not consider inter-subject variability and assumes that a brain response to a stimulus-type will be in the same topographical and temporal range for all subjects. Consequently, direct application of this analysis is not appropriate across different experimental groups. Further discussion of how this can be extended to between-group analysis is presented in a later chapter.

2.3.1 Latency

One of the main criteria for identifying ERPs lies in their latencies after stimulus presentation. This motivates the first goal of the framework: to locate any signal disparity between two conditions that is prominent enough to shift the classification accuracy away from chance level at a given latency. This signal disparity is attained by constraining features fed into the machine learning algorithm to data samples within a window in time.

A sliding window approach is taken for choosing features in the time domain. A window needs to be strict enough so that there is no major smearing of accuracies across time; moreover, it also needs to be sufficiently large to account for jitter (different response latencies across different trials) commonly observed in ERPs. Several window sizes were tested including ones of 6, 11, and 21 samples (a sample corresponds to around 8 ms where data is downsampled to 128 Hz prior to processing). This was set to be 6 samples as an arbitrary length where all results are reported on windows of that size. A window is taken starting at the first sample corresponding to -200 ms before stimulus onset and then moved 1 sample for each iteration totaling in 149 overlapping windows. A window of features corresponding to a given class and trial was a matrix of size C x S where C is the number of channels (64) and S is the number of samples in a single window (6). These data are passed into the pipeline starting with feature extraction and ending with model validation before moving on to the next window in sequence. An in-depth description of the pipeline and its procedures is in section 2.4.

Results produced by this step are visualized by plotting cross validated accuracies through time (corresponding to shifting windows) with error bars representing the 95% confidence intervals. A visual indication of where an ERP is detected can be seen where the confidence intervals of post and pre-stimulus don't overlap; brain responses before stimulus onset are supposedly unrelated to the experiment. Retaining pre-stimulus data also serves as a convenient validity check, because if the accuracy prior to stimulus presentation is above chance level, it means the data have been handled in an erroneous way earlier in the pipeline.

It should also be noted that the concept of a growing window could be analyzed where a window's start point is the first sample and with each iteration the length of the window

is incremented by one sample. This would serve to show at which point in time a distinguishable signal starts to appear between conditions. This has been briefly tested, but it proved to have high variance due to the learning algorithm's inability to deal with large numbers of features that, in many cases, severely outnumber the observations. Growing windows are susceptible to type II errors, especially when dealing with late latency ERPs, as higher accuracies given by characteristic signal differences would be counteracted by a decrease in accuracy caused by the increasing number of non-contributing features earlier in the window span.

2.3.2 Topography

There are two likely directions by which assessing the topographical distribution of an ERP can be tackled. The first lies in applying the same approach from the previous subsection where iterations are run on whole trials but are limited to data from a single (or smaller subset of) electrode(s); a classification between the two conditions would yield above-chance accuracies for electrodes containing a characteristic differentiating signal. While more intuitive and beneficially independent, the large increase in features, from the expanded time-dimension to a full trial, yielded low chance-level accuracies for most conditions. Thus, a second approach was taken where the latency information learned about the detected signals, from the previous subsection, is incorporated into feature selection. The finalized module iterated through all electrodes, selecting samples that are only within the best time window(s) selected in the time-variant process. A vector of features corresponding to 1 trial within a given condition was of length six before being passed into further feature extraction submodules. This component's results can also be easily assessed visually using topographical plots as shown in in the next chapter where above chance level

accuracies are attainable in signals recorded from defined regions on the scalp.

The main drawback to this approach is that only individual electrodes are assessed. This granular nature of the features does not permit topographical combinations of more than one electrode's data. In order to maintain a fast and relatively simple framework, this has been a trade-off that is acknowledged. Possible expansions on the framework that allow both brute-force combinatorial, and greedy feature mixing are briefly discussed in the future directions section.

2.3.3 Consistency

Due to the variable nature of EEG/ERP trials, it is a non-trivial problem to classify single trials with considerable accuracy. Since the main focus of this thesis is to highlight where and when an ERP arises between differing conditions, a hybrid approach was taken to extract relevant information in an automated manner. The intuition is that for a certain number of individual ERP-containing trials of the same experimental condition, a difference can be detected when comparing to baseline trials. ERPs that are consistent between both types of trials will not be detected; hence, only highlighting targeted responses of a paradigm. This is the same approach taken when using conventional analysis methods, provided that ERP windows are specified for the comparison. In contrast to using grand averages, however, a systematic averaging process is taken before each iteration of the classifier generation/validation phase. The framework starts by assessing accuracies on a dataset of single trials. The data is then processed so that one observation is an average of four random trials within the same condition. That process is repeated for averages of 8, 16, and 32 trials. The resulting accuracies can be interpreted as how consistent an ERP is. An ERP that only shows above chance level classification accuracies for 16 averaged

trials and above would indicate that the ERP is small or is not elicited in the same manner every time a corresponding stimulus is presented. For example, the P300 is known to have a strong signal that is consistent across trials which in turn would give high accuracies with observations formed by averages of less than 8 trials.

The use of consistency in the proposed pipeline is analogous to effect size in traditional approaches. For instance, a small difference between two conditions is less likely to be accurately detected given low SNR data. As larger numbers of trials are averaged, SNR increases and smaller effects are prominent enough to be detected by the learning algorithm.

In terms of module sequence, consistency is analyzed while processing both earlier measures; averages of trials are assessed once during latency analysis, then again within single-electrode topography extraction. For most plots and figures presented in this thesis, an average of 8 trials per observation is the assumed default visualization.

2.4 Machine Learning Methods

Following preprocessing by the module outlined in section 2.2, a given dataset was passed to the machine learning analysis module. The expected input to the module was an EEG/ERP dataset split into two tensors, corresponding to two experimental conditions (N), each with T x C x S dimensions where N was the number of conditions, T was the number of trials, C was the number of channels, and S was the number of samples. In cases when the two tensors had an unequal number of trials, sampling without replacement was done from the larger tensor to attain the same number of trials (as the smaller one). The following sequence of submodules were run twice consequently to: 1) extract latency information from the dataset; and 2) use information from (1) to extract topographical information. For details on the intuition behind the two runs, refer back to section 2.2. During a latency data

mining run, the pipeline was run on all sliding windows (in sequence) spanning 6 samples from 200 ms prior to stimulus onset and until 1000 ms after. On the other hand, a topography mining run looped through all individual electrodes constrained by a 6 sample window of interest. The underlying algorithm for the machine learning module can be found in pseudo-code form in 2.1.

The first step in the analysis module began with aggregating trials into the observations to be analyzed by the rest of the pipelines. Aside from the single-trial run, which just outputs the data as they were, the submodule divided the number of trials into X observations where X = floor($\frac{T}{n}$) , T was the number of randomly sampled trials from one condition, and $n\varepsilon\{2,4,8,16\}$ was the number of aggregated trials.

Procedure 2.1 Framework algorithm for a given dataset

Given data $X_{i,j,k}$ and labels Y_i where i is trial, j is channel, k is sample
Sample without replacement $X^{a}_{i1,j,k}$ and $X^{b}_{i2,j,k}$ for class a and b respectively where i1 = i2
for all sliding windows of 5 samples (latency) or electrodes at window of interest (topography) **do**
    extract features from X_i1,X_i2
    Premute and merge both classes and generate labels Y
    for i=0, loop until i >= 20 **do**
        Split to training and testing datasets
        model = trainSvm(X_{train}, Y_{train})
        Run MonteCarloBootstrap($model, X_{test}, Y_{test}$)
        i = i + 1
    end for
end for
function MONTECARLOBOOTSTRAP(model,X,Y)
    Accuracies = model.predict(X) == Y
    for i = 0, loop until i >= 1000 **do**
        meanAccuracies[i] = mean(sampleWRep(Accuracies))
    end for
    return confInt(meanAccuracies,95), mean(meanAccuracies)
end function

Initially, an approach utilizing a multitude of features in combination with a feature selection algorithm was used. From each dataset, the following features were calculated: power spectral density (PSD), cross power spectral density (CPSD), autocorrelation, cross-correlation, kurtosis, skewness, and mean. These features were originally calculated on larger windows of 50 samples to increase the frequency resolution. Features were generated for each trial to form N matrices of dimensionality T x F where N was the number of conditions, T was the number of trials, and F was the number of features. Features were passed through the mutual information based feature selection algorithm minimum redundancy maximum relevance (mRMR) which selects non-redundant features that maximize the difference between the two classes (Peng et al., 2005). mRMR is a supervised method as it has access to class labels within the training set. The feature selection algorithm required the discretization of the input data. Following recommendations of the original author, data were divided to 5 bins as follows: below 2 SDs of the data, between 2 SDs and 1 SD below mean, between 1 SD below and 1 SD above the mean, between 1 SD and 2 SDs above the mean, higher than 2 SDs above the mean. This process was applied to the training set where 50 features are selected to maximize differences between classes (calculated on testing set).

This, however, proved to be processing-heavy and didn't offer a significant improvement over smaller subsets of features. Several runs were made on different feature combinations where the fastest, highest scoring combination was to be used for later analysis. The finalized features comprised the amplitude values for the given samples, their mean, variance, kurtosis, and skewness. All further analyses and results discussed in the present thesis were done utilizing only those 4 feature types. After feature extraction was completed, the two tensors corresponding to the two analyzed conditions were each split into

a training and testing set. A training (testing) set contained an equal number of observations from each condition and was comprised of 70(30)% of observations from the original dataset.

As the main goal of this study was to formulate a difference detection automated algorithm, improving accuracy ratings beyond the point of significance was not the main focus. Therefore, intricate methods that require considerable attention to hyperparameter and design were discarded for a simpler and more robust alternative. The classifier used for the present thesis was the support vector machine (SVM). The MATLAB version of libsvm library (version 3.12) was used. The linear SVM was used with the cost parameter set to 1. The training set was passed to the classifier as a matrix N of dimensions $O \times F$ where F_i was the feature vector of an observation O_i. After training was completed, the testing set was passed through the generated model for label prediction. The labels predicted were then assessed as presented in depth in the next section. This process (dataset split, training, prediction, and validation) was run 20 times where all values were averaged post loop completion. The repetition of the above procedure was in place to ascertain that a particular random split of the training/test sets did not erroneously bias the results.

2.5 Validation Methods

2.5.1 Pre-analysis permutation and randomization

A certain degree of bias in an analytical process can be introduced by maintaining the same ordering of observations. For instance, if a machine learning algorithm is always given the same training set, the goal of improving accuracies becomes finding a set of hyperparameters, algorithms, and features that work exactly for that ordering to yield an increase

in model efficiency. Since a researcher normally runs the data through a pipeline several times, this bias increases considerably. To counteract that, several permutation points were added before major analysis nodes. This is prominent in three parts: choosing trials in cases of condition skewness, averaging trials to form new observations, and dividing the dataset into testing and training sets. The first occurs when there is a mismatch between the number of observations in the two classes. For instance, in an MMN study the number of standard trials are outnumbered by the number of deviants; this is a requirement for the ERP to be elicited. To maintain an equal number of observations for each class, the dominant class is randomly sampled without replacement for the number of observations in the smaller class. Undergoing that process strengthens the generalization of results as there is a higher chance of including more of the captured data in the analysis. Secondly, when averaging multiple trials to form individual observations, bias can ensue due to time-driven effects. For example, a subject might show stronger P300 responses in the beginning of an experiment in contrast to near the end; averaging randomly ordered trials counteracts that bias. The last works to eliminate bias given by constant splits of the data at arbitrary points and providing a higher chance that an observation is seen both as a training point and as a testing point in different iterations.

2.5.2 Monte Carlo bootstrapping

In order to assess the effectiveness of a model in discerning the difference between two conditions, there needs to be a measure that is robust to random variation in results. This need is accentuated by the many comparisons and models being trained on the different time windows, electrodes, and trial aggregations, which increase the likelihood of both type I and II errors. The approach taken was to use Monte Carlo bootstrapping on each

of the trained models to generate both a mean value over multiple label predictions, and confidence intervals which serve as the link to traditional statistical significance in the present thesis.

This method follows closely what Ramkumar et al. (2013) utilized in validating the results of machine learning applications on MEG data.The validation submodule takes a learned model, and a testing set. It initializes by generating the predictions of the model given the training data and comparing them with the labels generating a binary vector of correctly vs. incorrectly classified observations. It then proceeds to sample with replacement from the binary vector another vector equal in length and averages it to form a single mean accuracy. This process is repeated 999 times where all means are later observed and the 95% confidence intervals are calculated. By definition, if the confidence intervals do not overlap with chance level, a significant difference has been found. Moreover, it is possible to compare these intervals with data portions that are known not to contribute to meaningful signals as a confirmatory check during processing time.

Chapter 3

Results

3.1 Coarticulatory violation

Two different runs of this dataset through the pipeline were carried out. The first corresponded to dividing the data into the two main conditions, congruent and incongruent coarticulation, disregarding the effects of specific vowels or consonants. Differences between the two conditions were minimal and barely visible in the grand average (see figure 3.5). In this analysis, the data was split where each half contained all remaining trials (after preprocessing) for all subjects. The run yielded no significant differences between the two conditions as a whole.

The second run had the two classes constrained by either the consonant or vowel in a stimulus word. Since the number of significant results was small in comparison with all possible sub-conditions, only significant, or near-significant results, are discussed in detail.

No difference in signal was detected when constraining the dataset by vowels. This was characterized by accuracy confidence intervals that overlap, sometimes completely, through the post-stimulus interval with ones from the pre-stimulus presentation period. That was

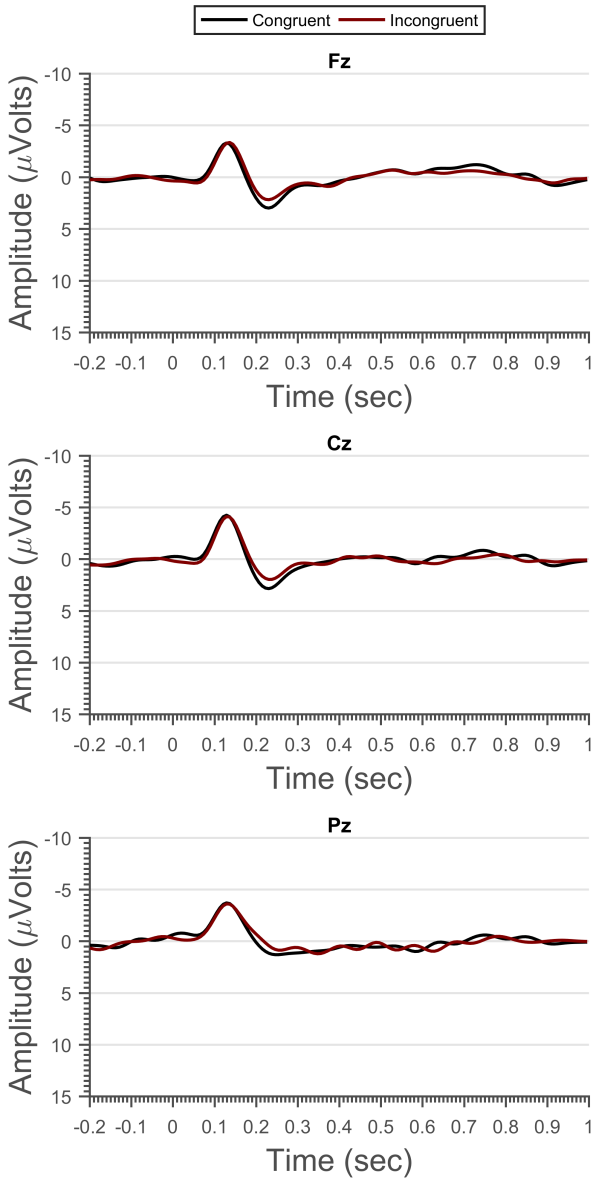

Figure 3.5: Brain responses to congurent and incongruent stimuli in the coarticulation violation paradigm for all consonant and vowel types. Equal number of trials for each condition was sampled and then averaged across all subjects to form the grand averages displayed. Trials extend from 200 ms before stimulus onset to 1000 ms after. Due to space constraints and for ease of visibility, only the medial line electrodes of Fz, Cz, and Pz are plotted.

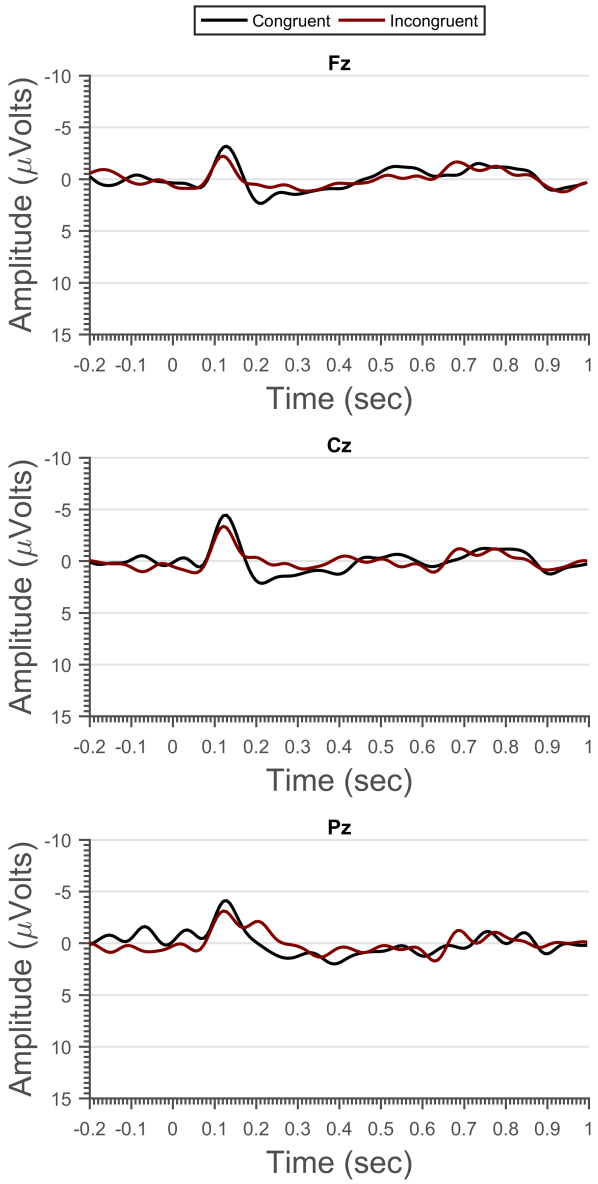

Figure 3.6: Brain responses to congurent and incongruent stimuli in the coarticulation violation paradigm constrained to the words beginning with the /p/ consonant. Equal number of trials for each condition was sampled and then averaged across all subjects to form the grand averages displayed. Trials extend from 200 ms before stimulus onset to 1000 ms after. Due to space constraints and for ease of visibility, only the medial line electrodes of Fz, Cz, and Pz are plotted.

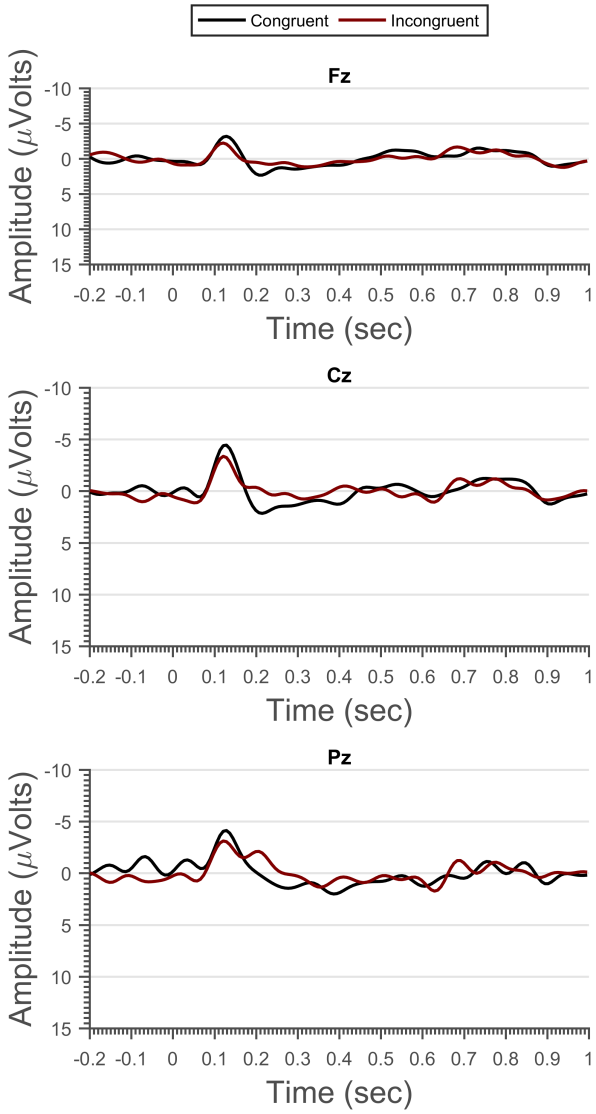

Figure 3.7: Brain responses to congurent and incongruent stimuli in the coarticulation violation paradigm constrained to the words beginning with the /t/ consonant. Equal number of trials for each condition was sampled and then averaged across all subjects to form the grand averages displayed. Trials extend from 200 ms before stimulus onset to 1000 ms after. Due to space constraints and for ease of visibility, only the medial line electrodes of Fz, Cz, and Pz are plotted.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	55.05	61.54	48.56
Two	54.68	63.89	45.59
Four	60.43	73.05	47.64
Eight	59.16	77.41	40.83
Sixteen	69.51	91.92	45.38

Table 3.1: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs intensity deviant) in the coarticulatory violation dataset constrained by the consonant /p/ at the highest time point close to the PMN region: 184 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	54.84	60.02	49.58
Two	56.21	63.51	48.88
Four	62.88	73.04	52.26
Eight	59.67	74.05	44.64
Sixteen	66.2	84.29	46.9

Table 3.2: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (congruent vs incongruent) in the coarticulatory violation dataset constrained by the consonant /t/ at the highest time point pertaining to the PMN region: 238 ms post stimulus onset. Results are reported for all five averaging settings.

also manifest in confidence intervals crossing random chance. Since the temporal aspect of the signal showed no accuracy peaks, the topographical step was omitted.

When constraining by the type of consonant, \t\ showed a significant accuracy peak in the 230 ms post-stimulus onset region (see figure 3.7 and 3.8 for grand average and analysis results, respectively). No other peaks showed a significant difference to baseline or chance level. The peak was observed to be more negative in the coarticulatory violation (incongruent) condition in comparison to correct-spliced words (congruent). The bump in accuracy can be seen in single trial analysis, but the confidence intervals fail to rise above chance level. As the number of averaged trials per observation rises, the accuracy rises to significance (see table 3.2). The effect on accuracy of the increased trials-averaged per observation can be seen in figure 3.9.

The /p/ sound showed a visible difference between the two conditions when inspecting the grand averages (see figure 3.6). This difference was not significant according to the framework proposed in this thesis (figure 3.8). The visually observable difference is argued to have dissipated after the reduction of trials to an equal number per condition. The effect of randomly sampling trials followed by averaging to form observations is shown in figure3.9. It can be argued that a signal is observed in the averages-of-eight run, but the confidence intervals cross the 50% boundary (see table 3.1). The constant increase in accuracy with increased averaging, failing in larger averages due to the variance caused by low number of total observations, provides evidence to suggest that a larger dataset can result in a positive result. In other words, the present framework's conservative approach is likely to have been responsible for a false negative in this particular sub-condition analysis.

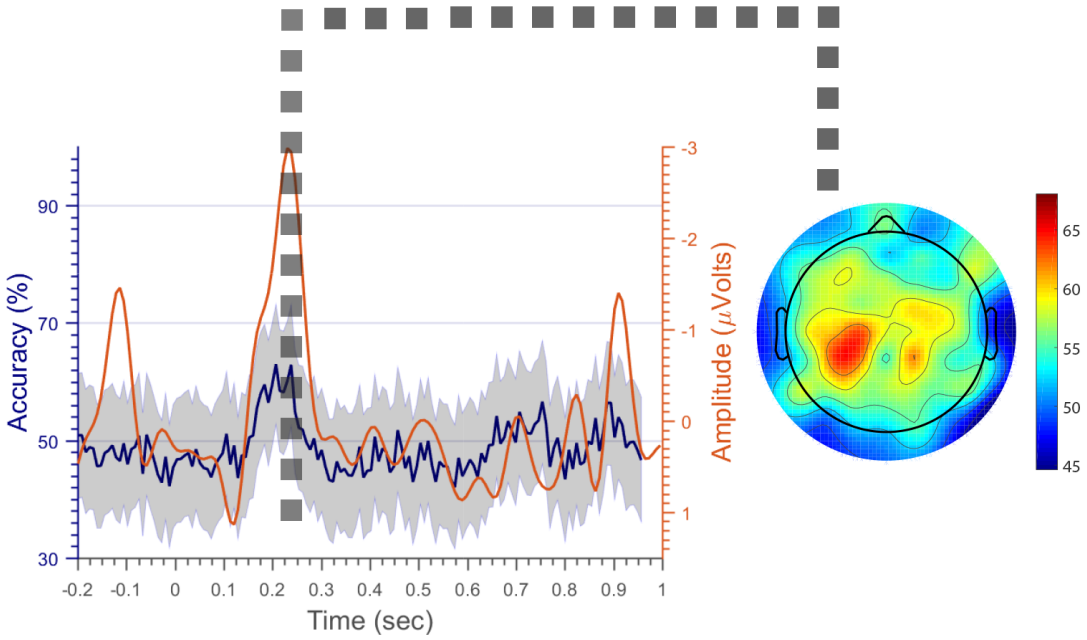

Figure 3.8: Differences between elicited brain responses to congruent and incongruent deviant tones in the coarticulation violation paradigm constrained to words starting with the consonant /t/. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference of the Cz electrode between responses to incongruent and congruent vowel sounds' grand averages. The accuracy maximum corresponding to the PMN is analyzed topographically to generate the topography plot of electrode accuracies (in %).

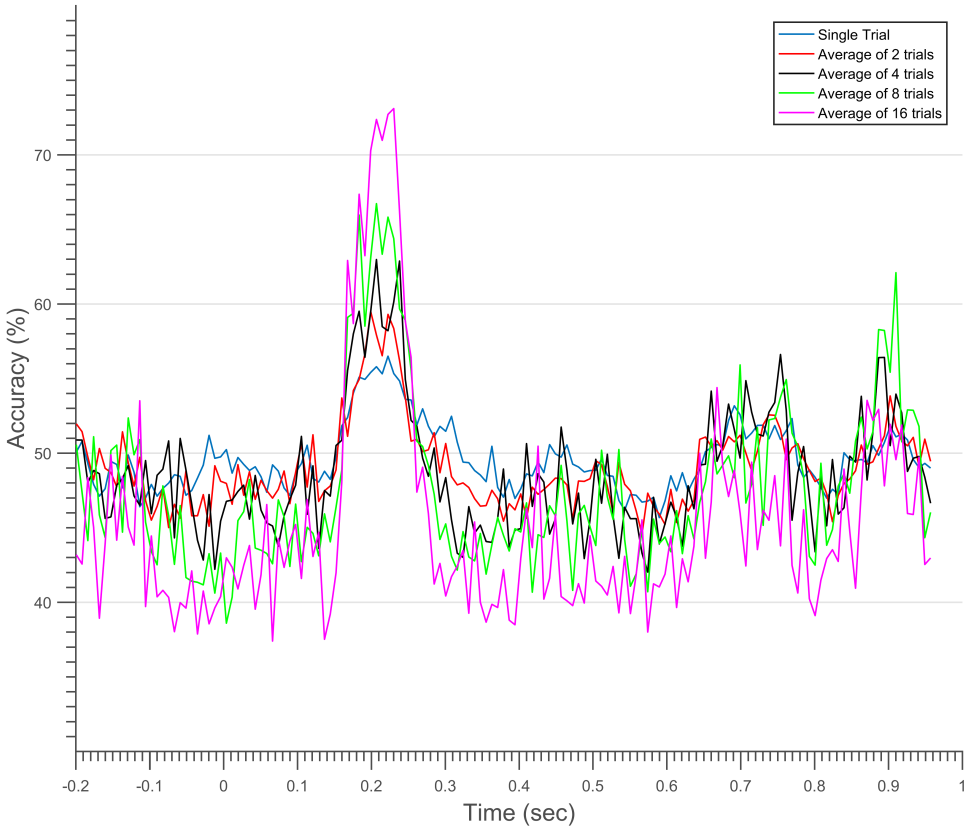

Figure 3.9: The accuracy of correctly classifying between congruently and incongruently -spliced words starting with the /t/ consonant in the coarticulatory violation paradigm. The five curves show accuracies using the proposed methods across the 5 different averaging settings.

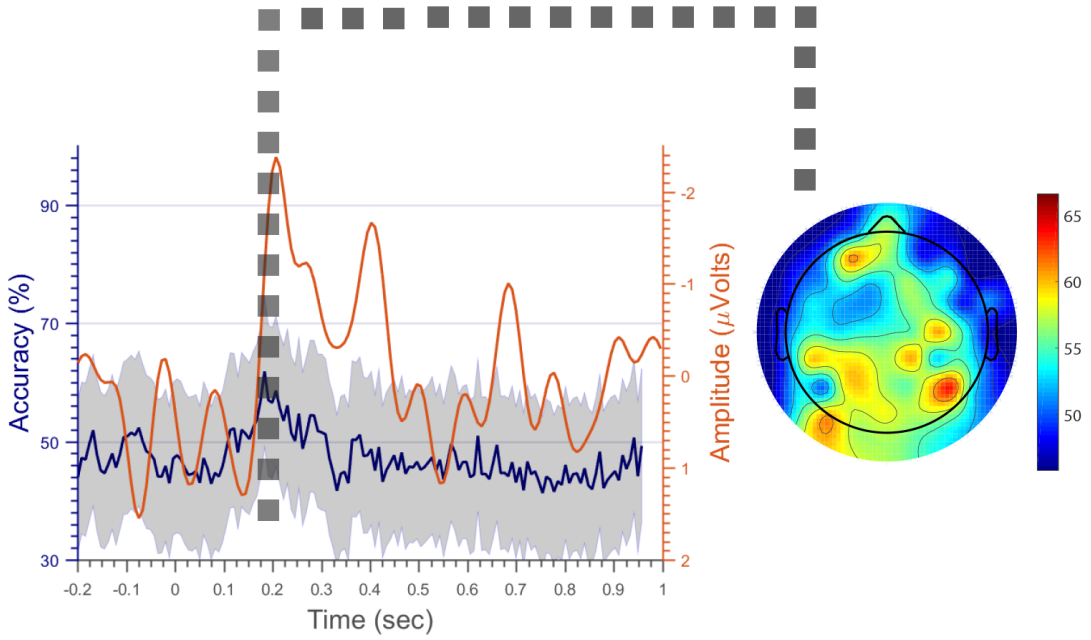

Figure 3.10: Differences between elicited brain responses to congruent and incongruent deviant tones in the coarticulation violation paradigm constrained to words starting with the consonant /p/. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference of the Cz electrode between responses to incongruent and congruent vowel sounds' grand averages. The accuracy maximum corresponding to the PMN is analyzed topographically to generate the topography plot of electrode accuracies (in %).

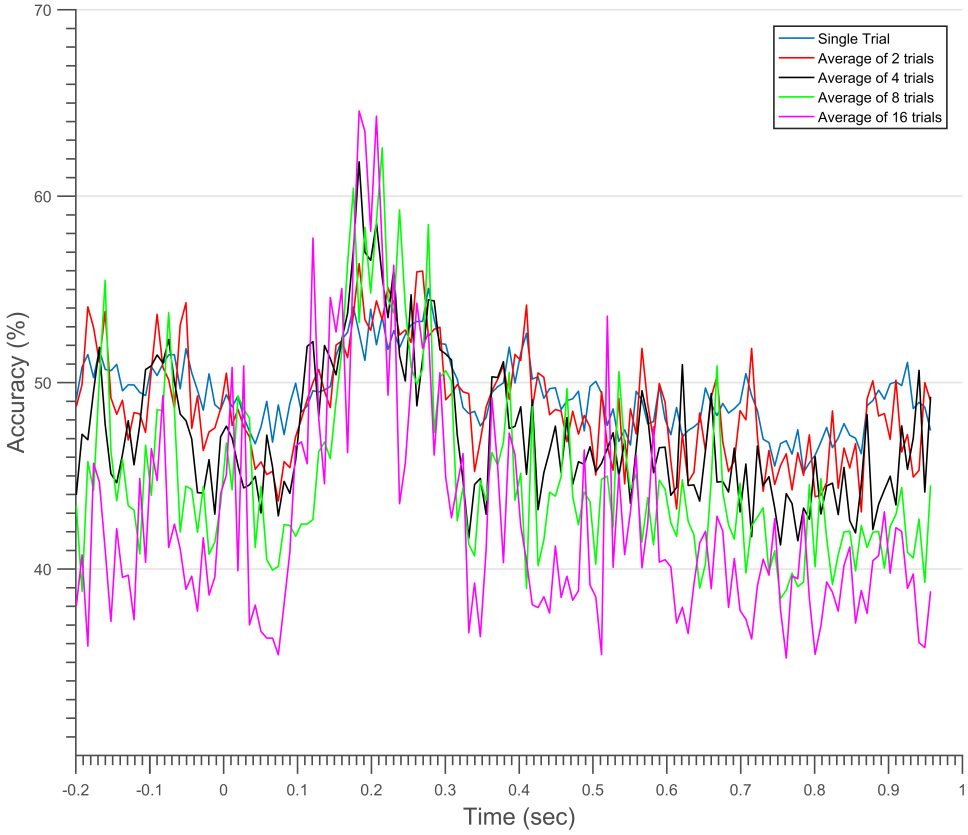

Figure 3.11: The accuracy of correctly classifying between congruently and incongruently -spliced words starting with the /p/ consonant in the coarticulatory violation paradigm. The five curves show accuracies using the proposed methods across the 5 different averaging settings.

3.2 Mismatch negativity

Since the paradigm involved 3 different types of deviants, the algorithm was run separately for each against standard tones. Grand averages for all conditions exhibited visually identifiable responses as seen in figure 3.12. In the latency analysis, all three types of deviants showed significant intervals around both the MMN and the N100 ERP regions compared to the standard condition. Additionally, another region of significance manifested around the 200 ms region and was identified as the P200. The difference waves between responses to deviants and standard tones as recorded by the Cz electrode, chosen for visualization due to its central head position, can be seen as the red plots in figures 3.17, 3.13, and 3.15.

Results from the latency analysis on deviant responses are shown in figures 3.17, 3.13, and 3.15. The blue curves correspond to the accuracies of models generated using features extracted from 6 consecutive samples (corresponding to approximately 47 ms). The abscissa for the curve corresponds to the initial time of a particular window and extends to the full window length for feature extraction. The shaded region surrounding the curve corresponds to 95% CI of accuracies when classifying between the two conditions (classes). The blue plot is taken from the average of 4 trials run for consistency across the present thesis; a larger number would cause accuracy saturation (almost 100%) for some comparisons, while lesser trial average would not show a significant difference for others. Confidence intervals for topographies are not included in the the aforementioned plots for visual simplicity. It is important to note that the MMN dataset contains four times the trials present in the P300 dataset due to the difference in number of presented stimuli. This is expected to cause more stability over higher trial-averages and smaller confidence intervals in general.

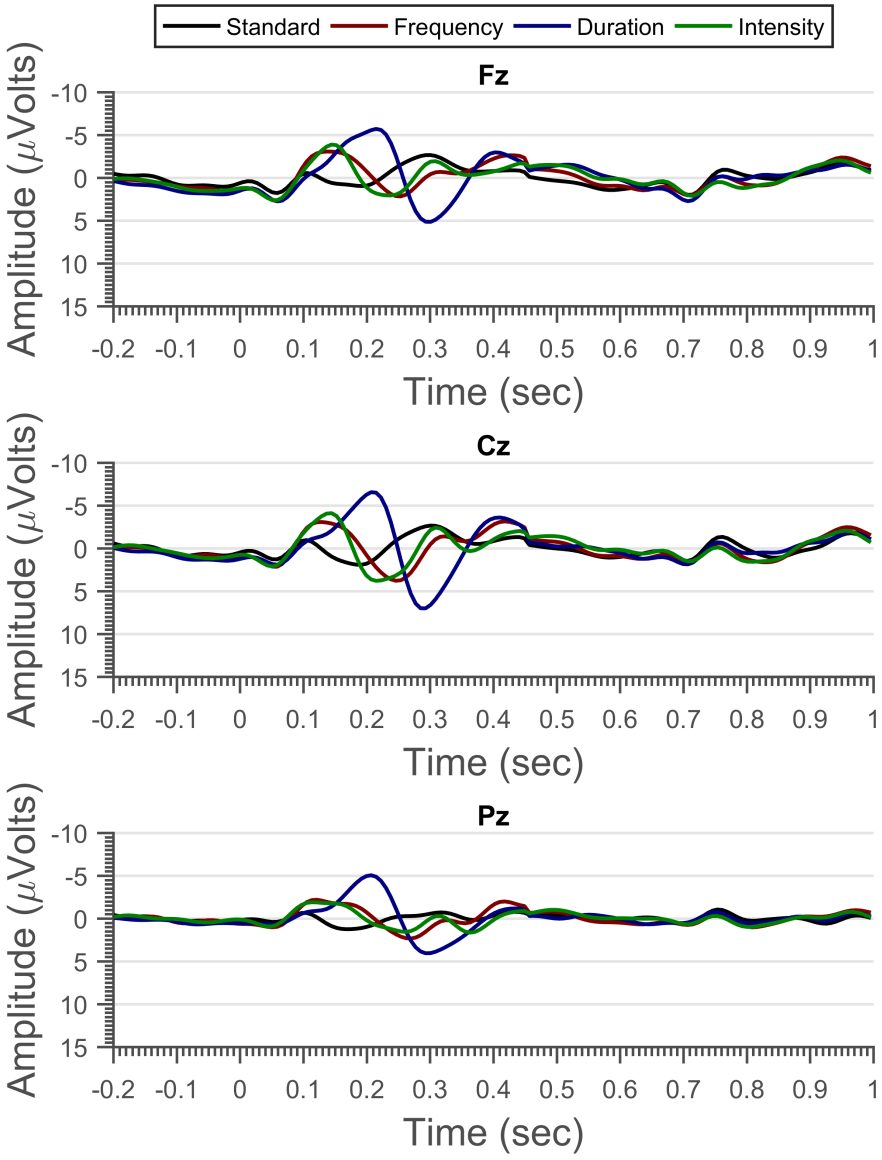

Figure 3.12: Brain responses to the four types of stimuli in the MMN paradigm. Equal number of trials for each condition was sampled and then averaged across all subjects to form the grand averages displayed. Trials extend from 200 ms before stimulus onset to 1000 ms after. Due to space constraints and for ease of visibility, only the medial line electrodes of Fz, Cz, and Pz are plotted here.

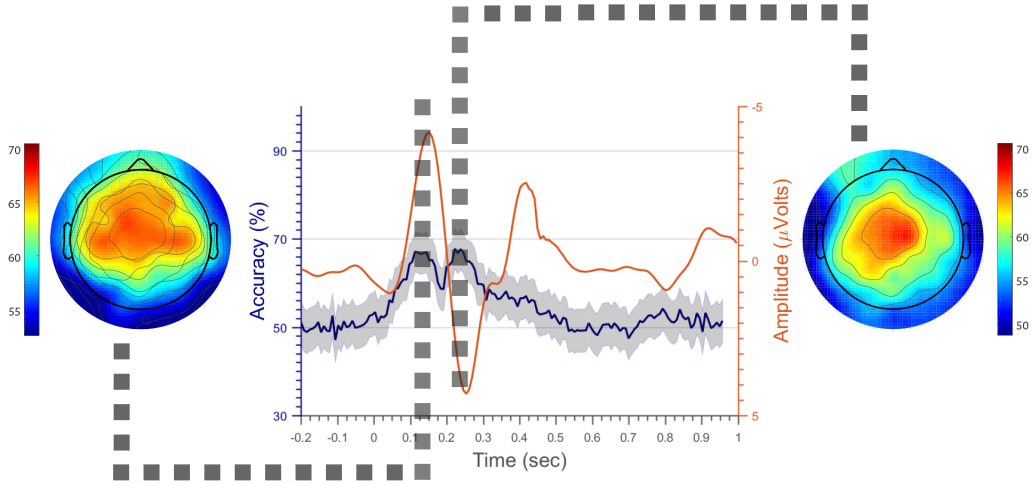

Figure 3.13: Differences between elicited brain responses to standard and frequency deviant tones in the MMN paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to frequency deviants and standard tones. The accuracy maxima corresponding to the P2 and MMN are analyzed topographically to generate the right and left topography plots of electrode accuracies (in %), respectively.

3.2.1 Frequency deviants

For the frequency deviant, significant accuracy peaks manifested at the expected latencies corresponding to the the MMN and a later positive-going signal as seen in figure 3.13. In further discussion of this positive wave, it was identified as the P200. Reasons for identifying it as the P200, as opposed to the P300, are discussed in detail in the next chapter. Accuracies for trial averages at the maximal points detected for these two ERPs can be seen in tables3.4 and 3.3.

The results of the five different trial-averaging approaches can be seen in figure 3.14.

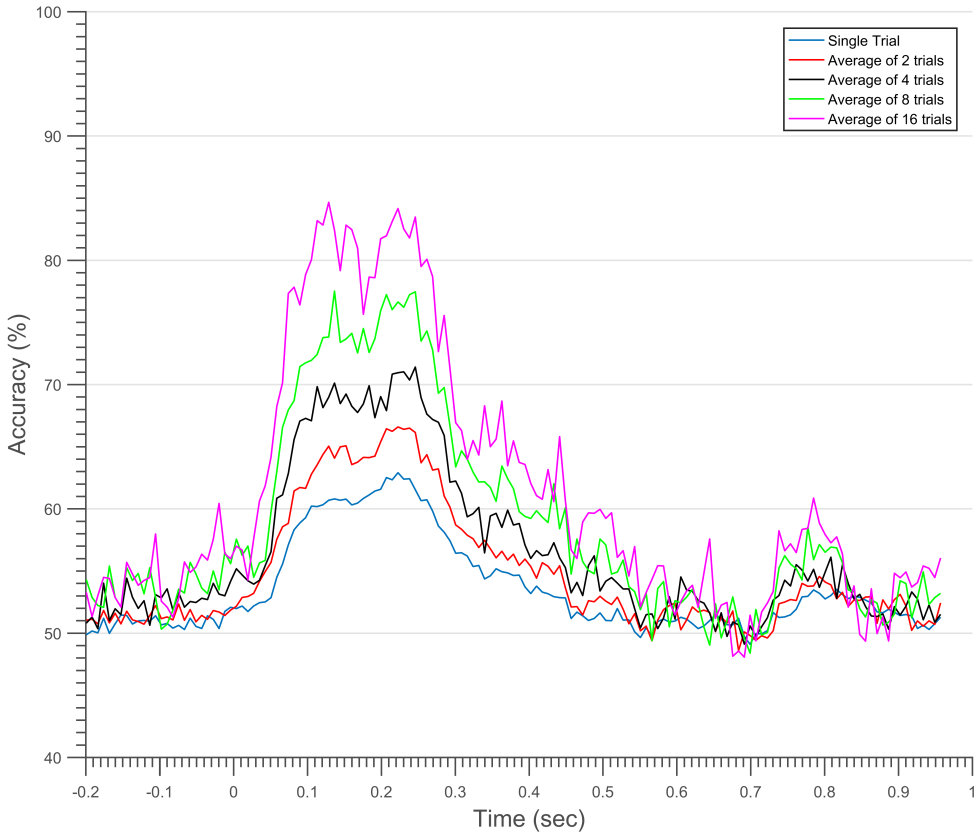

Figure 3.14: The accuracy of correctly classifying a frequency deviant from a standard tone trial in the MMN dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	61.53	63.92	59.12
Two	66.14	69.36	62.84
Four	71.4	75.78	66.88
Eight	77.46	83.13	71.55
Sixteen	83.47	90.51	75.96

Table 3.3: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs frequency deviant) in the MMN dataset at the highest time point pertaining to the P200 region: 246 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	60.79	63.21	58.39
Two	64.07	67.37	60.75
Four	70.11	74.53	65.59
Eight	77.51	83.13	71.67
Sixteen	82.4	89.47	74.6

Table 3.4: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs frequency deviant) in the MMN dataset at the highest time point pertaining to the MMN region: 137 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	73.18	75.33	71
Two	79.78	82.53	76.95
Four	88.87	91.85	85.73
Eight	92.93	96.18	89.17
Sixteen	97.68	99.8	94.62

Table 3.5: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs duration deviant) in the MMN dataset at the highest time point pertaining to the P200 region: 254 ms post stimulus onset. Results are reported for all five averaging settings.

For visual clarity the CIs were omitted in the plot where only the mean accuracies are reported for each curve. It can be noted that the signal is clearly detected starting from the single-trial approach, albeit with considerably small accuracies. Accuracies grow systematically until their maximum at the average-of-16 trials run.

Results show the two peaks as equal identifiers for class differences as accuracies match at around 61%. The polarity of the first peak can be confirmed as negative using the red curve, while the second peak is shown as positive in the difference wave. Note that a rise in accuracy (blue curve) at a particular point in time does not indicate an ERP at that particular point; however, it indicates that a difference between the conditions has been detected in the 38 ms window following that point. Concretely, a rise in accuracy (blue curve) is expected before the respective increase in amplitude (either negatively or positively) in the difference wave (red curve).

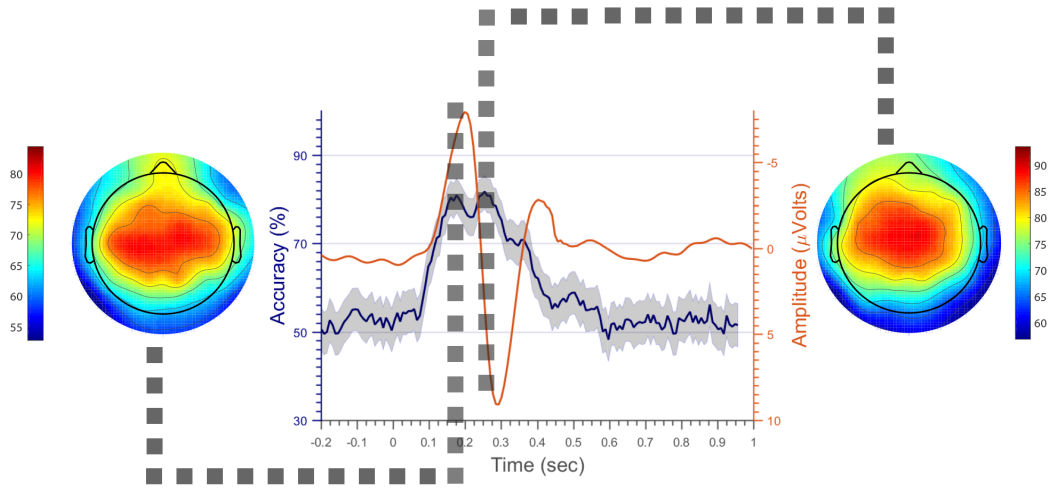

Figure 3.15: Differences between elicited brain responses to standard and duration deviant tones in the MMN paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to duration deviants and standard tones. The accuracy maxima corresponding to the P2 and MMN are analyzed topographically to generate the right and left topography plots of electrode accuracies (in %), respectively.

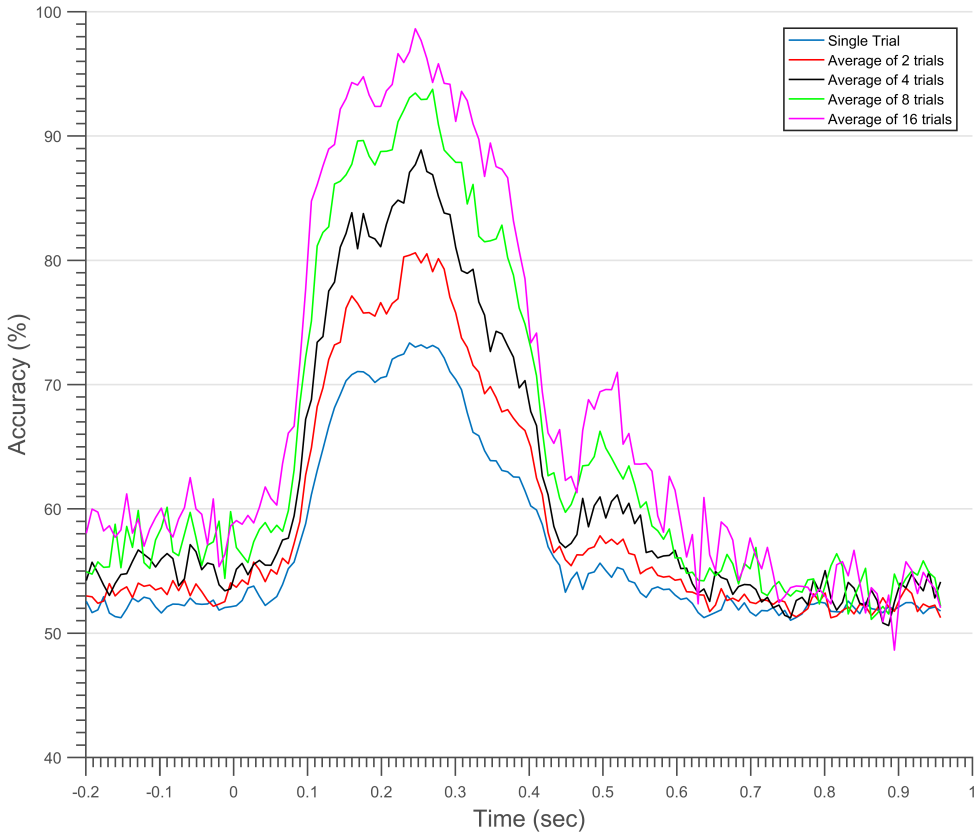

Figure 3.16: The accuracy of correctly classifying a duration deviant from a standard tone trial in the MMN dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	70.79	72.96	68.57
Two	77.13	79.98	74.2
Four	83.82	87.32	80.09
Eight	87.72	92.06	83
Sixteen	94.3	98.28	89.32

Table 3.6: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs duration deviant) in the MMN dataset at the highest time point pertaining to the MMN region: 160 ms post stimulus onset. Results are reported for all five averaging settings.

3.2.2 Duration deviants

Duration deviants produced comparable responses to the ones corresponding to frequency with a slight shift in latency observed in both accuracy and difference waves (figure 3.15). Brain responses to deviants were discernible from those to standard tones at the MMN peak with 84% accuracy when four trials averaged into each observation. The P200 proved to also contain differentiating features across the two conditions with accuracies matching and in some cases surpassing those seen for the MMN.

Similar to other conditions, accuracies showed a constant growth when increasing the number of trials averaged per observation with an increase in certainty as observed by smaller confidence intervals (see figure 3.16). Accuracies across the duration deviant showed the highest classification rates in the MMN dataset. The detailed results for both peaks of interest across the five averaging settings can be seen in tables 3.5 and 3.6.

Topographically, the MMN had a central distribution with lateral extension not unlike the frequency deviant's (see topography on the left in figure 3.15). Training on single

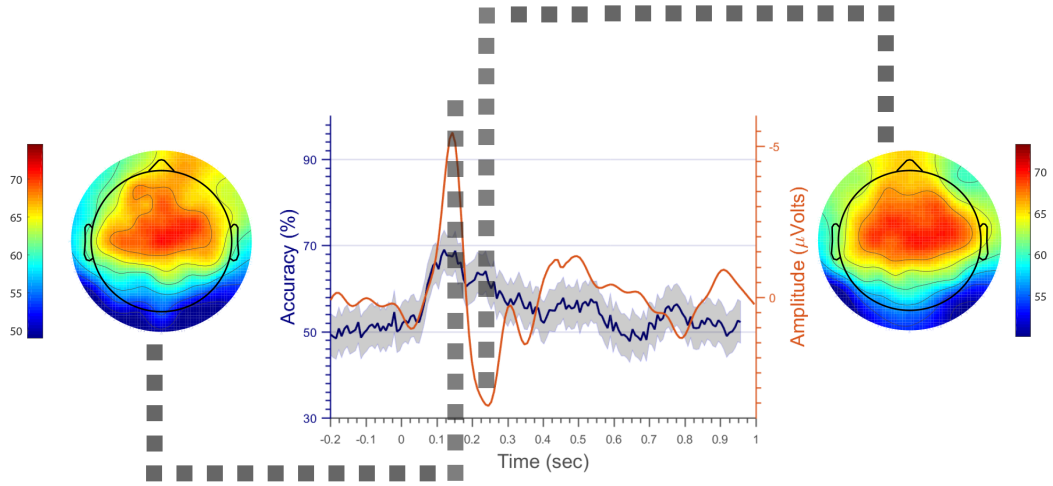

Figure 3.17: Differences between elicited brain responses to standard and intensity deviant tones in the MMN paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to intensity deviants and standard tones. The accuracy maxima corresponding to the P2 and MMN are analyzed topographically to generate the right and left topography plots of electrode accuracies (in %), respectively.

electrodes in the fronto-central region attained accuracies between 75-80% . The P200 peak showed a left, central distribution as can be seen in topography on the right (figure3.15).

3.2.3 Intensity deviants

The intensity deviants for the MMN dataset, not unlike the P300's, offered the least distinction in responses evident in lower accuracies across all runs. The latencies of both accuracy peaks were analogous to those found in the frequency deviants (see figure 3.17).

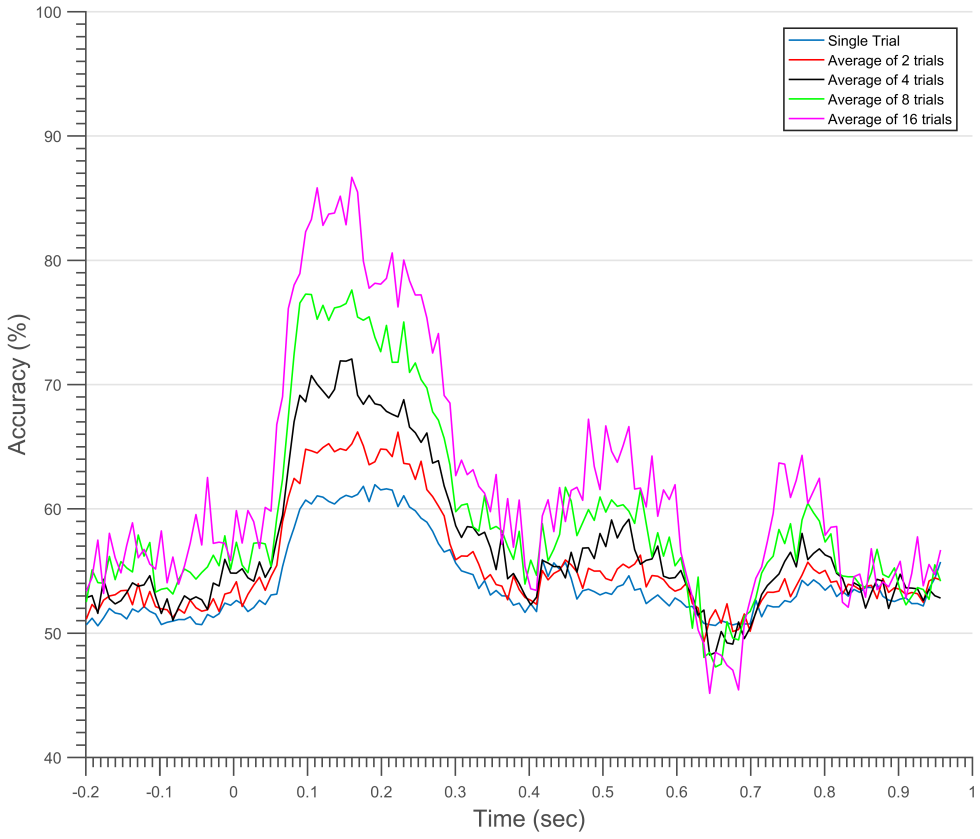

Figure 3.18: The accuracy of correctly classifying a intensity deviant from a standard tone trial in the MMN dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	61.05	63.43	58.65
Two	63.65	66.94	60.33
Four	68.78	73.29	64.22
Eight	75.03	81.02	68.86
Sixteen	80.02	87.63	71.79

Table 3.7: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs intensity deviant) in the MMN dataset at the highest time point pertaining to the P200 region: 231 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	60.39	62.8	57.99
Two	64.66	67.97	61.3
Four	70.72	75.12	66.17
Eight	77.24	82.83	71.16
Sixteen	83.28	90.15	75.86

Table 3.8: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs intensity deviant) in the MMN dataset at the highest time point pertaining to the MMN region: 106 ms post stimulus onset. Results are reported for all five averaging settings.

A consistent growth in accuracy was observed as the number of trials averaged per observation increased; however, both peaks of interest did not attain accuracies perceived for other deviants (see figure 3.18).

The topographies of both peaks show a spread across the scalp tending towards the fronto-central region as seen in both topographies in figure 3.17. Accuracies based on single electrodes were considerably closer to chance level than for other conditions. Mean accuracies and confidence intervals for both peaks can be seen in tables 3.7 and 3.8.

3.3 Attentional P300

Three latency analyses were run on the P3 dataset corresponding to comparisons between each deviant and standard tones. All three of these comparisons yielded peaks corresponding to ERPs expected after presentation of the P3 paradigm stimuli. The grand averages for deviants are plotted in figure 3.19 with standard tones over the three medial line electrodes: Fz, Cz, and Pz. The three ERPs expected in this paradigm (N100, MMN, and P300) can be seen across the head with slightly varying amplitudes and latencies. The difference waves given by subtracting responses to standard tones from the deviants, corrected for an equal number of averaged trials, are also presented as the red curves in figures 3.22, 3.20, and 3.24.

Since the P300 is considered to be a robust ERP, the number of trials in this paradigm was low (compared to the MMN dataset). This was accentuated by the progressively smaller training/testing sets being used with each iteration of averaging, reducing the number of observations by a factor of two each. As this is a direct cause for increased variance across different runs of the pipeline, the confidence intervals provided by the Monte-Carlo bootstrapping algorithm are especially important for this dataset's analysis.

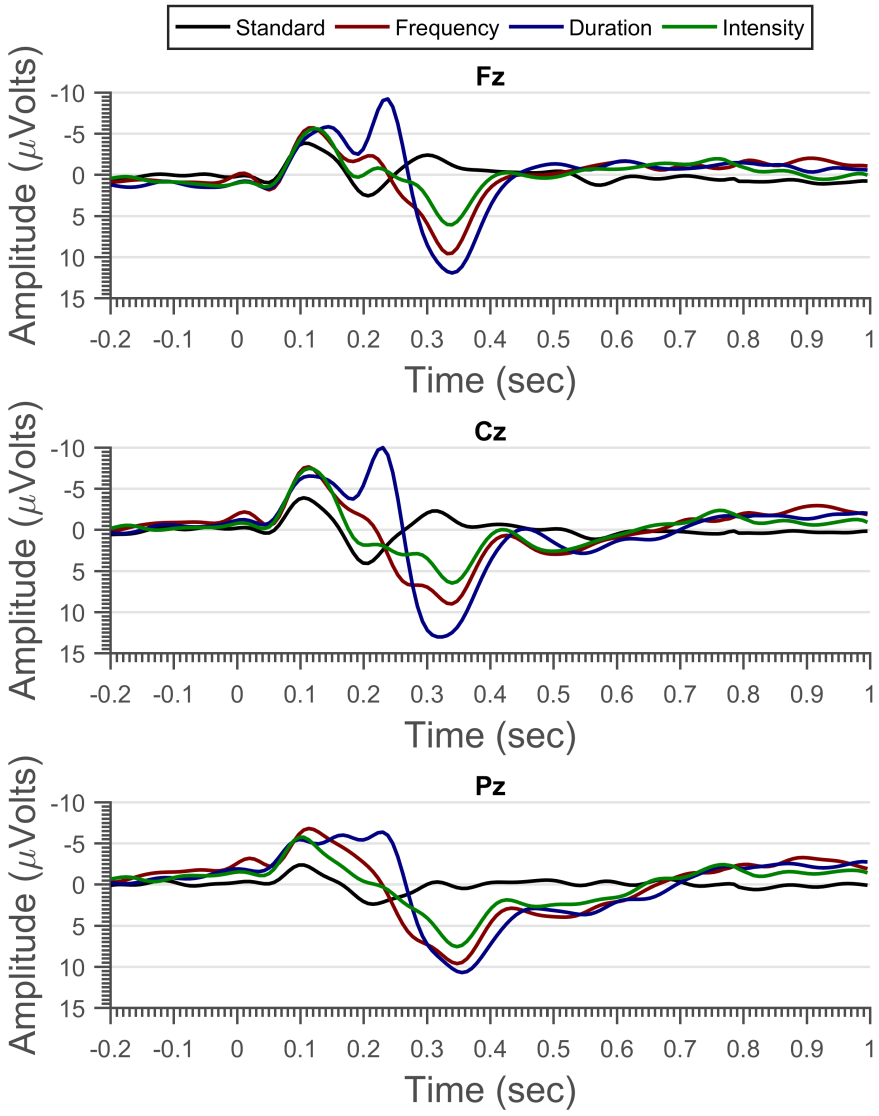

Figure 3.19: Brain responses to the four types of stimuli in the attentional P300 paradigm. Equal number of trials for each condition was sampled and then averaged across all subjects to form the grand averages displayed. Trials extend from 200 ms before stimulus onset to 1000 ms after. Due to space constraints and for ease of visibility, only the medial line electrodes of Fz, Cz, and Pz are plotted here.

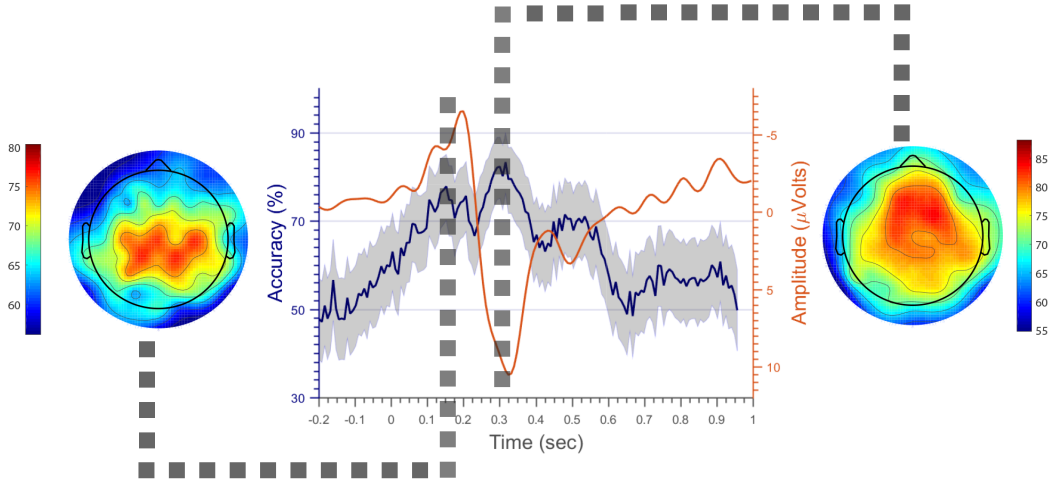

Figure 3.20: Differences between elicited brain responses to standard and frequency deviant tones in the P300 paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to frequency deviants and standard tones. The accuracy maxima corresponding to the N200 and P300 are analyzed topographically to generate the left and right topography plots of electrode accuracies (in %), respectively.

For all deviant types, an early negative component was detected that was slightly delayed compared to MMNs from the previous dataset with similar topographies. An initial categorization as late MMNs was discarded for reasons discussed in detail (see chapter 4). In this section, all instances of the early negativity will be referred to as N200s.

3.3.1 Frequency deviants

Brain responses to the frequency deviants show a spread of the early negativities (figure 3.20). The P300 component shows similar amplitude with a slight delay of around 25 ms

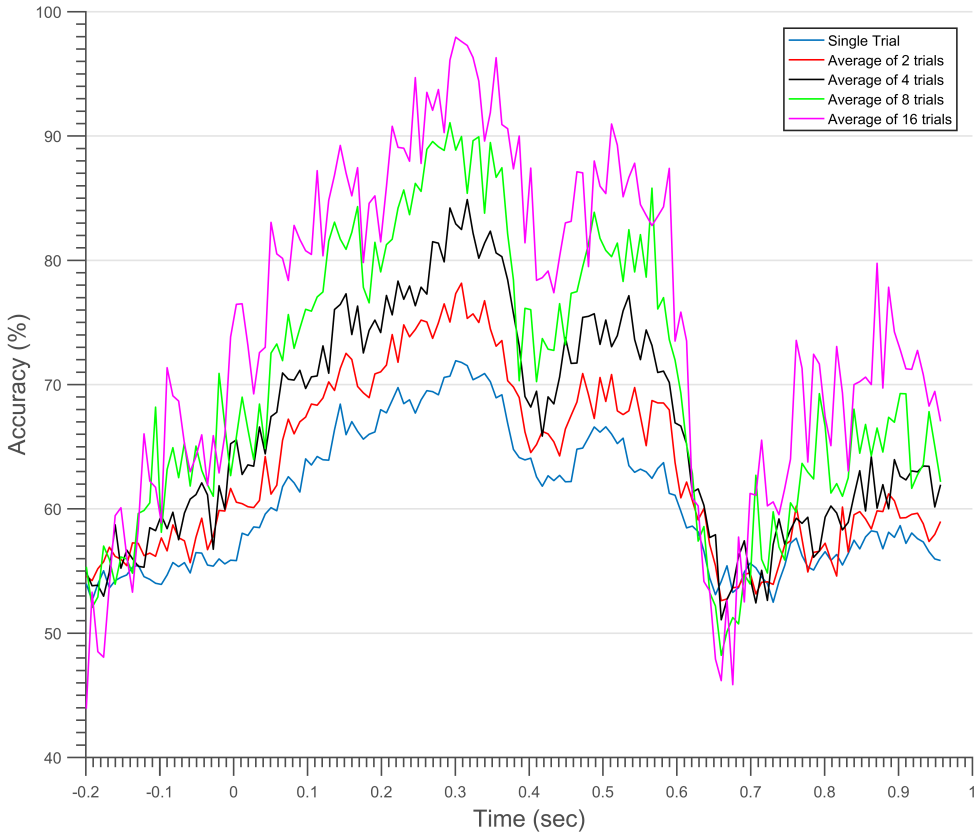

Figure 3.21: The accuracy of correctly classifying a frequency deviant from a standard tone trial in the P300 dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	71.52	75.68	67.26
Two	75.32	80.84	69.57
Four	84.88	91.17	77.91
Eight	85.38	93.91	75.59
Sixteen	97.28	100	92.04

Table 3.9: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs frequency deviant) in the P300 dataset at the highest time point pertaining to the P300 region: 316 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	65.96	70.31	61.46
Two	72.51	78.25	66.62
Four	77.3	84.71	69.32
Eight	80.87	90.55	70.18
Sixteen	87.02	97.96	73.52

Table 3.10: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs frequency deviant) in the P300 dataset at the highest time point pertaining to the N200 region: 152 ms post stimulus onset. Results are reported for all five averaging settings.

compared to the duration deviant's. The analysis modules discussed in this thesis were able to detect two main components corresponding to two time regions. This is observable by an increase in accuracy as the sliding window approaches these regions of signal difference. The peaking accuracies are at 152, 316 ms with polarities negative, and positive, respectively, as observed in the difference wave in figure 3.20.

A detailed look into the originating patterns in classification, when using the different number of trials per average, can be observed in figure 3.21. As expected, noise increases as the number of aggregations increases peaking at the average of 16 trials per observation run. As the signal to noise ratio in each observation rises, as expected from averaging, accuracies exhibit significant increases as well. Most importantly, the plot shows that even single-trial analysis of the P300 frequency deviants shows mean accuracies above 70%. On the other extreme end, averaging 16 trials for each observation was able to yield above 95% mean accuracy. Mean accuracies for each detected ERP component, along with their 95% confidence intervals, can be seen in tables 3.10 and 3.9.

Single-electrode analysis at the peaks of interest yielded the topography maps shown in figure 3.20. The earlier negative waveform is shown to have been elicited centrally with slight lateral extensions on both sides. The other map, corresponding to the P300, shows a fronto-central topography peaking in the Fz region.

3.3.2 Duration deviants

By visual inspection of the duration deviant difference waveform, 2 peaks can be identified: a 200 ms negativity, and a 300 ms positivity. The primary ERP observed in this experiment was the P300 and is seen to be close to a 10 μV difference between the two conditions at Cz. The negative early peak lies in the range of the MMN/N2b. The two observable ERPs

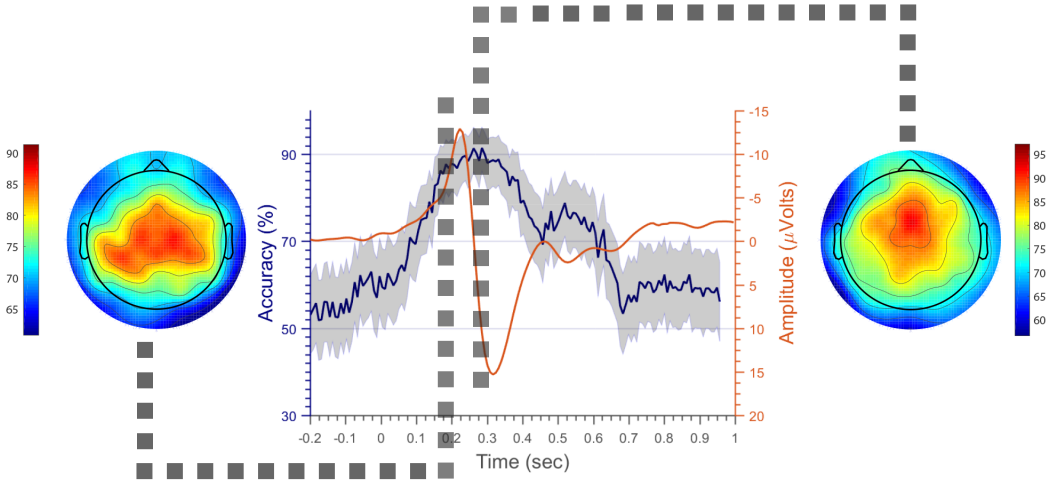

Figure 3.22: Differences between elicited brain responses to standard and duration deviant tones in the P300 paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to duration deviants and standard tones. The accuracy maxima corresponding to the N200 and P300 are analyzed topographically to generate the left and right topography plots of electrode accuracies (in %), respectively.

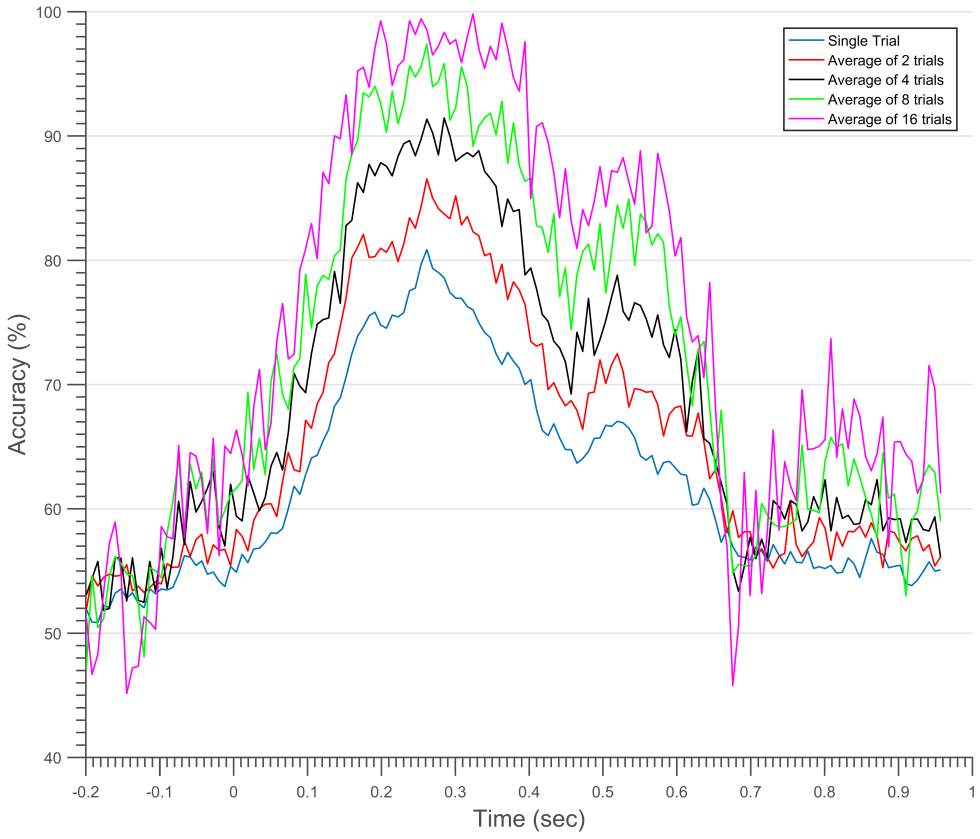

Figure 3.23: The accuracy of correctly classifying a duration deviant from a standard tone trial in the P300 dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	78.59	82.36	74.7
Two	83.71	88.41	78.79
Four	91.44	96.22	86.06
Eight	95.82	99.55	90.73
Sixteen	98.34	100	95.37

Table 3.11: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs duration deviant) in the P300 dataset at the highest time point pertaining to the P300 region: 285 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	74.78	78.73	70.68
Two	80.96	86.02	75.73
Four	87.84	93.45	81.42
Eight	92.61	98.41	85.18
Sixteen	99.25	100	97.78

Table 3.12: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs duration deviant) in the P300 dataset at the highest time point pertaining to the N200 region: 199 ms post stimulus onset. Results are reported for all five averaging settings.

are robustly detected by the ML analysis as can be seen in Figure 3.22 where accuracies are displayed taking the average of four trials as one observation. While duration deviants show a delayed negative peak, the P300 response in the grand average is shown to arise at exactly 300 ms post stimulus presentation; this shows an earlier brain response time compared to the other two deviants in this dataset.

The duration deviant provided the highest accuracies in both MMN and P300 datasets. This is especially notable in the incremental rise of classification accuracy as the number of averaged trials increases (see figure 3.23). Duration deviants elicited a salient P300 response with more than 90% classification accuracy on average of four trial runs and above (see table 3.11). Additionally, the detected N200 peak offered comparable, in some cases higher, accuracies as can be seen in table 3.12. It is important to note that compared to the slightly weaker robustness of the N200 peak in response the other deviant types, the duration deviant counterpart was as salient as the P300.

In terms of topography, the two observed peaks where analyzed as outlined in the methods chapter yielding two topography maps for the N200 (left), and P300 component (right) as seen in figure 3.22. The ability to discern the deviant from the standard at the early negative peak showed a fronto-central distribution that extends laterally to both sides. Features extracted from single electrodes in the middle of that region yielded up to 85% accuracy on the average of four run. The P300 component offered a strong fronto-central topography reaching 90% accuracy for observations constituting four trials each.

3.3.3 Intensity deviants

As can be seen in figure 3.24, the intensity deviant offered the lowest accuracies compared to the other deviants. The waveform was very similar to that of the frequency deviant's, but

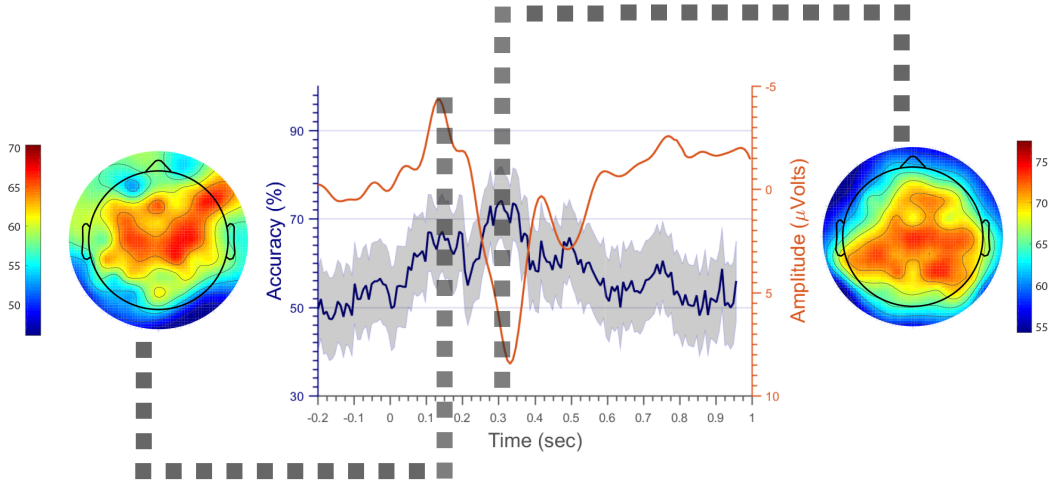

Figure 3.24: Differences between elicited brain responses to standard and intensity deviant tones in the P300 paradigm. Accuracies obtained from classification of sliding windows across time (abscissa) between both conditions are shown as the blue plot corresponding to the left axis. The shaded region corresponds to the 95% CIs as reported by the Monte Carlo bootstrapping submodule. The red plot's ordinate corresponds to the right axis and displays the difference at the Cz electrode between grand averaged responses to intensity deviants and standard tones. The accuracy maxima corresponding to the N200 and P300 are analyzed topographically to generate the left and right topography plots of electrode accuracies (in %), respectively.

Figure 3.25: The accuracy of correctly classifying a intensity deviant from a standard tone trial in the P300 dataset using the proposed methods and showing the differences across the 5 different averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	65.16	69.51	60.63
Two	68.7	74.77	62.52
Four	77.13	84.82	68.98
Eight	78.81	88.91	67.73
Sixteen	91.31	99.44	80.37

Table 3.13: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs intensity deviant) in the P300 dataset at the highest time point pertaining to the P300 region: 340 ms post stimulus onset. Results are reported for all five averaging settings.

# of trials per observation	Mean %	Upper bound %	Lower bound %
Single	64.68	69.06	60.18
Two	65.84	72.03	59.68
Four	71.26	79.27	62.55
Eight	73.42	84.55	61.55
Sixteen	75.86	90.74	59.54

Table 3.14: Mean accuracies, and their 95% confidence intervals, of the learned model in predicting the correct class (standard vs intensity deviant) in the P300 dataset at the highest time point pertaining to the N200 region: 160 ms post stimulus onset. Results are reported for all five averaging settings.

had small amplitudes overall in Cz with a 1.5 μV smaller P300, and a 1 μV smaller MMN. Furthermore, latencies were also homologous to their frequency deviant counterparts.

Addition of trials to averages per observation offered a consistent increase in accuracy (figure 3.25). Accuracies corresponding to the two peaks of interest were significantly lower for the N200 as seen in tables 3.13 and 3.14. This, however, was more apparent in the higher averages as the N200 had higher variance and lower mean accuracy compared to the P300.

The topographical analysis of the earlier peak of interest yielded a distribution similar to that elicited by the two other deviants. It can be clearly seen that the signal is much weaker than its counterparts in the other conditions as apparent by the low accuracies (see figure 3.24). The P300 shows a more spread signal topography and a drop in accuracy for average of four trial classification on single electrodes compared to other deviants.

Chapter 4

Discussion

Analysis of the three datasets using the proposed machine learning framework has shown results that both correlate with the literature, and shed light on aspects of ERP waveforms that are not normally analyzed. Although the relevance of the detected information is not within the focus of the present thesis, a discussion on the inferences from the data are included as an example for later application of the framework to other studies.

4.1 Dataset inferences

4.1.1 Coarticulatory Violation

The framework of analysis discussed in the present thesis was least efficient when dealing with this dataset. Utilizing the entire dataset as a 2-class classification problem yielded no results. Constraining observations to particular expected vowels, initial consonants, or both highlighted only one positive finding. The accuracy peak corresponding to the detected ERP in the incongruent condition while constraining the dataset to words beginning with

the /t/ sound is seen to correspond to the PMN (Connolly and Phillips, 1994). Accuracy peaks at the window starting 238 ms post-stimulus onset placing the detected signal at the characteristic PMN latency region. The consonant sound /p/ offered comparable results which did not pass the significance threshold. The peak detected in the /p/ constraint was also relatively early to be identified as a PMN.

Note that when constraining to the /t/ sound, all vowel combinations are left intact in the data. It can be postulated that certain vowel sounds and combinations can be more discernible, mirroring consonant differences, and possibly resulting in a stronger ERP response when sufficient constraints are enforced; however, the data provided had a limited number of trials that were spread across all subconditions. Constraining by both vowel and consonant left the dataset with trials too limited in number to proceed with the analysis pipeline. In general, the lack of distinction between the two classes when constraining by all but one consonant sound can correspond to either of the following: 1) the framework's limitation when dealing with a relatively small ERP, and a low number of trials, 2) the PMN was not consistently generated across trials/subjects in this study. Further experimentation with the dataset using the pipeline and using direct analogies of results yielded by classical statistical methods is required to reach an understanding of the results.

4.1.2 Mismatch Negativity

As expected, the differences between responses to each deviant and standard tones included a negative peak corresponding to the MMN topography and latency (Todd et al., 2008). These findings were characterized by a significant increase in the classification accuracy between a deviant and the standard condition as the sliding window approached 150-175 ms post stimulus onset. Latency was delayed for the duration deviant, consistent with

results also found by Todd et al. (2008). Polarity of these accuracy peaks were confirmed to be negative by visual inspection of the difference waves for each relevant comparison.

It was not determined whether the peak identified as the MMN is solely composed of that ERP. Changes in the features of the higher frequency tone, compared to the standard, is expected to change the elicited N100; thus, increased differentiability between the two conditions is expected to arise from sources other than the endogenous ERP (see Näätänen et al., 2005, Näätänen and Picton, 1987). A separate condition where the high frequency tone is used as a standard was not included in the paradigm, consequently a conclusion cannot be directly formulated on the contribution of each component. The same argument applies for the other two deviants.

Another ERP was found 100 ms following the MMN for each of the deviants. Inspection of the difference waves indicated its positivity. The ERP was found to be as salient to the distinction between deviant and standard trials as the MMN. The ERP was initially identified as the P200, as opposed to the P300; however, in depth discussion of several complications to that categorization is discussed below in the next section. Notably, the positivity is not found in the original study of the MMN paradigm (Todd et al., 2008).

Results indicated robust findings that match the expected MMN response with its respective latencies and topographies. Different deviant types produced distinguishable differences in the ERP waveform (Alho, 1995, Näätänen and Picton, 1987). Most notably, the intensity deviant elicited a weaker response compared to its counterparts. This was assumed to be due to the difference between the deviant and the standard tone being hard to distinguish for some subjects. That was later confirmed by examination of the behavioral responses, to the same intensity-deviant tones, in the P300 dataset.

It is important to reiterate that four of the participants exhibited averages that had positivities around the 300 ms region indicating a likely elicitation of P3as in some of the trials. No measures were taken during preprocessing to filter those participants out to conform to the robustness criterion of the method in cases where some confounding variables cannot be avoided. This inclusion, however, has undoubtedly affected the results for the MMN dataset unfavorably. Concretely, the supervised ML approach takes a passed label (deviant, for instance), as the truth value. If these labels are not correct, then the performance of the algorithm deteriorates due to both training on wrong examples, and/or erroneously counting a correctly-classified, mislabeled observation as incorrect.

4.1.3 P300

The P300 dataset offered the most robust responses that in turn had the highest classification accuracies among other datasets discussed in the present thesis. This was expected due to the ERP's robustness, large amplitude, and low intra-subject variance Polich (2007).

In terms of the P300 latency, all three deviants elicited peaks at the 300 ms region. Differences in latency across different sites were not analyzed in depth; however, visual inspection of figure 3.19 shows comparable latencies. Amplitudes and consistencies varied slightly across deviants: responses to duration deviants being the largest and most easily detectable, frequency deviants showing slightly smaller responses, and intensity deviants eliciting smaller, higher variance responses. Topography of the P300 response was evidently fronto-central, providing strong evidence against the elicitation of the posterior P3b. On inspection of prior literature, it can be argued that this type of response was due to the lack of a memory role in the cognitive processing of the participants (Polich, 2007). The only tone in memory was the standard one, anything deviating from that norm was regarded

as a different stimulus, and thus followed by the appropriate behavioral response.

The earlier negative response offered latencies similar to peaks elicited by the MMN paradigm across all conditions, although being consistently delayed. An initial postulation that these peaks correspond to elicited MMN was later invalidated. While it was possible to attribute that delay to variance, the consistency by which all peaks (see figure 3.19) have been shifted is unlikely to be random. Comparison with results from the study by Näätänen et al. (1982) provided evidence suggesting the negativity to be an N2b. Näätänen and associates have shown that the two components in the N200 region can be dissociated into both the MMN and the N2b; an ERP that is closely tied with the P3a and showing a similar fronto-central topography. Further discussion of this issue is presented in the next section.

Contrary to hypotheses, higher intensity tones did not provide comparable results to the other two deviant sounds. A possible hypothesis is that the difference in loudness between the standard and the intensity deviant was not consistently detected by the participants. This could either be due to inter or intra -participant variability which requires further single-subject analysis.

While the number of trials per subject and condition was comparatively low to that found in MMN dataset, classification accuracies between a P300-producing deviant and standard tones were high; frequency (duration) deviant trials were distinguishable at above 70% (80%) accuracy in the single-trial runs. The small number of trials, however, affected the variance of the results. Confidence intervals for each of the analyses run on this dataset were larger than their counterparts in the MMN dataset.

4.1.4 Salience of the Duration Deviant

Upon inspection of the two datasets where the duration deviant tone was presented as a stimlus, it can be noted that responses have similar characteristics (see figures 3.15, 3.22). This is especially important given that the other two deviants had different responses cross-paradigm. This raises the issue of whether the peaks observed in the grand averages, and detected by the proposed framework, correspond to the same ERPs; essentially conveying that for this deviant type, participant attentiveness did not factor in eliciting ERP components.

Concretely, there are four notable observations that can be seen in the data: 1) responses to the duration deviant in the MMN paradigm showed a positive peak similar, although smaller in size, to the one in the P300 paradigm, 2) a double-peaked negativity can be seen in the MMN region for both paradigms and especially for the P300 one (see figures 3.19, 3.12), 3) the early negative peaks detected in the P300 paradigm for both frequency and intensity deviant responses were less consistent compared to their respective P300 peaks, and 4) responses were most consistent to duration deviants for both paradigms.

These observations were found to coincide with data reported by Näätänen et al. (1982). The study compared elicited brain responses through mismatch with a standard tone by three types of deviants: proximal higher frequency, proximal lower frequency, and extremes of high frequency. Results for when participants were asked to attend to the stimulus and respond to a type (or all) of deviant(s) were reported. Moreover, the same stimuli were used in an inattentive paradigm where participants read a book during stimulus presentation. Although sound features and tasks differed, the extremes from Näätänen et al. (1982) elicited very similar responses to what can be seen for duration deviants. For the inattentive part of the experiment, a bimodal negativity was elicited by the extremes followed by a later

positivity identified as the P3a. These peaks increased in amplitude when participants were asked to actively detect deviants. Additionally, the bimodal negativity and P3a elicitation during inattentiveness were attenuated, sometimes completely, for the proximal deviants in the paradigm. These results follow closely the observations highlighted above, suggesting a particular salience of the duration deviant and relating it to the extreme stimuli. The comparable results also suggest that what has been identified as the generic N200, is an N2b. Additionally, the positive peak elicited by the duration deviant in the MMN paradigm can be argued to be an attenuated P3a, as opposed to the exogenous P200. Hence, it can be argued that it was difficult for participants to disregard duration deviants during the MMN paradigm.

Classification of the P200 peak for the two other deviants in the MMN paradigm remains uncertain. In contrast to the unimodal P200 peak, the N200 peak in the P300 paradigm elicited by both the intensity and frequency deviants was spread across time and small in amplitude. Categorization of that peak to either the MMN or the later N2b was not pursued; however, proximal deviant results from Näätänen et al. (1982) and the accuracy plateaus spanning both ERP time regions (see figures 3.20, 3.24) suggest the overlap of the two components.

4.1.5 Mismatch Negativity and the N100

As can be observed in both P300 and MMN experiments, the MMN elicited due to the presentation of the duration deviants are slightly delayed compared to the two other types. This can be attributed to the deviant tone being longer. In order for an observer to recognize that the tone is longer than its standard counterpart, one whole duration of the standard tone (50 ms) needs to pass in addition to a variable number of milliseconds that depends on a

subject's delay in perception of duration. That hypothesis is supported by comparing the MMN waveform in figures 3.15, 3.22 to 3.20, 3.24, 3.13, and 3.17. It can be observed that the timing for the peaks being compared are both within the region for both the N100 and the MMN. An assumption that the difference being detected is only a resultant of the MMN component is simply not true as the N100 is affected directly by the features of the sounds being presented Näätänen and Picton (1987). Thus, a clear identification of either ERP component and its role in signal disparity between two conditions is not attainable, limited by both the experiment design and utilized methods.

Since the stimuli are identical in features until the 50 ms mark for both the standard and duration deviant conditions, the comparison between the two signals should yield a disparity which only highlights the MMN (Näätänen et al., 2005); however, it can be argued that the tone still being presented at 50 ms (extending for 50 more milliseconds) would play a role in the addition of an exogenous component on top of the developing MMN. The exogenous effect is expected to be comparable to sustained potentials highlighted by Picton et al. (1978). However, since comparisons in that study were done across drastically differing tone lengths, an accurate estimate is difficult. If the sustained potential is assumed to be relatively small, compared to the MMN and the onset N100, the expected elicited signal would be an N100 of comparable characteristics to the standard response overlayed on a delayed MMN. A more in-depth study of the underlying response can include subtracting the standard tone responses from the duration deviants' and delaying the result by 50 ms. The produced signal should hypothetically correspond to the isolated MMN with a slight negative shift corresponding to the short sustained potential. That analysis was not within the scope of the present thesis and thus was left for future work.

4.1.6 P3a vs P200

Analysis done on the MMN dataset showed a consistent positive peak that arose in deviant conditions after the component identified as the MMN. The latency of the positive component categorize it as either the P3a (homologous to the P300 dataset), or the exogenous P200. The dissociation between the two components, or lack thereof, can then be established by the similarities between the peaks arising at the characteristic latency across the two datasets. While a complete dissociation and categorization of the component to one of the two ERPs would be the optimal outcome, another proposition is the possible overlap of the two to form the detected peak.

Upon inspection of the accuracy and difference waveforms, it is apparent that the positive peaks show different characteristics, especially for frequency and intensity deviants (see figures 3.13 vs 3.20, 3.15 vs 3.22, and 3.17 vs 3.24). For frequency and intensity deviants, the elicited positive peaks were more than 50 ms earlier in the MMN dataset compared to their counterparts in the P300 dataset. While this early latency has been observed in P3a literature (see Polich, 2007), different elicitation latencies for the same ERP (P3a) are difficult to explain given the identical stimuli across the two datasets. Inspection of topography shows a distinction between the two positive peaks elicited by the frequency deviants. This topography effect, however, is not observable for the other two deviants. These differences are seen to provide evidence suggesting the dissociation between the positive components elicited by frequency and intensity deviants across the two datasets; P3a responses are found in the P300 dataset, P200s are elicited or emitted in the MMN dataset.

Contrary to the differences observed for the intensity and frequency deviant responses,

duration deviants elicited comparable waveforms across the two datasets. Following the arguments discussed in subsection 4.1.4, it is postulated that the salience of duration deviants was homologous to the extremes in Näätänen et al. (1982); thus eliciting a P3a even during inattentiveness.

4.2 Future directions and uses

The present thesis discussed a basis for approaching EEG/ERP analysis in semi-automation utilizing ML and validation utilities. Efficiency and accuracy of the framework in the detection of ERPs was presented using 3 sample datasets. It was demonstrated that using minimal apriori knowledge, detailed descriptions of ERPs found in a dataset can be extracted using the three criteria focused on by the framework: consistency, latency, and topography. The capability of ML techniques in dealing with the high-dimensionality inherent in EEG was further confirmed by achieving high classification accuracies on both small trial averages, and single trials.

There are, however, many limitations and constraints that prevent the proposed techniques from being a full valid replacement for traditional methods. This section highlights several direct uses of the framework, in addition to future upgrades that would increase its suitability to a variety of research questions and industrial applications.

4.2.1 Class extension

A two class constraint was put on the data for all possible comparisons. While this was possible on the three datasets previously discussed, there are many instances where that falls short of analysis requirements. For instance, interactions between conditions and

other independent variables are hard to analyze given the skeleton structure of the pipeline discussed here. Moreover, consider the P300 dataset previously discussed, data was always split to a pair of classes (standard and one type of deviant). If the researcher required comparisons between each possible pair of conditions, a direct application of the proposed framework would generate $(4-1)! = 6$ different models with their separate three measures giving 18 items of information that reduce result clarity significantly as class comparisons overlap.

An intuitive resolution to this issue would be to replace basic binary class SVMs with multi-class versions. While this appears a trivial refactoring of classifier submodules, complications would arise. For instance in the P300 example, some conditions will be similar (intensity and frequency deviants) henceforth significantly reducing the overall accuracy of an inclusive model and contributing to an increase in type II errors.

4.2.2 ERP alterations through time

A main strength of the proposed analysis design was its capability of ERP detection utilizing a small number of trials. A traditional analysis design treats the average of a subject's trials as one data point (for every channel and condition). This approach is undertaken to increase the SNR significantly in exchange for the loss of finer details across the duration of an experiment (see chapter 1 for more details). This loss can be circumvented considerably by utilizing the suggested analysis pipeline.

The development of ERPs throughout an experiment can be extracted by the addition of a module encompassing all current ones, splitting a dataset through elapsed time during an experiment. This new module can extract the required information in one of two ways: 1) checking for differences between two conditions in the discretized time-regions, and 2)

comparing the deviant condition in the local time-split to a global ERP shape. The first will serve to highlight the progress of the cognitive processing without particular focus on a particular ERP which can be especially beneficial in exploratory studies; for instance, the change of coma patients' responses to stimuli through time. The second would be more appropriate in the study of particular ERP and their progress as the same stimuli are presented repeatedly to participants; a study of how the P300 changes under fatigue conditions, for example.

It is important to note that the idea of observing ERPs through a continuous time period is not new. On the contrary, this is the basis on which most current BCI algorithms work (Krusienski et al., 2006, Wolpaw and McFarland, 2004, Blankertz et al., 2004). Other methods were also developed following that approach both for diagnostic purposes and ERP-specific research (see Armanfard et al., 2016, Fallgatter et al., 1997).

4.2.3 Time-series specialization

Current used submodules implement methods that were not developed with time-series signal analysis as a goal application. This limitation was handled by assuming that samples can be treated and analyzed in isolation from others along the time axis of a given EEG trial. Although utilizing this simplification allowed for fast analysis time, there are two main aspects this pipeline fail to address.

Firstly, the current iteration of the method offer no insight on the interactions between components that interact across time. For instance, in a hypothetical case where the N1-P2 complex has strong interactions with a late ERP (400 ms post-stimulus onset as an example), a generated model would only be able to capture the relevance of the earlier component(s) in isolation. A "smarter" method would be able to extract information about

the entire waveform as a whole, and hence, uncovering a more intricate layer of inter-component interactions within conditions.

Stemming from the first point, a main drive for delegating intricate datamining procedures to automated learning machines is for the machine to solve the problem as a whole. An expert in the field employing, in some cases, arbitrary discretization boundaries is by no means an optimal solution but merely one in favor of practicality. This is analogous to issues that arise in traditional EEG/ERP methods where arbitrary windows are taken around ERPs in question as preprocessing in order to perform analyses of variation. An excellent example of how a novel statistical modeling method, namely generative additive models, has managed to offer more information on the data can be found in Meulman et al. (2015).

Two main propositions arise when trying to resolve the aforementioned issues. 1) Application of deep learning on entire trials where the network would be complex enough to capture possible correlations and interactions across both time, and location (electrodes). 2) Abandoning the entire approach of trial segmentation and utilizing time-series-specific algorithms that excel at recognizing trends in continuous data, mostly the long-short-term-memory network (Hochreiter and Schmidhuber, 1997, Davidson et al., 2007). While offering a continuous time interface is beneficial to EEG/ERP analysis, it must be noted that a main criterion for choosing the correct method needs to account for global sequences in time and not just their local shapes; an ideal model should be able to differentiate between two different ERPs with similar shapes and polarities while being given no information about stimulus onset time.

4.2.4 Single-subject grading

The proposed method was created to handle a dataset as a whole with no restrictions or information gathered on smaller entities, aside from two compared conditions. This is an issue when handling datasets with multiple conditions, as discussed in an earlier subsection. More importantly, it does not allow for direct inter-subject comparisons. This problem is two fold: firstly, the current way of measuring component consistency and jitter comes as a percentage value representing the accuracy of discerning two experimental conditions. This does not lend itself to other forms of comparisons where two subjects have both: a lower number of trials for a proper analysis, and a possible variance in latency needing proper statistical constraints to test for significance. The other issue stems from the fact that there is no clear indication of what a specific accuracy truly means aside from comparing it to the baseline's.

For instance, assume a sufficiently large number of trials is present for two subjects per condition in a particular dataset. Given that we know one of the two subjects is a healthy control (HC), how is it possible to compare the non-control (NC) to the HC's results? A trivial state would be if the NC shows no significance peaks, while the HC does. However, complications arise if both subjects show significance but on different accuracy ranges; how much difference would be due to a true effect, and which would be attributed to variance? Moreover, in cases where it is not a subject's component amplitude that differs in accuracy, but the latency at which the accuracy peak manifests, it would also not be clear if the change proves significant. Mainly, collapsing several trials to a couple of datapoints for comparisons needs to be done in properly designed manner that does not discard useful information in the process and provides a clear indication of how different two subjects, or a subject and a population, are.

4.2.5 Domain extension of output units

Due to resource constraints, many research projects are often limited to analysis of data either constrained by either a small number of subjects or few acquisition repetitions. This is often justified due to experimental design choices which limit experiments to strictly cross-section or longitudinal designs. While this is a well-suited approach for self-contained studies, there have been movements towards making datasets freely available on online repositories for further analyses, aggregation into studies with wider scopes, and tackling different hypotheses. A main obstacle when dealing with such a varied group of datasets, is that analysis techniques are at a risk of failing to capture information that generalizes well. This is a primary focus of using traditional statistics to generate inferences that can be replicated given another sample of data. While the field of statistics has defined formal structures for generalization, which also often fail, a pure machine learning approach can not directly be claimed to be the same. Measures of significance captured by the proposed framework can be justifiably proven to show signals that are empirically different. However, a complete dissociation between a difference detected due an underlying brain signal, as opposed to bias caused by recording error for instance, is lacking.

A preliminary step in generalization of the knowledge, is confirming the ability to apply models learned from a set of participants to others which have not been involved in model creation. A direct application of attaining this generalization would come in the form of diagnostic tools which are capable of extracting a participant's brain response to certain types of stimulus and comparing them to a previously collected response dataset (condensed to a learned model).

Generalization is also important for intra-subject variability. For example, a participant's state can greatly affect recorded responses. A subject's level of consciousness, prior

hours of sleep, fatigue, and age are some of these factors. Other facets of variance are session-related. A subject's data can be different across different sessions due to surrounding noise, electrical interference, or different hardware setups. The goal of extending the generalization capabilities is to attain a level of abstraction of the data that is automatically computed and which transforms a given EEG signal to an invariant form, less affected by session/participant -specific features.

Chapter 5

Conclusions

In the present thesis, a semi-automated analysis framework of EEG/ERP datasets has been presented. The method utilizes a modular pipeline centered on machine learning theory and algorithms. Validation approaches are implemented to ensure correctness of provided results, corrected for numerous runs of each submodule on a given dataset. The framework was applied on three EEG/ERP datasets that were recorded as parts of other independent studies. The framework was shown to provide clear indication of the consistency, topography, and latency of each ERP expected in each of the corresponding datasets. The pipeline outlined by the framework provides the aforementioned ERP component features while requiring minimal human-expert intervention. The compelling results reported on each of the discussed datasets demonstrate the capabilities of the analysis scheme; moreover, enhancements to the framework, outlined in chapter four, can be directly used in basic, industrial, and clinical research with both ERP and EEG methodologies.

Bibliography

Alho, K. (1995). Cerebral generators of mismatch negativity (mmn) and its magnetic counterpart (mmnm) elicited by sound changes. *Ear and hearing*, 16(1):38–51.

Archibald, L. and Joanisse, M. F. (2011). Electrophysiological responses to coarticulatory and word level miscues. *Journal of Experimental Psychology: Human Perception and Performance*, 37(4):1275.

Armanfard, N., Komeili, M., Reilly, J. P., Mah, R., and Connolly, J. F. (2016). Automatic and continuous assessment of erps for mismatch negativity detection.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and M??ller, K. R. (2011). Single-trial analysis and classification of ERP components - A tutorial. *NeuroImage*, 56(2):814–825.

Blankertz, B., Müller, K.-r., Curio, G., Vaughan, T. M., Schalk, G., R, J., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schröder, M., and Birbaumer, N. (2004). The BCI Competition 2003:. *Online*, XX:100–106.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international 5.9/10*, pages 341–345.

Browne, M. W. (2000). Cross-Validation Methods. 132:108–132.

Buduma, N. (2014). Deep learning in a nutshell.

Commons, W. (2013). Extrema example original.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process*, 36(94):287–314.

Connolly, J. F. and Phillips, N. A. (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *Journal of Cognitive Neuroscience*, 6(3):256–266.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cowan, N., Winkler, I., Teder, W., and Näätänen, R. (1993). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *Journal of experimental psychology. Learning, memory, and cognition*, 19(4):909–921.

Davidson, P. R., Jones, R. D., and Peiris, M. T. (2007). Eeg-based lapse detection with high temporal resolution. *IEEE Transactions on Biomedical Engineering*, 54(5):832–839.

Desmedt, J. E. and Tomberg, C. (1990). Topographic analysis in brain mapping can be compromised by the average reference. *Brain Topography*, 3(1):35–42.

Dien, J. and Santuzzi, A. M. (2005). Application of Repeated Measures ANOVA to High-Density ERP Datasets: A Review and Tutorial. *Event-Related Potentials. A Methods Handbook*, 4(March):57–82.

Dou, D., Frishkoff, G., Rong, J., Frank, R., Malony, A., and Tucker, D. (2007). Development of NeuroElectroMagnetic ontologies (NEMO): A Framework for Mining Brainwave Ontologies. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, page 270.

Fallgatter, A., Brandeis, D., and Strik, P. W. (1997). A robust assessment of the nogo-anteriorisation of p300 microstates in a cued continuous performance test. *Brain topography*, 9(4):295–302.

Gow, D. W. and McMurray, B. (2007). Word recognition and phonology: The case of english coronal place assimilation. *Papers in laboratory phonology*, 9:173–200.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220.

Kramer, S. (2014). Neural Responses Demonstrate The Dynamicity Of Speech.

Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2006). A comparison of classification techniques for the P300 Speller. *Journal of neural engineering*, 3(4):299–305.

Kutas, M. and Federmeier, K. D. (2014). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *NIH Public Access*, pages 621–647.

Larochelle, H., Larochelle, H., Bengio, Y., Bengio, Y., Lourador, J., Lourador, J., Lamblin,

P., and Lamblin, P. (2009). Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research*, 10:1–40.

Lefebvre, C. D., Marchand, Y., Eskes, G. A., and Connolly, J. F. (2005). Assessment of working memory abilities using an event-related brain potential (ERP)-compatible digit span backward task. *Clinical Neurophysiology*, 116(7):1665–1680.

Lohninger, H. (1999). *Teach/Me data analysis: single user edition; examples from all fields of science, 25 fully interactive applets, 25 animations and slide shows. CD-ROM*. Springer.

Luck, S. J. (2014). *An introduction to the event-related potential technique*.

McQueen, J. M., Norris, D., and Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5):1363.

Meijer, E. H., Smulders, F. T. Y., Merckelbach, H. L. G. J., and Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, 66(3):231–237.

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., and Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLoS One*, pages 1–31.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Morlet, D. and Fischer, C. (2014). MMN and novelty P3 in coma and other altered states of consciousness: A review. *Brain Topography*, 27(4):467–479.

Näätänen, R. (1992). The mismatch negativity. *Attention and brain function*, pages 136–200.

Näätänen, R., Jacobsen, T., and Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. *Psychophysiology*, 42(1):25–32.

Näätänen, R., Paavilainen, P., Titinen, H., Jiang, D., and Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology*, 30(5):436–450.

Näätänen, R. and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Näätänen, R., Simpson, M., and Loveless, N. (1982). Stimulus deviance and evoked potentials. *Biological psychology*, 14(1-2):53–98.

Newman, R. L. and Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom-up and top-down effects on spoken word processing. *Biological Psychology*, 80(1):114–121.

Parvar, H., Sculthorpe-Petley, L., Satel, J., Boshra, R., D?Arcy, R. C., and Trappenberg, T. P. (2015). Detection of event-related potentials in individual subjects using support vector machines. *Brain Informatics*, 2(1):1–12.

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18(1):49–65.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information

criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Perrin, F., Garcıa-Larrea, L., Mauguière, F., and Bastuji, H. (1999). A differential brain response to the subject's own name persists during sleep. *Clinical Neurophysiology*, 110(12):2153–2164.

Perrin, F., Schnakers, C., Schabus, M., Degueldre, C., Goldman, S., Brédart, S., Faymonville, M.-E., Lamy, M., Moonen, G., Luxen, A., Maquet, P., and Laureys, S. (2006). Brain response to one's own name in vegetative state, minimally conscious state, and locked-in syndrome. *Archives of neurology*, 63(4):562–569.

Picton, T. W., Woods, D. L., and Proulx, G. (1978). Human auditory sustained potentials. i. the nature of the response. *Electroencephalography and clinical neurophysiology*, 45(2):186–197.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148.

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., and Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *Journal of Neuroscience*, 33(18):7691–7699.

Rasmussen, C. (2006). Gaussian processes for machine learning. *International journal of neural systems*, 14(2):69–106.

Ravan, M., Hasey, G., Reilly, J. P., MacCrimmon, D., and Khodayari-Rostamabad, A. (2015). A machine learning approach using auditory odd-ball responses to investigate the effect of Clozapine therapy. *Clinical Neurophysiology*, 126(4):721–730.

Sculthorpe-Petley, L., Liu, C., Ghosh Hajra, S., Parvar, H., Satel, J., Trappenberg, T. P., Boshra, R., and D'Arcy, R. C. N. (2015). A rapid event-related potential (ERP) method for point-of-care evaluation of brain function: Development of the Halifax Consciousness Scanner. *Journal of Neuroscience Methods*, 245:64–72.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958.

Teplan, M. (2002). Fundamentals of EEG measurement. *Measurement Science Review*, 2:1–11.

Todd, J., Michie, P. T., and Jablensky, A. V. (2003). Association between reduced duration mismatch negativity (MMN) and raised temporal discrimination thresholds in schizophrenia. *Clinical Neurophysiology*, 114(11):2061–2070.

Todd, J., Michie, P. T., Schall, U., Karayanidis, F., Yabe, H., and Näätänen, R. (2008). Deviant Matters: Duration, Frequency, and Intensity Deviants Reveal Different Patterns of Mismatch Negativity Reduction in Early and Late Schizophrenia. *Biological Psychiatry*, 63(1):58–64.

Wolpaw, J. R. and McFarland, D. J. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17849–17854.