LOGISTIC REGRESSION MODELS: FEATURE SELECTION FOR P300 DETECTION IMPROVEMENT

Sadraque S. Viana*, Danilo M. Batista* and Danilo B. Melges**

* Undergraduate Program in Control and Automation Engineering, - UFMG, Belo Horizonte, Brazil
** Graduate Program in Electrical Engineering - UFMG, Belo Horizonte, Brazil

e-mail: viana.sadraque@gmail.com

Abstract: *Brain computer interfaces are devices that enable people to communicate using their brain activity, being useful for those suffering with neurodegenerative diseases. Many BCI use the P300 evoked response. However, the detection of this potential is very difficult, since it presents low signal-to-noise ratio. This paper studies which time features present in the electroencephalogram can improve the detection rates of this response. For this purpose, logistic regression models were used to assess the significance of the following features for P300 representation: most positive peak, most negative peak, latencies of these peaks, area and RMS value between them. These parameters were evaluated on averages of 600 ms windows that could or not contain the potential. The results showed that positive peak, RMS value and both latencies significantly improve P300 detection rates.*

Keywords: *P300, brain-computer interface, feature selection, logistical regression.*

Introduction

Brain computer interfaces (BCI) are defined as communication systems that a person can use to send messages or commands through alternative pathways different from those offered by muscles and peripheral nerves [1].

BCI are possible solutions for people that suffer with neurodegenerative diseases such as: amyotrophic lateral sclerosis, brainstem stroke, brain or spinal cord injury, cerebral palsy and others [1]. Especially for severally disabled subjects, BCI can enhance autonomy and promote improvements on life quality. Recent applications of BCI include prosthesis control, motorized wheelchairs and home automation [2].

Many BCI are based on the P300 evoked response potential (ERP) [3] (Figure 1), which is characterized by a positive voltage peak ranging from 2 to 5μV at about 300ms to 600ms post-stimulus and can be recorded at the scalp by means of electroencephalogram (EEG). It appears with better signal-to-noise ratio (SNR) at Fz, Cz and Pz derivations [4] and arises as a response to visual, somatosensory or auditory stimuli, when a significant target stimulus is presented infrequently among non-target stimuli, in a random way.

When compared to other assistive technology devices, such as those that use muscular responses like eye movement, P300 based BCI present lower communication rates. However they are a useful alternative, mainly when the stage of motor disability makes the use of devices based on biosignals with better SNR impossible [5].

Detecting P300 is a big challenge, since its pattern is masked by spontaneous EEG. An important step for detecting patterns and classifying data is to perform feature extraction and selection [6].

For this reason, many P300 time features [7, 8, 9] have been used in order to detect which EEG signal averages (coherent means) contain the P300, such as those listed in Table 1.

However, few works focused on identifying the best features to distinguish between the two average classes: P300 and spontaneous EEG average (target and non-target).

Hence, the objective of this work is identifying time features that significantly improve the probability of P300 detection. This goal was accomplished by performing statistical tests on the coefficients of logistic regression models. This strategy can be used to select which features should be used as inputs of classification algorithms in future works.

Modeling Binary Response Variables

Logistic Regression Models – The response variable considered can assume only two classes: target and non-target. To describe the behavior of this kind of variable, logistic regression models are used, whose response function is defined as [10]:

$$E(y_i) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{-x_i\beta}}, \qquad (1)$$

where $E(y_i)$ is the expected value of the response variable y_i, x_i is the explanatory variable (feature), β is the coefficient associated with the feature.

Adding multiple explanatory variables, Equation 1 becomes:

$$E(y_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}} \qquad (2)$$

Equation 2 models the relationship between y_i and the explanatory variables x_i. The expected value can be used as a probability score in order to predict the outcome of the response variable [11].

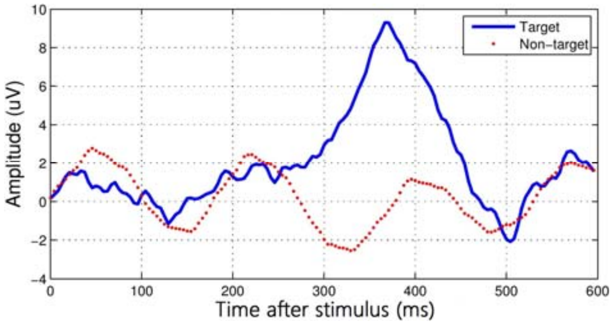

Figure 1: P300 signal average.

Table 1: Features of interest

Feature	Description
Most Positive Peak (PPeak)	The most positive peak in the average.
Most Negative Peak (NPeak)	The most negative peak in the average.
Area Under Peaks (Area)	Value of the area under the curve defined from NPeak to PPeak.
RMS Value	RMS for the curve from NPeak to PPeak.
Positive Peak Latency (PLat)	Factor that assumes values from 1 to 6 relative to different ranges of time occurrence of PPeak in the average. Thus, 1corresponds to 0...100ms, 2 to 100…200ms, and so on.
Negative Peak Latency (NLat)	Idem PLat description, but for the NPeak.

Tests on Individual Model Coefficients – With the model described by Equation 2, one may want to find out the statistical significance of each coefficient β_k, leading to the following hypothesis:

$$H_0: \beta_k = 0$$
$$H_1: \beta_k \neq 0 \quad (3)$$

To test the above hypothesis, a t-like statistic called the Wald statistic can be derived [8]:

$$Z_0 = \frac{\beta_k}{se(\beta_k)}, \quad (4)$$

which has a standard normal reference distribution.

It can be seen that, as the standard error grows, Z_0 will tend to zero. So, a coefficient must have a low standard error (in comparison to its absolute value) in order to be statistically significant.

Thus, tests on Z_0 may be used to assess the statistical significance of each feature, providing information on how it contributes to the ability of the model to distinguish between classes. The combination of features was tested by multiplying the corresponding model coefficients.

Methodology

Data Collection – The experiment was executed using the database provided by [12], with data from 24 healthy volunteers, 16 men and 8 women, aged between 19 and 25 years.

The EEG was recorded with the reference on the right earlobe and with ground lead at the right mastoid. Signals were band-filtered between 0.1Hz and 60Hz, and sampled at 256Hz.

For each volunteer, 21 symbols were presented, using the oddball paradigm [9] for P300 elicitation. The symbols are disposed in a 6x6 matrix (Figure 2). Each row and each column is intensified during 100 ms and appears blank for 75 ms, with the order of intensification being randomized. When a row or column that contains the desired symbol is intensified, P300 elicitation is expected to occur [5, 7]. As such, for each desired symbol there are twelve averages of rows and columns: two of them contain the P300 (target stimuli) and ten are spontaneous EEG (non-target stimuli).

Two averages of target and two of non-target stimuli (picked up randomly) from Cz derivation were selected for each desired symbol. The features listed in Table 1 were extracted from these averages.

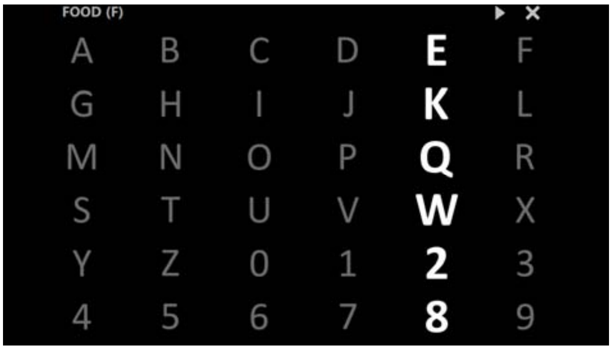

Figure 2: Row-column paradigm for P300 elicitation.

Experimental Design – A logistic regression model was used to predict the probability that the P300 occurred in the selected average window. This approach allowed identifying if changes on feature (Table 1) values would improve the detection rates of P300.

If a given feature does not significantly alter the probability of P300 occurrence in the average, the corresponding model's coefficient will fail its Wald statistic tests. Therefore, the hypotheses defined for the experiment are those described by Equation 3, which were tested with a confidence interval of 0.01.

The generated model was also used to analyze how the interaction between features changes the detection rates of P300. This was accomplished by fixing the values of all but one of the features and then predicting how the probability of P300 detection changes accordingly to the feature that was not fixed.

Results

The statistical test showed that PPeak ($p<0.007$), RMS value ($p<0.003$), PLat4 ($p<0.001$), PLat6 ($p<0.003$) and NLat4 ($p<0.002$) are significant with a 0.01 confidence interval, when each parameter is considered individually.

Figure 3 shows the influence of PPeak on P300 detection rates. The solid lines represent the probability scores as the value of this feature varies, for each level of PLat and NLat. The translucent shadows around each line are the confidence intervals of probability scores. Hence, the narrower the shadow (smaller confidence interval), the more significant is the feature.

PPeak, for example, is a good predictor of P300 occurrence, presenting a significant interaction ($p<0.001$) for PLat = 3, 4, 5 and 6 (latency range from 200 to 500 ms). This is evidenced by the fact that these curves (Figure 3, top) present high detection probability scores with narrow confidence intervals for PPeak values over 7 μV.

The interaction with NLat (Figure 3, bottom) was significant ($p<0.001$) only when it occurred in the interval between 300 and 400 ms, which corresponds to NLat = 4.

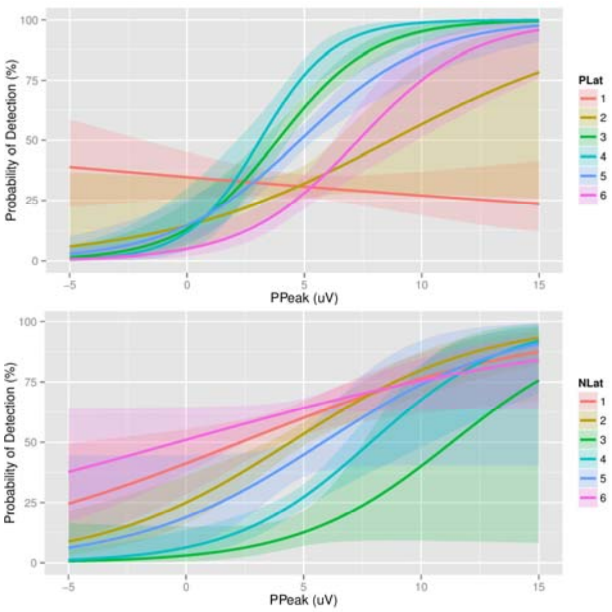

Figure 3: Probability of P300 detection when PPeak varies for different ranges of PLat (top) and NLat (bottom).

NPeak (Figure 4, top) has a strong interaction ($p < 0{,}001$) for PLat = 3, 4 and 5, since for values under -10 μV, detection probability scores over 75% are achieved.

However, the interaction between NPeak and NLat was not significant (Figure 4, bottom).

The Area presented no significance when considered individually. However, a significant interaction ($p < 0.01$) was found for PLat = 3 and 4 (Figure 5, top). No significant interaction with NLat was found (Figure 5, bottom).

The interaction between the RMS value and the Positive Latency was significant ($p < 0.001$) for PLat =3 and 4 (Figure 6, top), but the interaction with NLat was not significant (Figure 6, bottom).

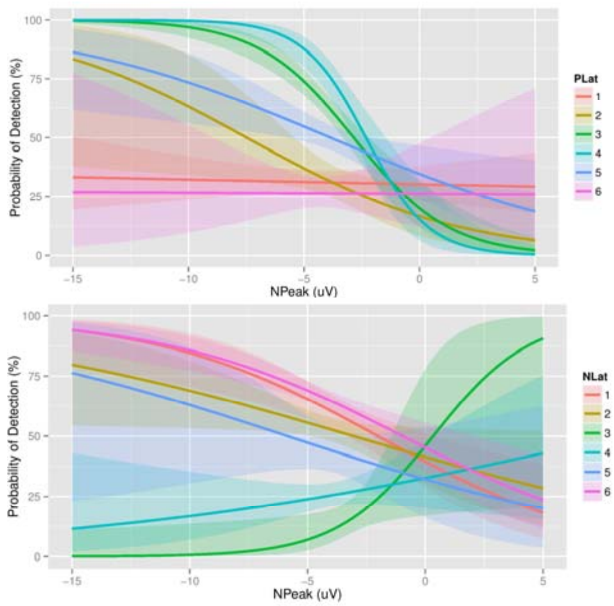

Figure 4: Probability of P300 detection when NPeak varies for different ranges of PLat (top) and NLat (bottom).

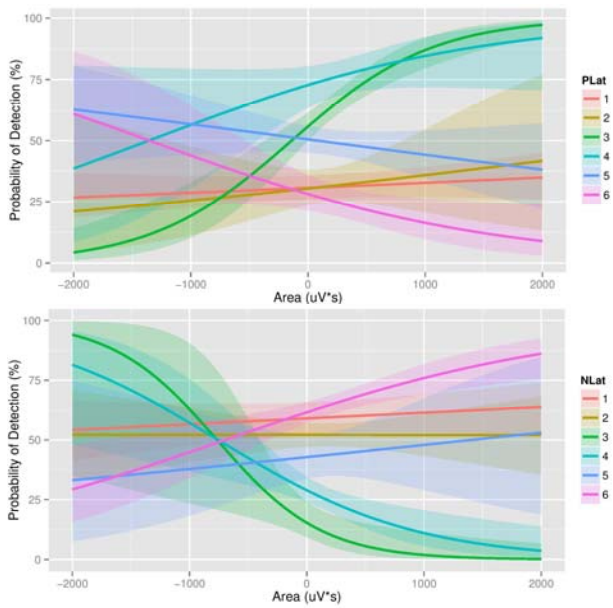

Figure 5: Probability of P300 detection when Area varies for different ranges of PLat (top) and NLat (bottom).

Discussion

The P300 is primarily characterized by a positive peak [4], which is reflected by the significance of PPeak as a predictor of its occurrence. This result corroborates with the use of this feature in other works [7, 9].

Because it measures the magnitude of a varying quantity, the RMS value may reveal that target stimuli presents greater spread from zero than non-target ones.

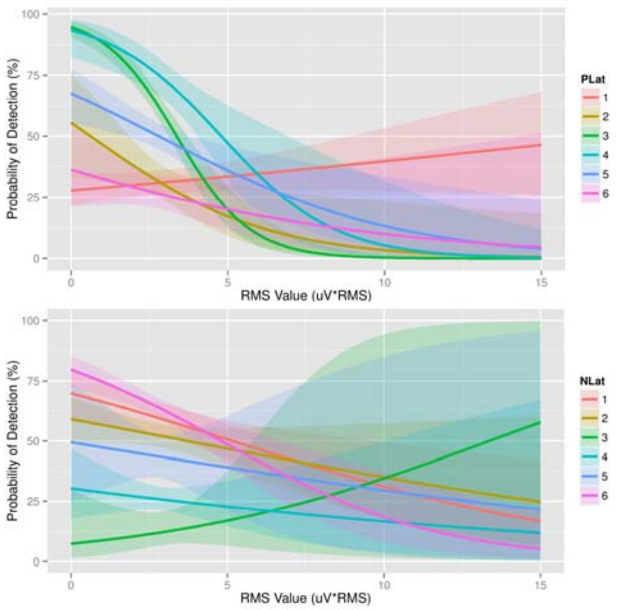

Figure 6: Probability of P300 detection when RMS varies for different ranges of PLat (top) and NLat (bottom).

The significance found on this feature shows that it may improve the distinction between the two classes, and can be used as an input of classification algorithms.

Area calculation sums the parts above and subtracts the parts below zero. As P300 (Figure 1) presents positive and negative peaks for both classes, the area calculation can tend to low values for both target and non-target averages.

This feature has been previously used in other works [7, 8], but our results suggest it may not be suitable for differentiating between P300 and spontaneous EEG.

P300 doesn't present a prominent negative peak, so this feature alone may not be representative of target stimuli, justifying its lack of significance.

Besides the amplitude of the positive peak, P300 is also characterized by the latency at which this peak occurs, at about 300 ms post-stimulus [4].

The interaction between all features with PLat and NLat provides information about the relationship between magnitude and timing of signals. Its significance shows that the use of spatiotemporal filters, such as reported by [2], is a good way to differentiate between classes.

Although Area and NPeak may not aid to identify target stimuli, they can help to identify non-target ones. This may justify the significance of the interaction between these features with PLat and NLat.

Conclusion

In this paper, logistic regression models were used to identify which time features improve P300 detection rates. This was carried out by performing statistical tests on the model's coefficients. The value of the most positive peak and the RMS value between peaks were found to be good predictors of P300 occurrence, which can be justified by the differences between P300 and spontaneous EEG waveforms.

The interaction between factors was also found to significantly improve the detection rates of P300. This may be justified because they provide information on the relationship between amplitude and time of potentials, helping to differentiate between the classes.

Finally, we conclude that logistic regression models can be used to perform feature selection for P300 detection improvement.

References

[1] Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G, Vaughan T. Brain-computer interfaces for communication and control. Clinical Neurophysiology. 2002; 113(6):167–791.

[2] Sellers E, Krusienski D, McFarland D, Vaughan T, Wolpaw Jr. A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance. Biological Psychology. 2006; 73(3): 242–252.

[3] Bashashati A, Fatourechi M, Ward RK, Birch GE. A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. Journal of Neural Engineering. 2007; 4(2):R32.

[4] Polich J. Updating p300: An integrative theory of P3a and P3b. Clinical Neurophysiology. 2007; 118(1): 2128 – 2148.

[5] McFarland DJ, Sarnacki WA, Townsend G, Vaughan T, Wolpaw JR. The P300-based brain computer interface (BCI): Effects of stimulus rate. Clinical Neurophysiology. 2011, 122(4): 731-737.

[6] Guyon I, Gunn S, Nikravesh M, Zadeh L. Feature Extraction, Foundations and Applications. 1st ed: Springer; 2006.

[7] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology. 1988; 70(1): 510–523.

[8] Allison BZ, Pineda JA. ERPs evoked by different matrix sizes: implications for a brain computer interface (BCI) system. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2003; 11(2): 110-113.

[9] Kaper M, Ritter H. Generalizing to new subjects in brain–computer interfacing. In: Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; San Francisco, United States; 2010. p 4363–4369.

[10] Myers RH, Montgomery DC, Vining GG, Robinson TJ. Generalized linear models: with applications in engineering and the sciences. 1st ed: John Wiley & Sons; 2012.

[11] Bhandari M, Joensson A. Clinical research for surgeons. 1st ed: Thieme; 2009.

[12] Ledesma-Ramirez C, Bojorges-Valdez E, Yanez-Suarez E, Saavedra C, Bougrain L, Gentiletti GG et al. An open-access P300 speller database. In: 4th International BCI Meeting; 2010.