Towards the detection of Error-Related Potentials and its integration in the context of a P300 Speller Brain-Computer Interface

A. Combaz[a], N. Chumerin[a], N.V. Manyakov[a], A. Robben[a], J.A.K. Suykens[b], M.M. Van Hulle[a]

[a] *K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie, Herestraat 49, B-3000 Leuven, Belgium*
[b] *K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium*

Abstract

A P300 Speller is a Brain-Computer Interface (BCI) that enables subjects to spell text on a computer screen by detecting P300 Event-Related Potentials in their electroencephalograms (EEG). This BCI application is of particular interest to disabled patients who have lost all means of verbal and motor communication. Error-related Potentials (ErrPs) in the EEG are generated by the subject's perception of an error. We report on the possibility of using these ErrPs for improving the performance of a P300 Speller. Overall 9 subjects were tested, allowing us to study their EEG responses to correct and incorrect performances of the BCI, compare our findings to previous studies, explore the possibility of detecting ErrPs and discuss the integration of ErrP classifiers into the P300 Speller system.

1. Introduction

Brain Computer Interfaces (BCIs) are aimed at creating a direct communication pathway between the brain and an external device, bypassing the need for an embodiment. In the last few years, research in the field of BCI has witnessed an important development (see [1, 2]) and it is nowadays regarded as a very promising application of neuroscience. Indeed, such systems can provide a significant improvement of the quality of life of neurologically impaired patients suffering from pathologies such as amyotrophic lateral sclerosis, brain stroke, brain/spinal cord injury, etc.

In invasive BCIs, a micro-electrode array is implanted in the brain (mainly in the motor or premotor frontal areas [3] or into the parietal cortex [4]), while in non-invasive BCIs, mostly *electroencephalograms* (EEGs) are recorded from the scalp. There are several types of EEG-based BCIs. For example, some are based on *Steady State Visually Evoked Potentials* (SSVEPs, [5]): they work by detecting the activity of the brain at a specific frequency corresponding to the flickering frequency of a visual stimulus (see [6, 7] for applications). Another type of BCIs relies on the detection of mental tasks (imagination of right/left hand movements, calculation, word association, etc), which are detected through *Slow Cortical Potentials* [8], *Readiness Potential* [9] and *Event-Related Desynchronization* [10].

The BCI presented here belongs to another category; it is based on the detection of the *P300 Event-Related Potential* (ERP: a stereotyped electrophysiological response to an internal or external stimulus, [11]). This brain potential is elicited in the context of an oddball paradigm: when a subject perceives two types of events, one of which occurs only rarely, the rare event will elicit in the EEG an ERP with an enhanced positive-going component at a latency of about 300 ms (the P300 ERP, [12]).

The first spelling system based on the detection of this ERP was introduced in 1988 by Farwell and Donchin [13]. This application is nowadays one of the most studied BCIs and the work presented here deals with a similar system. The P300 Speller allows subjects to spell words by focusing on the desired characters shown in a matrix while the rows and columns of the matrix are consecutively and randomly intensified (Fig. 1-left). The intensification of a row or column containing the target symbol will elicit a P300 ERP and, by detecting this ERP, the BCI is able to identify the target row and column and thus to retrieve the symbol the subject has in mind.

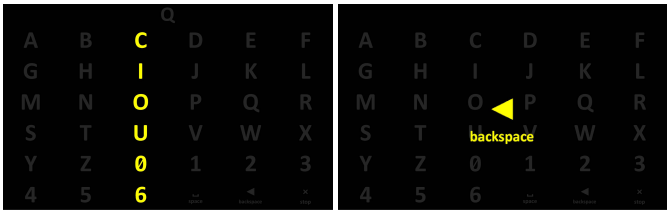


Figure 1: User display for the P300 Speller BCI. Left: intensification of a column of the matrix display. Right: Feedback moment (the identified symbol is displayed on the screen).

Ideally, one sequence of intensifications of each row and column would be enough to identify the target symbol. Unfortunately, the low signal-to-noise ratio of the P300 ERP makes it almost undetectable in a single trial. It is therefore common practice to repeat several times the sequence of intensifications, in order to average the EEG responses and increase the signal-to-noise ratio. Depending on the number of repetitions, this approach can lead to a dramatic increase of the time needed to communicate a symbol. It is thus important to work on robust and efficient feature extraction and classification techniques to reduce this number of repetitions.

An elegant way to improve the performance of a BCI is the detection of the so-called *Error-related Potentials* (ErrPs).

ErrPs were suggested to be generated in the anterior cingulate cortex with a spatial distribution over the fronto-central regions of the scalp, and related to the subject's perception of an error [14, 15]. If the first studies on the presence of an ErrP in the EEG were dealing with brain responses to errors made by the subject himself [16, 17], more recent work discusses the presence of such potential in the context of a BCI, when the user realizes that the interface failed to recognize properly his intention [15, 18–21]. This latter phenomenon is what we will refer to as ErrP in the article. In [19], ErrPs were observed in the context of a vertical cursor controlled with *mu*/*beta* waves, while in [18] it was in the context of a simulated BCI, where the subject manually delivers commands to move a horizontal cursor. This experiment was successfully reproduced in a situation where the BCI was still simulated with an *a priori* error rate but this time the subjects were performing movement imagination [15]. To our knowledge only researchers from the Politecnico di Milano University [20, 21] recently presented some work on the error potential in the context of a P300 Speller.

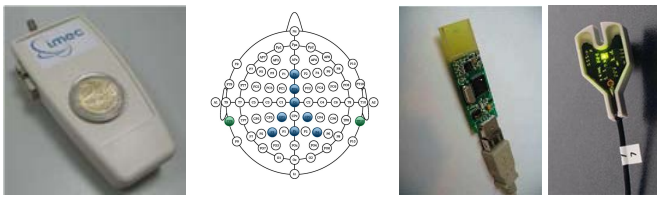
This paper is an extension of the work presented in [22] and reports on a study performed in our laboratory where 2 series of experiments were conducted. In the first one, 6 subjects were tested on the P300 Speller developed by our group [22–24], and their EEG responses to correct and incorrect feedback (*i.e.*, the moment when the BCI displays what it identifies as the target symbol, see Fig. 1-right) were recorded. Each subject performed one session of maximum 2 hours during which he/she used the system to type several words of his/her choice. For the second series of experiments, 3 new subjects were recruited and the same recording sessions as just described were repeated 6 to 7 times over a maximum period of 2 weeks for each one of them. The data from the first series of experiments allowed us to observe the ErrP elicited in the EEG of the participants and to compare our observations with the ones reported in the studies previously mentioned. The second series of experiments allowed us to study the possibility of detecting the ErrP by building several classifiers using training data and measuring their accuracy on test data. We finally discuss the possibility and interest of including an ErrP detection tool into the P300 Speller.

2. Data acquisition

2.1. Material

The EEG recordings were performed using a prototype of an ultra low-power 8-channels wireless EEG system, which consists of two parts: an amplifier coupled with a wireless transmitter and a USB stick receiver (Figs. 2a, 2c). The data are transmitted with a sampling frequency of 1 kHz for each channel. The prototype was developed by imec [25]. We used a braincap with large filling holes and sockets for active Ag/AgCl electrodes (ActiCap, Brain Products, Fig. 2d).

The recordings were collected with eight electrodes placed over the frontal, central and parietal areas of the brain, namely in positions *Fz*, *FCz*, *Cz*, *CP1*, *CP2*, *P3*, *Pz* and *P4* according to the international 10-20 system (Fig. 2b). The reference and ground electrodes were positioned on the left and right mastoids, respectively (*TP9*, *TP10*). The recording sites were the same as in [22] (see also [26, 27] for some guidelines on how to choose the recording sites).



(a) (b) (c) (d)

Figure 2: (a) Wireless 8 channels amplifier. (b) Locations of the electrodes on the scalp. (c) USB stick receiver. (d) Active electrode.

The visual stimulation consisted of a matrix of 6-by-6 symbols (Fig. 1). For both the training and testing stages, each sequence of intensifications consisted in the highlighting of each row and column of the matrix only once and in random order. Each highlighting lasted for 100 ms, followed by 100 ms of no intensification. All recordings and stimulation were performed with MATLAB, the display of the stimuli and their precise timing was achieved using the *Psychophysics Toolbox Extensions* [28, 29].

2.2. Recording session

Each recording session lasted between one and two hours, and maximum effort was made to keep the subject fully concentrated; the experiments were stopped when the participants started to feel tired.

The first step of the experiment was to familiarize the subject with the P300 Speller BCI and to train the system to recognize the P300 ERP. Hence, *prior* to any "mind-spelling", we performed a training session during which the participants were asked to focus consecutively on 8 symbols randomly selected by the interface. An indication of the symbol to focus on was first presented to the subject, then the random sequence of intensifications of all the rows and columns was repeated 10 times and, finally, the target symbol was presented to the subject in the middle of the screen for 2 seconds (feedback moment, Fig. 1-right). This was repeated for all 8 symbols.

Based on the data recorded during the training session, we built a classifier for the detection of the P300 ERP. The signals were beforehand filtered between 0.3 and 15 Hz (zero-phase 3rdorder Butterworth filter), and then cut into 800 ms epochs starting from the stimuli onsets. Those epochs were then average-downsampled to 80 data points (each new data point corresponds to the average of the signal over a 10 ms window) and finally, the data of the same classes were averaged over the desired number of trials (corresponding to the desired number of repetitions of the sequence of intensification for the spelling mode).

For each trial (stimulus), we thus have 8 channels $\times$ 80 data points = 640 features to classify as a response to either a target stimulus or a non-target stimulus. A linear *Support Vector Machine* (SVM, [30, 31]) with a 10-fold cross-validation and a linesearch for the optimization of the regularization parameter was built from those training features. Training the linear

Table 1: Details of the performances of each participant to the first series of experiments.

Subject	Gender	Age	Number of words typed	Total number of typed symbols	Total number of mistyped symbols (%)
S1	M	24	5	32	6 (19%)
S2	F	23	7	65	10 (15%)
S3	M	34	5	37	7 (19%)
S4	M	27	7	59	16 (27%)
S5	F	22	9	60	13 (22%)
S6	M	29	7	56	19 (34%)

Table 2: Details of the performances of each participant to the second series of experiments.

Subject	Gender	Age	Total number of typed symbols	Total number of mistyped symbols (%)	Session	Number of typed symbols	Number of mistyped symbols (%)
S7	M	27	659	171 (26%)	1	83	23 (28%)
					2	98	25 (26%)
					3	100	35 (35%)
					4	128	35 (27%)
					5	124	29 (23%)
					6	126	24 (19%)
S8	F	24	963	114 (12%)	1	93	12 (13%)
					2	146	24 (16%)
					3	177	12 (7%)
					4	110	14 (13%)
					5	154	17 (11%)
					6	139	21 (15%)
					7	144	14 (10%)
S9	F	24	758	121 (16%)	1	128	12 (9%)
					2	134	21 (16%)
					3	92	22 (24%)
					4	140	21 (15%)
					5	133	21 (16%)
					6	131	24 (18%)

SVM on 2000 feature vectors with the modified finite Newton method proposed in [32] typically took around one minute.

In the second step of the experiment, the subjects used the P300 Speller with the previously built classifier. This classifier was applied online to the data in order to detect the P300 ERP and identify the target symbols. They would first use the system with 10 repetitions of the sequence of intensifications, in order to make them confident about the accuracy of the system. Most of them spelled their first word with no mistake. As the aim was to record EEG responses to erroneous feedback, we then reduced this number of repetition to 5, 4 and even to 3, depending on how accurately the subjects were typing.

2.3. Experiment design

As mentioned earlier, two series of experiments were performed. For the first one, 6 healthy subjects (4 male, 2 female, age 22–34, 5 right handed and 1 left handed) were recruited. They all performed one session during which they spelled between 32 and 65 symbols with a number of errors comprised between 6 and 19 (see Table 1).

For the second series of experiments, three new subject were tested (2 female, 1 male, age 24–27, 2 right handed and 1 left handed); they performed between 6 and 7 sessions. The detail of those sessions are presented in Table 2.

We will refer to the subjects as S1-S6 for the ones participating in the first study and S7-S9 for the ones who engaged in the second study.

3. First study: presence of an ErrP

This section reports on the results obtained from the first series of experiments; we present here the average EEG responses to correct and incorrect feedback, compare our observations with results from recent studies, and assess of the statistical significance of the difference between those two type of feedback responses.

3.1. The shape of the ErrP

The averaged EEG responses to correct and incorrect feedback for each subject at electrode *FCz* and the grand average over all subjects for each electrode are plotted in Figs. 3 and 4 (the signal were filtered between 0.5 and 15 Hz with a zero-phase 3rdorder Butterworth filter). We also plotted the *error-minus-correct* difference potentials (difference between averaged responses to erroneous feedback and averaged responses to correct feedback).

In [19], Schalk et al. observed a *error-minus-correct* difference consisting of a positive potential that peaked about 180 ms followed by a negative potential (4 subjects were tested). In [15] and [18], this difference was characterized by a first positive peak at 200 ms after the feedback, followed by two larger negative and positive peaks at about 250 ms and 320 ms and a wider negative peak at 450 ms after the feedback (5 subjects were tested). Finally in [20] and [21] (2 and 5 subjects tested, respectively), this *error-minus-correct* difference showed a negative peak occurring at about 300 ms followed by a positive peak at around 400 ms after the feedback.

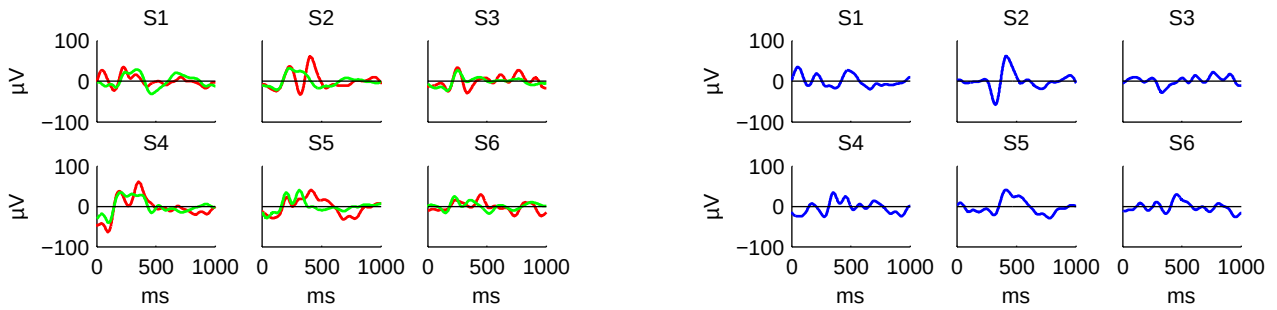


Figure 3: EEG responses for each subject at electrode location FCz for 1 second from the feedback onset. Left: EEG responses averaged over all the correct (green) and erroneous (red) feedbacks. Right: averaged *error-minus-correct*. Units are ms for the x-axes and µV on the y-axes

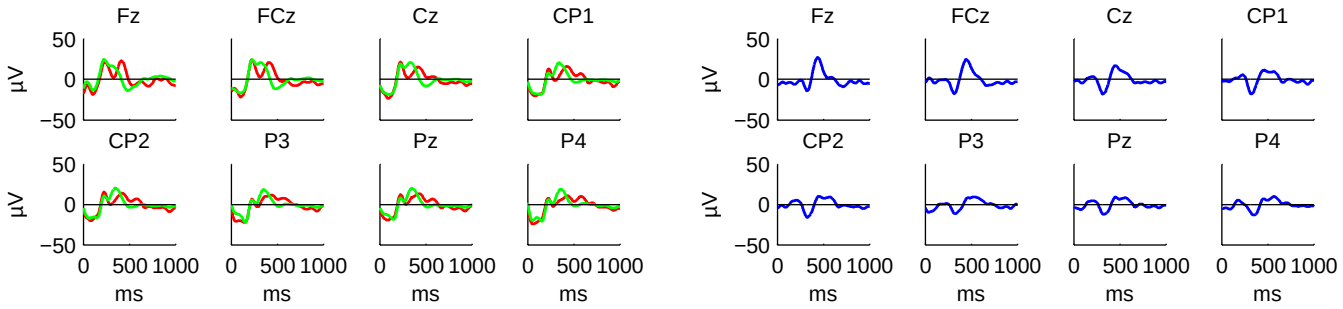


Figure 4: EEG responses averaged over all subjects at each electrode location for 1 second from the feedback onset. Left: EEG responses averaged over all the correct (green) and erroneous (red) feedbacks. Right: averaged *error-minus-correct*. Units are ms for the x-axes and µV on the y-axes

All three studies show quite different results concerning the shape of the ErrP. As they all involve different tasks for controlling the BCI, this could explain the differences. In [19], the subjects were first trained to control their *mu* and *beta* waves so as to later on use them to manipulate a cursor. In [15, 18], the subjects were also controlling a cursor through a manual command in [18] and a motor imagination paradigm in [15] (imagination of left hand and right foot movements). Lastly, in [20, 21], the subject's task was, as in our case, to count the number of occurrences of a target stimulus.

Another potentially influential parameter could be the nature of the feedback: the brain might respond differently when reaching (or not) a target with a cursor [19] than when moving the cursor towards (or away from) the target [15, 18] or to the display of a (non) desired symbol [20, 21]. If a cursor moving task could involve the motor area of the brain, a spelling task might involve language related cognitive processes, and this could lead to a different error feedback processing.

Moreover, each study uses a different time line and presentation mode for the display of the feedback: in [19] the feedback is presented after the subject performed the task (7–8 seconds) and consists of 3 flashes of the detected target within 3 seconds; in [15, 18] the feedback correspond to the movement of the cursor every 2 seconds and in [20, 21] the target symbol is presented for 2 to 3 seconds after 15 seconds of stimulation and 1 second of pause. The importance given to the feedback by the subject as well as the frustration that an error would generate could influence the shape of the recorded EEG response and might depend on how often the feedback is presented and on whether it corresponds to the achievement of the task (detecting a symbol, reaching a target [19–21]) or just one step of the task (moving towards a target [15, 18]).

In our case, when looking at the grand average *error-minus-correct* (Fig. 4-right), we can observe a negative peak followed by positive one at about 320 ms and 450 ms respectively. Those peaks are most prominent at the electrode sites *Fz* and *FCz*. These results are in concordance with [20, 21] where a similar P300 Speller as the one presented here was used.

3.2. Statistical significance

In order to assess the significance of the difference between responses to erroneous feedback and responses to correct feedback, we analyzed the data of each subject at the electrode location *FCz*. We first "average-downsampled" the signals from 1000 Hz to 100 Hz. Then, for each time step $i = 1, ..., N$, and for all M trials of a given subject, we calculated the coefficient of determination $R(i)^2$ (square of the correlation coefficient, [33]) indicating the fraction of the total variance of the EEG feedback responses x_{ki}, that was explained by the class y_k of the corresponding trial k (correct feedback versus erroneous feedback):

$$R(i)^2 = \frac{\mathrm{cov}(X_i, Y)^2}{\mathrm{var}(X_i)\,\mathrm{var}(Y)},\ i = 1, \ldots, N$$
$$\text{with} \begin{cases} X_i = \{x_{ki},\ k = 1, \ldots, M\} \\ Y = \{y_k \in \{-1, 1\}, k = 1, \ldots, M\} \end{cases} \quad (1)$$

These values are plotted in Fig. 5 (left). While due to the low signal-to-noise ratio of the EEG signals, the values of this coefficient of determination remain quite low, we can still observe

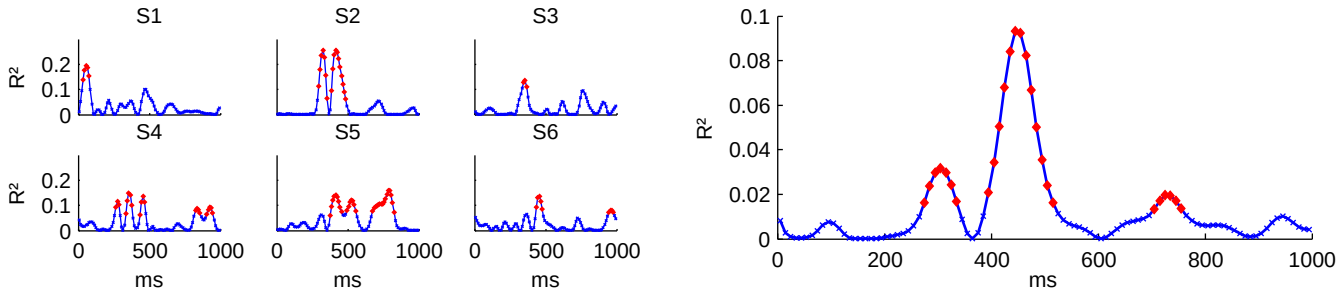


Figure 5: Coefficient of determination versus time after feedback onset for each subject independently (left) and for all subjects together (right). The red diamonds indicates the values for which the permutation test gives a p-value lower than 0.05. Units are ms for the x-axes and the y-axes represent the value of R^2.

some peaks along the time line. Some of those peaks have the same latency time as the negative and positive peaks that we accounted for as ErrP in the EEG feedback responses. To have an idea about how significant those peaks are with respect to the labeling as ErrP and non-ErrP, we performed a permutation test at each time point (significance level 0.05, [34]). For most subjects (except subject S1), the time intervals corresponding to at least one of the 2 peaks associated with the ErrP were statistically significant (red diamonds on Fig. 5-left). The same study was performed by regrouping the data from all subjects together and the coefficient of determination for both time intervals appeared statistically significant (Fig. 5-right).

If those results suggest an apparent discriminability between EEG responses to both kinds of feedback, the high variability between the responses of the same type among trials and participants indicates the necessity of training a classifier to recognize the ErrP for each subject. In the case of a P300 Speller, this can be problematic due to the long time needed to acquire a sufficient amount of training and testing data to build and assess the accuracy of the classifier. And, as shown by the comparison of the results from [15, 18–21], the shape of the ErrP seems to be closely related to the type of paradigm used for the BCI. Thus, we should collect the training data in the exact context in which we want to detect the ErrP.

4. Second study: classifying the feedback responses

This section reports on the results obtained from the second series of experiments. We aim here at gaining insight into the possibility of correctly classifying the EEG responses of a particular subject as ErrP (incorrect feedback) and non-ErrP (correct feedback). We are also interested in the amount of training data required to reach an acceptable accuracy. for this purpose, we performed, with 3 new subjects, 6 to 7 recording sessions similar to the ones presented in the previous section. We could thus gather for each subject an amount of EEG responses to erroneous feedback large enough to first confirm the shape and statistical significance of the *error-minus-correct* presented in the previous section, and then to carry on with an attempt to correctly classify our EEG data.

4.1. ErrP Shape and statistical significance

In order to confirm the observations of our first study, we performed the same analysis on the new data. The EEG responses

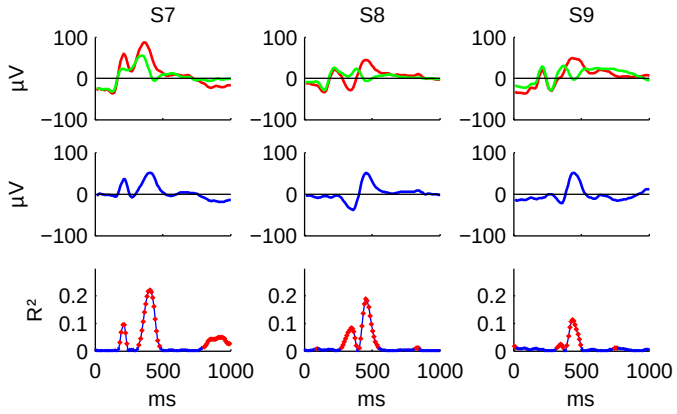


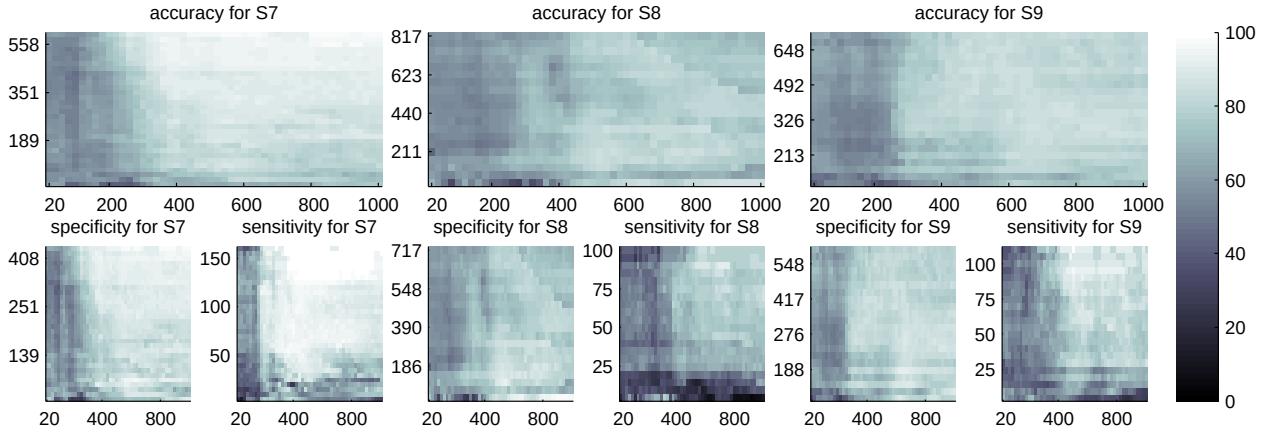
Figure 6: EEG responses for each subject at electrode location FCz for 1 second from the feedback onset. Top: EEG responses averaged over all the correct (green) and erroneous (red) feedbacks. Center: *error-minus-correct*. Bottom: Coefficient of determination versus time after feedback onset for each subject. The red diamonds indicates the values for which the permutation test gives a p-value lower than 0.005.

to feedback, the *error-minus-correct* and the correlation coefficients with their statistical significance are shown in Fig. 6 for each of the three subjects for the electrode site *FCz*.

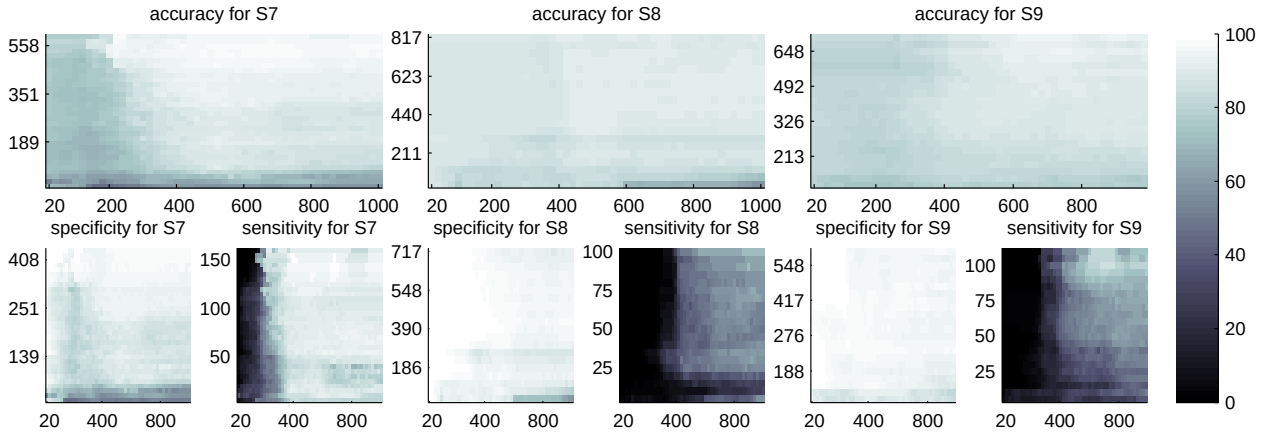
While, for the first study, the variability in the shape of the EEG responses to feedback across subjects could be explained by the low signal to noise ratio of the EEG combined with the low number of responses to erroneous feedback, the data here are in a sufficient amount to rule out such an explanation.

When looking at Fig. 6, the common feature which we can observe for responses to correct and incorrect feedback for all 3 participants is that those responses are composed of 2 positive peaks appearing respectively, between 200 and 220 ms and between 340 and 450 ms after the feedback onset. For all 3 subjects, the second peak is higher, broader and appears later in the case of a response to an incorrect feedback than in the case of a response to a correct feedback; this results in the positive peak that we can observe in the *error-minus-correct* between 410 and 460 ms after the feedback onset for all 3 subjects.

The negative peak observed in the *error-minus-correct* at around 350 ms for subjects S8 and S9 is due to the difference in latency of this second peak between the responses to correct and erroneous feedback and, in the case of subject S8, also to



(a) Accuracy specificity and sensitivity for all 3 subjects for the FLDA with respect to the selected time window and the amount of training data



(b) Accuracy specificity and sensitivity for all 3 subjects for the SVM with respect to the selected time window and the amount of training data

Figure 7: Accuracy, specificity and sensitivity for all 3 subjects and both classifying techniques. The x-axis represents on each graph the size of the time window taken into account for the classification (from 20 to 1000 ms starting from the feedback onset). The y-axes relates to the amount of training data used to build the classifiers: for the the accuracy graphes, it represents the total number of training data, for the specificity graph, this is the number of non-ErrP data contained in the training set and for the sensitivity graph, the number of ErrP data contained in the training set.

the fact that the negative deflection between the 2 positive peaks in the case of an incorrect feedback is stronger than in the case of a correct one (this is why this negative peak in the *error-minus-correct* is stronger for subject S8, while her response to incorrect feedback is very similar to that of subject S9). These differences are not present in the responses for subject S7; this explains why, in his case, we do not observe any significant negative peak in the *error-minus-correct*. However, for this subject, the first positive peak in the response to feedback is significantly higher in the case of an error than in the correct case; leading to a positive peak in the *error-minus-correct* at around 200 ms after the feedback onset. This was observed neither in our other subjects nor in [20, 21]. The fact that this subject was the only non-naïve participant to the experiment might explain this difference, although, as each subject performed 6 to 7 sessions, they all quickly became experienced with the system and the shape of their EEG responses to feedback were quite stable across sessions.

This seems to indicate a significant variability in the way the brain processes errors across subjects for a given context. Nevertheless, in our case, the concern is the fact that the observed differences between both types of feedback are significantly different, giving us good hopes about the possibility of classifying accurately those EEG responses.

4.2. Classifying the feedback responses

We aimed at studying the influence of the amount of training data on the accuracy of the classification of the EEG responses as ErrP and non-ErrP. For this, we built several datasets for each subject with an increasing number of training data. Each dataset was composed of a training and a testing dataset respectively used to build the classifier and measure its accuracy. Each data point was associated to a label corresponding to its class (+1 for ErrP and −1 for non-ErrP).

The data were taken in the same order as they were recorded (to stay close to reality). To build the first training set, we started from the data corresponding to the first feedback and included the following ones, until we reached 5 responses to incorrect feedback. The rest was used as the first test set. To build the other datasets, we incremented this method by adding each time the group of data containing the 5 next responses to incorrect feedback. We made sure that the last testing set con-

tained at least 10 responses to incorrect feedback. In this way, we obtained 32 datasets for subject S7, 20 for S8 and 22 for S9 (the y-axes on Figs. 7a and 7b show the corresponding amount of training data).

As a preprocessing step, the EEG signals from the 8 channels were filtered between 0.5 and 30 Hz (zero-phase 4^{th}order Butterworth filter) and downsampled to 250 Hz. We performed the ErrP classification using a *Fisher Linear Discriminant Analysis* (FLDA) and a linear *Support Vector Machine* (with a 10-folds cross-validation for the optimization of the regularization parameter, [32]) on a growing time window starting from the feedback onset until 1 second after, with steps of 20 ms (this corresponds to the x-axes on Figs. 7a and 7b).

The results are plotted for each subject in Fig. 7a for the FLDA and in Fig. 7b for the SVM. The color corresponds to the classification accuracy (from 0 to 100% as the color bars indicate). Due to the fact that the datasets are highly unbalanced (the amount of ErrP data represents between 12% and 26% of the data) and that it is very important to minimize the number of false negatives (amount of non-ErrP classified as ErrP), we also plotted for each participant the specificity (proportion of non-ErrP data that are correctly classified) and the sensitivity (proportion of ErrP data that are correctly classified).

Our interest in assessing not only accuracies but also specificities and sensitivities lies in the fact that, for strongly unbalanced dataset, a high accuracy could be due to a substantial bias towards the class with the largest training size, thus leading to a classification of most data as belonging to this class. This could not be observed by only looking at the global accuracy.

This is also why we chose to compare performances of the SVM and FLDA classifiers; indeed if SVMs are known to react to unbalanced training dataset by creating a relative bias towards the most represented class (see [35]), the LDA tends to be less influenced by such a disproportion (see [36]). This behavior seems to be confirmed by the results plotted in Fig. 7: the results from the linear SVM show for all 3 subjects very high values for the specificity and relatively lower values for the sensitivity (mainly for subjects S7 and S9). Whereas the results from the FLDA show more balanced performances for specificity and sensitivity. Although for the global accuracy, the linear SVM seems to outperform the FLDA.

It would be tempting to stress the much better sensitivity coming from the FLDA with respect to the SVM (after all, our aim is to detect correctly ErrP responses), however, one should keep in mind that before managing to identify properly ErrP responses, the first task of the classifier is to avoid the misclassification of non-ErrP responses. The first reason for this is that, as mentioned earlier, non-ErrP responses are more present and have thus a stronger influence on the global accuracy. The second reason is the frustration that ensues from the misclassification of an EEG response to the feedback of a correctly detected symbol; if this happens too often, the subject might not even consider the advantage of detecting ErrPs.

Those results also show that the amount of training data seems to be more influential on the sensitivity than on the specificity (certainly due to the unbalanced dataset) and that, in order to reach a sensitivity stabilized over 50%, we need training datasets containing at least 25 instances of ErrP for the FLDA classifier and 50 for the SVM (except in the case of subject S8 who gets high sensitivity values almost immediately, although we observe a decrease in the sensitivity between 25 and 50).

Concerning the influence of the time window, we observe, in most cases, an important increase in the performance around 350 ms which corresponds to the time interval of the statistically significant peaks of the *error-minus-correct* for all 3 subjects (cf. Fig. 6)

5. Discussion

5.1. ErrP detection for the Mind Speller

The question arises now about how to use an ErrP detection tool in the particular context of the P300 Speller. To detect the target symbol, the classification algorithm computes a score for each row and column of the matrix and then selects the best row and column. From those scores, we can deduce a ranking of all the symbols of the matrix. One simple strategy could be, after ErrP detection, to simply repeat the sequence of intensifications, with eventually a lower number of repetitions, and to update this ranking. But this would lead to an important increase in the time taken to communicate the symbol. Another strategy could be, when the presence of an ErrP is detected, to select the second best symbol according to the classifier's ranking. This approach has the advantage of not increasing the stimulation time. This is supported by the fact that in many cases of wrong symbol detection, we could observe in our experiments that at least the column or the row of the target symbol was correctly identified. When looking at Table 3, we can see that for subjects S2 and S8, in 70% of the case when a letter was misspelled, the correct symbol was ranked in second position by the classifier. Table 3 also shows how such a strategy could improve the performance of the Mind Speller for all 9 subjects. We can observe a substantial theoretical increase in the typing performance (up to 15%).

However, the accuracy presented in the last column of Table 3 assumes a perfect detection of the ErrP. In a more realistic case, we have to keep in mind that not all ErrPs are correctly detected and new mistakes can appear when responses to correct feedback are wrongly classified. From the classification results of our second study, we measured the gain of this method based on ErrP classification and symbol ranking compared to the original results with no ErrP detection. In Fig. 8, we present, for each subject and both classification methods, the accuracy difference on the test sets between those two approaches with respect to the amount of training data. The results are presented for the classifiers trained and tested on EEG data with a time window of 700 ms length starting from the feedback onset. The y-axis indicates thus the accuracy gain in percent of the new method with respect to the original accuracy, so that negative values actually indicate a loss in typing accuracy. The gain for the linear SVM classifier (green curves) and from the FLDA (red curves) can be compared to the maximum theoretical gain (assuming a perfect ErrP classification, blue curves). We can observe that for all 3 subjects the SVM outperforms the FLDA,

Table 3: Details of the performances of each participant to both series of experiments. The 4th column details the number of mistyped symbols for which the real target was ranked in second position by the classifier (and the proportion of this number with respect to the total number of mistyped symbols). The last column shows the theoretical new accuracy, assuming a ideal ErrP classification and performing the selection of the secondly ranked symbol when an ErrP is detected.

Subject	Gender	Age	Total number of typed symbols	Typing accuracy	Number of target ranked in second position (percentage of total mistake)	New accuracy assuming perfect ErrP detection
S1	M	24	32	81%	4 (67%)	94%
S2	F	23	65	85%	7 (70%)	95%
S3	M	34	37	81%	3 (43%)	89%
S4	M	27	59	73%	9 (56%)	88%
S5	F	22	60	88%	8 (62%)	92%
S6	M	29	56	66%	7 (37%)	79%
S7	M	27	659	74%	93 (54%)	88%
S8	F	24	963	88%	80 (70%)	96%
S9	F	24	758	84%	77 (64%)	94%

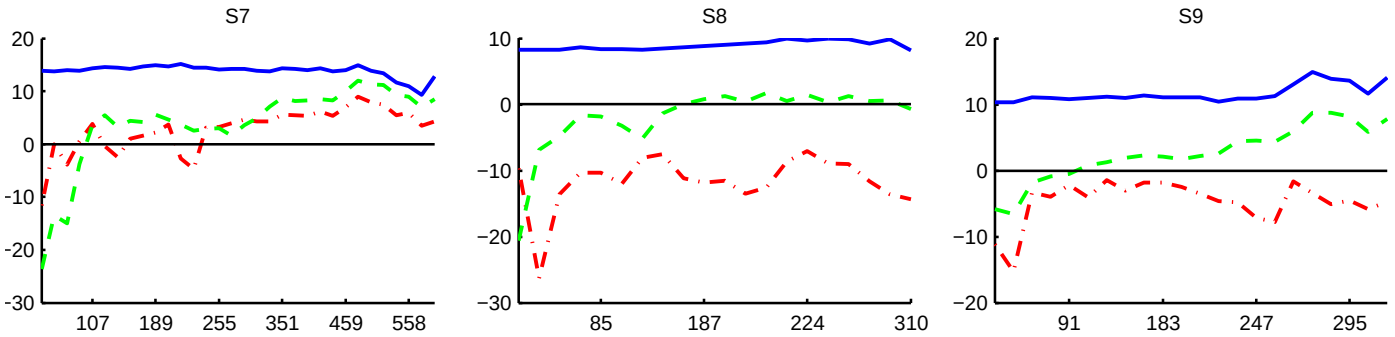


Figure 8: Gain in typing accuracy (%, *y*-axes) on test data with respect to the size of the training set for subjects S7 S8 and S9. The *x*-axes represent the total amount of training data. The classifications were done for a time window of 700 ms starting from the feedback onset. The red curves (dash-dotted lines) represent this gain for the FLDA, the green ones (dashed lines) for the linear SVM and the blue ones (solid lines) illustrate the maximum possible gain, assuming a perfect ErrP/non-ErrP classification, for each test set.

although for subject S8, even with a high amount of training data, the maximum gain remains close to zero. Subjects S7 and S9 seem to benefit from the SVM ErrP detection with an amount of training data starting from around 100 and 250 respectively. Subject S7 is the only one who seem to benefit from the ErrP detection based on FLDA; for the 2 other subjects, this method leads to a decrease of their typing performances. These results illustrate the importance of minimizing the proportion of false negatives, and the fact that having a classifier biased towards the non-ErrP class is at our advantage in the specific context of ErrP detection for the P300 Speller.

5.2. Combining ErrP detection with other strategies

One can think of several ways to improve the ErrP detection accuracy in the context of the P300 Speller: a possibility would be to weight the scores of each symbol with an *a priori* probability of occurrence given the previous symbol and the typing language (*e.g.*, Dasher [37]). Combining this approach with an ErrP classification algorithm might lead to a better error detection. The disadvantage of such a technique is that it would be language specific and not usable for proper nouns or non-text based communication (*e.g.*, icon-based communication).

Another way would be to use results from the P300 Speller classifier itself. As previously mentioned, the classifier ranks the symbols according to their scores. We could infer from those scores a "certainty measure" of the classifier depending how high the score of the first ranked letter is with respect to the other scores. This "certainty measure" could then be used as a *prior* probability of error to weight the results of the ErrP classifier.

5.3. Conclusion

A first step towards the integration of ErrP detection in the P300 Speller BCI was presented. Besides the undeniable practical advantages of ErrP detection, the necessity of gathering enough training data, the importance of minimizing the amount of false positives in a single trial detection and the strong noisy component of the EEG signals make this task very challenging. If, from a practical point of view, performing hours of training in order to build an ErrP classifier is not acceptable for a commercial device, a solution would be to let the user to utilize the BCI and become familiar with the device before enhancing it with the ErrP detection tool once a sufficient amount of training data has been collected. Combining the ErrP detection with other techniques based or not on EEG processing might help reduce the difficulty of the task. In such a case, the contribution of all techniques involved should be measured, so as to know whether the ErrP detection takes a determinant role in the correction of typing errors.

Acknowledgements

AC and AR are supported by a specialization grant from the Agentschap voor Innovatie door Wetenschap en Technolo-

gie (IWT, Flemish Agency for Innovation through Science and Technology). NC is supported by IST-2007-217077. NVM is supported by the Flemish Regional Ministry of Education (Belgium) (GOA 10/019). JAKS acknowledges support of FWO G.0588.09, G.0302.07, CoE EF/05/006, GOA-MANET, IUAP DYSCO. MMVH is supported by research grants received from the Financing program (PFV/10/008) and the CREA Financing program (CREA/07/027) of the K.U.Leuven, the Belgian Fund for Scientific Research – Flanders (G.0588.09), the Interuniversity Attraction Poles Programme – Belgian Science Policy (IUAP P6/29), the Flemish Regional Ministry of Education (Belgium) (GOA 10/019), and the European Commission (IST-2007-217077), and by the SWIFT prize of the King Baudouin Foundation of Belgium. The authors wish to thank Refet Firat Yazicioglu, Tom Torfs and Chris Van Hoof from imec Leuven for providing us with the wireless EEG system.

References

[1] Editorial Comment: Is this the bionic man?, Nature 442 (2006) 109.

[2] P. Sajda, K.-R. Muller, K. Shenoy, Brain-Computer Interfaces [from the guest editors], IEEE Signal Processing Magazine 25 (2008) 16–17.

[3] M. A. L. Nicolelis, Brain-Machine Interfaces to Restore Motor Function and Probe Neural Circuits, Nature Reviews Neuroscience 4 (2003) 417–422.

[4] B. Pesaran, S. Musallam, R. A. Andersen, Cognitive Neural Prosthetics, Current Biology 16 (2006) R77–R80.

[5] H. Segers, A. Combaz, N. V. Manyakov, N. Chumerin, K. Vanderperren, S. Huffel, M. M. Hulle, Steady State Visual Evoked Potential (SSVEP) - Based Brain Spelling System with Synchronous and Asynchronous Typing Modes, in: 15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011), volume 34 of *IFMBE Proceedings*, Springer Berlin / Heidelberg, 2011, pp. 164–167.

[6] P. Martinez, H. Bakardjian, A. Cichocki, Fully Online Multicommand Brain-Computer Interface with Visual Neurofeedback Using SSVEP Paradigm, Computational Intelligence and Neuroscience (2007) 94561.

[7] M. Cheng, X. Gao, S. Gao, D. Xu, Design and Implementation of a Brain-Computer Interface with High Transfer Rates, IEEE Transactions on Biomedical Engineering 49 (2002) 1181–1186.

[8] N. Birbaumer, A. Kübler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, H. Flor, The Thought Translation Device (TTD) for Completely Paralyzed Patients, IEEE Transactions on Rehabilitation Engineering 8 (2000) 190–193.

[9] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, G. Curio, The Non-Invasive Berlin Brain-Computer Interface: Fast Acquisition of Effective Performance in Untrained Subjects, NeuroImage 37 (2007) 539–550.

[10] J. R. Wolpaw, D. J. McFarland, T. M. Vaughan, Brain-Computer Interface Research at the Wadsworth Center, IEEE Transactions on Rehabilitation Engineering 8 (2000) 222–226.

[11] S. Luck, An introduction to the Event-Related Potential Technique, MIT Press, 2005.

[12] W. S. Pritchard, Psychophysiology of P300, Psychological Bulletin 89 (1981) 506–540.

[13] L. Farwell, E. Donchin, Talking Off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Brain Potentials, Electroencephalography and Clinical Neurophysiology 70 (1988) 510–523.

[14] C. B. Holroyd, M. G. Coles, The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity, Psychological Review 109 (2002) 679–709.

[15] P. Ferrez, J. del R. Millán, EEG-Based Brain-Computer Interaction: Improved Accuracy by Automatic Single-Trial Error Detection, in: Advances in Neural Information Processing Systems, volume 20, MIT Press, Cambridge, MA, 2008, pp. 441–448.

[16] W. J. Gehring, M. G. H. Coles, D. E. Meyer, E. Donchin, A Brain Potential Manifestation of Error-Related Processing, Electroencephalography and Clinical Neurophysiology, Supplement 44 (1995) 261–272.

[17] B. Blankertz, C. Schäfer, G. Dornhege, G. Curio, Single Trial Detection of EEG Error Potentials: A Tool for Increasing BCI Transmission Rates, in: Artificial Neural Networks ICANN 2002, volume 2415 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2002, pp. 1137–1143.

[18] P. Ferrez, J. del R. Millán, Error-Related EEG Potentials Generated During Simulated Brain Computer Interaction, IEEE Transactions on Biomedical Engineering 55 (2008) 923–929.

[19] G. Schalk, J. R. Wolpaw, D. J. McFarland, G. Pfurtscheller, EEG-based Communication: Presence of an Error Potential, Clinical Neurophysiology 111 (2000) 2138 – 2144.

[20] B. D. Seno, M. Matteucci, L. T. Mainardi, Online Detection of P300 and Error Potentials in a BCI Speller, Computational Intelligence and Neuroscience (2010) 307254.

[21] G. Visconti, B. D. Seno, M. Matteucci, L. T. Mainardi, Automatic Recognition of Error Potentials in a P300-Based Brain-Computer Interface, in: 2008 Proceedings of the 4th International Brain-Computer Interface Workshop & Training Course, pp. 238–243.

[22] A. Combaz, N. Chumerin, N. V. Manyakov, A. Robben, J. A. K. Suykens, M. M. Van Hulle, Error-Related Potential Recorded by EEG in the Context of a P300 Mind Speller Brain-Computer Interface, in: 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 65–70.

[23] N. Chumerin, N. V. Manyakov, A. Combaz, J. A. K. Suykens, R. F. Yazicioglu, T. Torfs, P. Merken, H. Neves, C. Van Hoof, M. M. Van Hulle, P300 Detection Based on Feature Extraction in On-line Brain-Computer Interface, in: KI 2009: Advances in Artificial Intelligence, volume 5803 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2009, pp. 339–346.

[24] A. Combaz, N. V. Manyakov, N. Chumerin, J. A. K. Suykens, M. M. Van Hulle, Feature Extraction and Classification of EEG Signals for Rapid P300 Mind Spelling, in: 2009 International Conference on Machine Learning and Applications (ICMLA), pp. 386–391.

[25] R. F. Yazicioglu, P. Merken, R. Puers, C. Van Hoof, Low-Power Low-Noise 8-Channel EEG Front-End ASIC for Ambulatory Acquisition Systems, in: 2006 Proceedings of the 32nd European Solid-State Circuits Conference, pp. 247–250.

[26] J. N. Mak, Y. Arbel, J. W. Minett, L. M. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, D. Erdogmus, Optimizing the p300-based brain-computer interface: current status, limitations and future directions, Journal of Neural Engineering 8 (2011) 025003.

[27] D. Krusienski, E. Sellers, D. McFarland, T. Vaughan, J. Wolpaw, Toward enhanced P300 speller performance, Journal of Neuroscience Methods 167 (2008) 15–21.

[28] D. H. Brainard, The Psychophysics Toolbox, Spatial Vision 10 (1997) 433–436.

[29] D. G. Pelli, The videotoolbox software for visual psychophysics: Transforming numbers into movies, Spatial Vision 10 (1997) 437–442.

[30] N. Cristianini, J. Shawe-Taylor, Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[31] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De,Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific Pub Co Inc, 2002.

[32] S. S. Keerthi, D. DeCoste, A modified Finite Newton Method for Fast Solution of Large Scale Linear SVMs, Journal of Machine Learning Research 6 (2005) 341–361.

[33] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[34] D. S. Moore, G. P. McCabe, W. M. DuckWorth, S. L. Sclove, The Practice of Business Statistics Companion Chapter 18: Bootstrap Methods and Permutation Tests, W. H. Freeman, 2003.

[35] Y.-M. Huang, S.-X. Du, Weighted Support Vector Machine for Classification with Uneven Training Class Sizes, in: 2005 Proceedings of the 4th International Conference on Machine Learning and Cybernetics, volume 7, pp. 4365–4369.

[36] P. Xu, P. Yang, X. Lei, D. Yao, An Enhanced Probabilistic LDA for Multi-Class Brain Computer Interface, PLoS ONE 6 (2011) e14634.

[37] D. J. Ward, D. J. C. MacKay, Fast Hands-free Writing by Gaze Direction, Nature 418 (2002) 838.