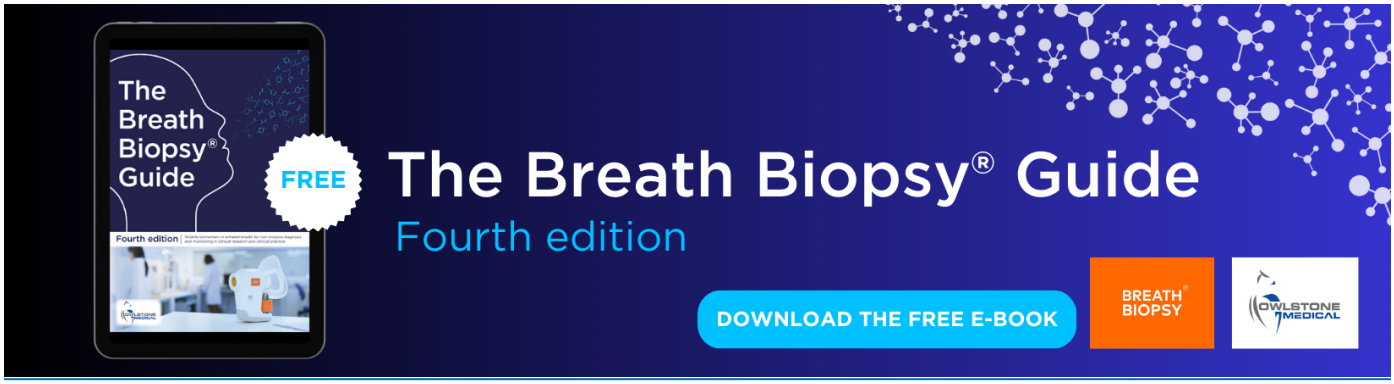
PAPER

Self-calibration algorithm in an asynchronous P300-based brain–computer interface



IOP Publishing Journal of Neural Engineering
J. Neural Eng. **11** (2014) 035004 (9pp) doi:10.1088/1741-2560/11/3/035004

Self-calibration algorithm in an asynchronous P300-based brain–computer interface

F Schettini[1,2], **F Aloise**[1,2], **P Aricò**[1,2], **S Salinari**[2], **D Mattia**[1] **and F Cincotti**[1,2]

[1] Neuroelectrical Imaging and BCI Lab, IRCCS Fondazione Santa Lucia, Rome, Italy
[2] Department of Computer, Control, and Management Engineering, University of Rome 'Sapienza', Rome, Italy

E-mail: f.schettini@hsantalucia.it

Received 31 August 2013, revised 1 April 2014
Accepted for publication 8 April 2014
Published 19 May 2014

Abstract

Objective. Reliability is a desirable characteristic of brain–computer interface (BCI) systems when they are intended to be used under non-experimental operating conditions. In addition, their overall usability is influenced by the complex and frequent procedures that are required for configuration and calibration. Earlier studies examined the issue of asynchronous control in P300-based BCIs, introducing dynamic stopping and automatic control suspension features. This report proposes and evaluates an algorithm for the automatic recalibration of the classifier's parameters using unsupervised data. *Approach.* Ten healthy subjects participated in five P300-based BCI sessions throughout a single day. First, we examined whether continuous adaptation of control parameters improved the accuracy of the asynchronous system over time. Then, we assessed the performance of the self-calibration algorithm with respect to the no-recalibration and supervised calibration conditions with regard to system accuracy and communication efficiency. *Main results.* Offline tests demonstrated that continuous adaptation of the control parameters significantly increased the communication efficiency of asynchronous P300-based BCIs. The self-calibration algorithm correctly assigned labels to unsupervised data with 95% accuracy, effecting communication efficiency that was comparable with that of supervised repeated calibration. *Significance.* Although additional online tests that involve end-users under non-experimental conditions are needed, these preliminary results are encouraging, from which we conclude that the self-calibration algorithm is a promising solution to improve P300-based BCI usability and reliability.

Keywords: brain–computer interface (BCI), asynchronous control, self-calibration algorithm, unsupervised calibration, P300 event-related potential

1. Introduction

Brain–computer interface (BCI) systems allow users to communicate with the environment and control electrical devices without using muscles and nerves [1]. One of the principal aims of BCI technology is to restore communication and interaction with the external world in people with severe motor disabilities [2], and other applications of BCI technology have recently been proposed, such as rehabilitation [3], additional control channel [4], and games and monitoring applications [5].

Despite the considerable scientific advancements in recent years, unresolved issues remain that prevent the widespread use of these systems as assistive technologies (AT). To shrink the gap between BCI systems and other augmentative and alternative communication technologies, BCI systems should have greater reliability and have simple configuration and calibration procedures [6]. Moreover, the throughput speed should be faster, and the operation mode should match daily life necessities.

Of the physiological signals that are usable as control features for a BCI, the P300 is an event-related potential (ERP)

that is widely used for communication and environmental control, because it allows one to select an item of interest between a set of available choices with relatively little effort (e.g. no user training, short calibration sessions, possibility to display several items at once). ERPs vary widely between subjects and within the same subject [7, 8]. Further, external factors, such as light, noise, and stimulation modalities [9, 10], and 'internal' factors, such as attentional level and fatigue, can affect the morphology of these potentials [11, 12]. As noted by Thompson *et al* [13], these factors influence the reliability of BCI systems, who reported variability in the morphology of the P300 potential across BCI sessions.

The tuning of parameters that control the system should be updated frequently to ensure peak performance. However, frequent explicit recalibration of the system (i.e. supervised acquisition of data to train the classifier) is time-consuming and frustrating for users. Thus, classification methods [14–16] for partial and complete unsupervised learning in P300-based BCIs and user-friendly solutions [17] have been proposed to simplify or eliminate the configuration and calibration processes. However, these approaches were tested in brief controlled BCI sessions (1–2 h), and inter-session variability was not assessed.

Moreover, the proposed methods did not address two important issues to decrease the gap between BCI and AT input devices: (*i*) BCIs should implicitly withhold control when the user is not attending to the interface, even without an explicit mechanism to enter a pause mode; and (*ii*) BCIs should dynamically adapt the speed of selection to the subject's skills (dynamic stopping) and provide an appropriate tradeoff between recognition accuracy and speed, allowing the system to maintain a high level of communication efficiency. The asynchronous classifier that was proposed by Aloise *et al* [18] resolved these issues, increasing the communication efficiency of P300-based BCI systems for communication and environmental control applications [19].

This study (i) determined whether a repeated (automatic) update of the classifier's parameters across BCI sessions increased the system's performance in terms of accuracy and communication efficiency and (ii) proposed and evaluated a self-calibration algorithm to label data that were acquired in the unsupervised modality. The algorithm will be used to update the classifier parameters without the need for an explicit calibration session.

2. Materials and methods

2.1. Experimental protocol

Ten healthy subjects participated in this study (five males and five females, mean age 25 $\pm$ 3 years). All subjects had previous experience with P300-based BCI systems and had normal or corrected-to-normal vision. Scalp EEG signals were recorded (g.USBamp, gTec, Austria, 256 Hz) from eight scalp positions (Fz, Cz, Pz, Oz, P3, P4, PO7 and PO8—[20]), referenced to the right earlobe and grounded to the left mastoid. The stimulation interface comprised a 6 $\times$ 6 matrix that contained alphanumeric characters (P300 speller—[21]). The recordings of stimulation and EEG data were managed by the BCI2000 framework [22].

Visual stimulation consisted of the pseudorandom intensification of rows and columns on the interface: target stimuli comprised the intensification of the row or column that contained the character that was attended to by the subject, and non-targets were the intensification of any other row or column. Each row and column was intensified for 125 ms, and 125 ms elapsed between the end of a stimulus and the onset of the subsequent one (inter-stimulus interval). A stimulation *sequence* consisted of the consecutive intensification of every row and column on the interface, for a total of 12 stimuli (2 targets and 10 non-targets).

The term *trial* refers to a set of eight successive repetitions of the stimulation sequence, relating to the same target character. A *run* is an uninterrupted series of six trials, followed by a pause in EEG acquisition. A *session* consisted of six control runs and two no-control runs. During each *control* run, 6 characters were prompted as the target; within each session, all 36 characters of the interfaces were prompted exactly once. During the two *no-control* runs, EEG data were acquired while the subject was in a voluntary no-control state: subjects were required to gaze at a fixed cross in the middle of the interface and ignore the surrounding stimulation. In one of the two no-control runs, subjects were also required to solve simple arithmetic problems that were posed by the experimenter [18].

Each subject underwent five recording sessions in the same day at 10:00 AM, 12:00 PM, 2:00 PM, 4:00 PM and 6:00 PM. Data for 180 control trials and 60 no-control trials were collected per subject. The subjects were required to wear the EEG cap for the entire day. Before each session, the experimenter checked the correct position of the cap on the subject's scalp, the electrode-scalp impedance value (which was kept below 10 kΩ) and the quality of the EEG signal. Each session lasted approximately 1 h, and between consecutive sessions, the subject could perform daily activities, such as working, studying, talking with friends and eating.

2.2. EEG pre-processing

The 8-channel EEG signal was segmented into 800 ms epochs, starting at the onset of each stimulus. The epochs were then downsampled, replacing each segment of 12 samples with their mean and then reducing an epoch to 17 samples. The resulting 8 $\times$ 17 data arrays were concatenated, creating a 136-element feature vector v_f for each stimulus. The classifier was trained on the resulting set of feature vectors, each associated with the label of a *target* or *non-target* stimulus. In addition, epochs that were related to no-control periods were included in the training set by labeling them as non-target epochs.

Stepwise linear discriminant analysis was performed to identify the most significant features and estimate the weight w of the linear classifier (non-significant features were given a weight $w_i = 0$). The number of maximum iterations of the algorithm was set to 60. For each iteration, features with p-value <0.1 were added to the model, whereas those with p-value >0.15 were removed [23]. The score y for each stimulus was calculated as the linear combination of the feature vector multiplied by the classifier's weight ($y = w^T v_f$).

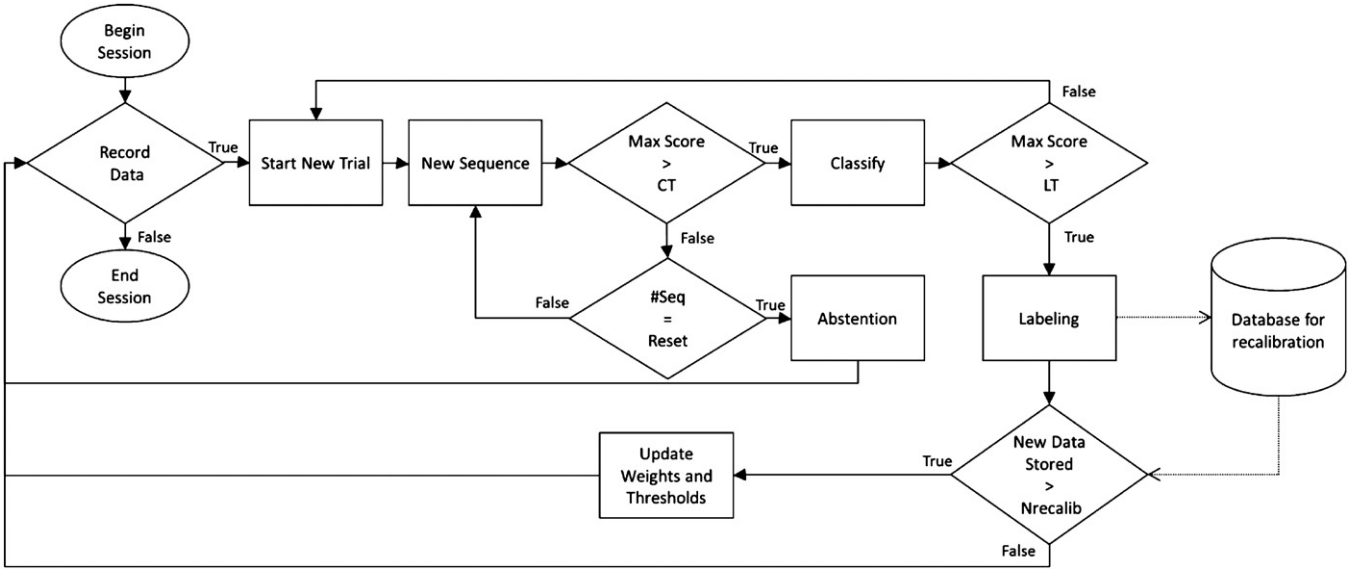


Figure 1. Flowchart of the self-calibration algorithm.

2.3. Self-calibration algorithm

The self-calibration algorithm performs (online) unsupervised labeling of data and processes them to automatically update the classifier weights. The proposed method relies on introducing two threshold sets in the classifier. The first will be denoted as the classification threshold (CT), which will permit dynamic stopping and control-suspending features (section 2.3.1). The second set of thresholds will be called the labeling threshold (LT), which will be used to decide the sequences that can be reliably labeled for continuous updating of the classifier's weights (section 2.3.2).

Figure 1 shows a flowchart of the self-calibration algorithm. At the outset, the classifier's parameters and thresholds are defined using data from the previous session (or from an explicit calibration session for initial use of the system). Every time a new stimulation sequence is delivered, the scores for each stimulation class are computed and compared with CTs. When the CT is exceeded for the row and column classes (i.e. a character is tentatively selected), the differences in scores are estimated (see section 2.3.2). If they also exceed the LT (the epochs that relate to the current trial), they are labeled per the classification result and stored for further recalibration. When a predefined number of new epochs ($N_{\text{recalibration}}$) from the online session is stored, the same amount of the oldest epochs is removed from the training dataset, and the classifier weights and threshold values are updated. The $N_{\text{recalibration}}$ value is set as 5% of the number of epochs in the recalibration database. Recalibration might be performed each time a new epoch is added to the recalibration database ($N_{\text{recalibration}} = 1$), if the computational power of the system allows for it.

Conversely, $N_{\text{recalibration}}$ might be set to 100% of the recalibration database's size—i.e. recalibration is performed only when a completely fresh dataset is available. An offline simulation, carried out during a pilot study involving three healthy subjects, has shown that the 5% value might be an effective compromise between update frequency and the computational requirements.

2.3.1. Classification thresholds. The classification thresholds were defined as described in [18]. Their values were recomputed when a new set of $N_{\text{recalibration}}$ epochs was available.

For each stimulation sequence in the training set, 12 y_{stim} scores were computed: 6 for the rows and 6 for the columns on the interface. Within each trial, the stimulus scores of the first through eighth sequence seq were accumulated (summed up), yielding $Y_{\text{stim}}[\text{seq}] = \sum_{i=1}^{\text{seq}} y_{\text{stim}}[i]$, $(\text{seq} = 1, \ldots, 8)$. For each stimulation sequence, the maximum score $M[\text{seq}] = \max_{\text{stim}} \{Y_{\text{stim}}[\text{seq}]\}$ was extracted, and a label that was equal to 1 (target) or 0 (non-target or no-control) was assigned to it. Thus, we defined a distribution of the maximum scores for each stimulation sequence.

Each distribution was used to plot a receiver-operating characteristic (ROC—[24]) curve. To reduce the false positive rate (FPR) and ensure a high true positive rate (TPR), the threshold was selected at the intersection between the ROC curve and the segment that joined the point (0.05, 0.5) with (0, 1), because the former provides a tradeoff between FPR (maximum 5%) and TPR (minimum 50%) and the latter represents the optimal classification (FPR = 0% and TPR = 100%).

2.3.2. Labeling thresholds. With regard to defining the labeling thresholds, a different procedure was applied to the training data to ensure a high level of robustness to false positives.

Starting from the scores that were accumulated according to the number of stimulation sequences that was delivered in each trial, scores that were related to the rows and columns stimuli were sorted, and the differences between the highest

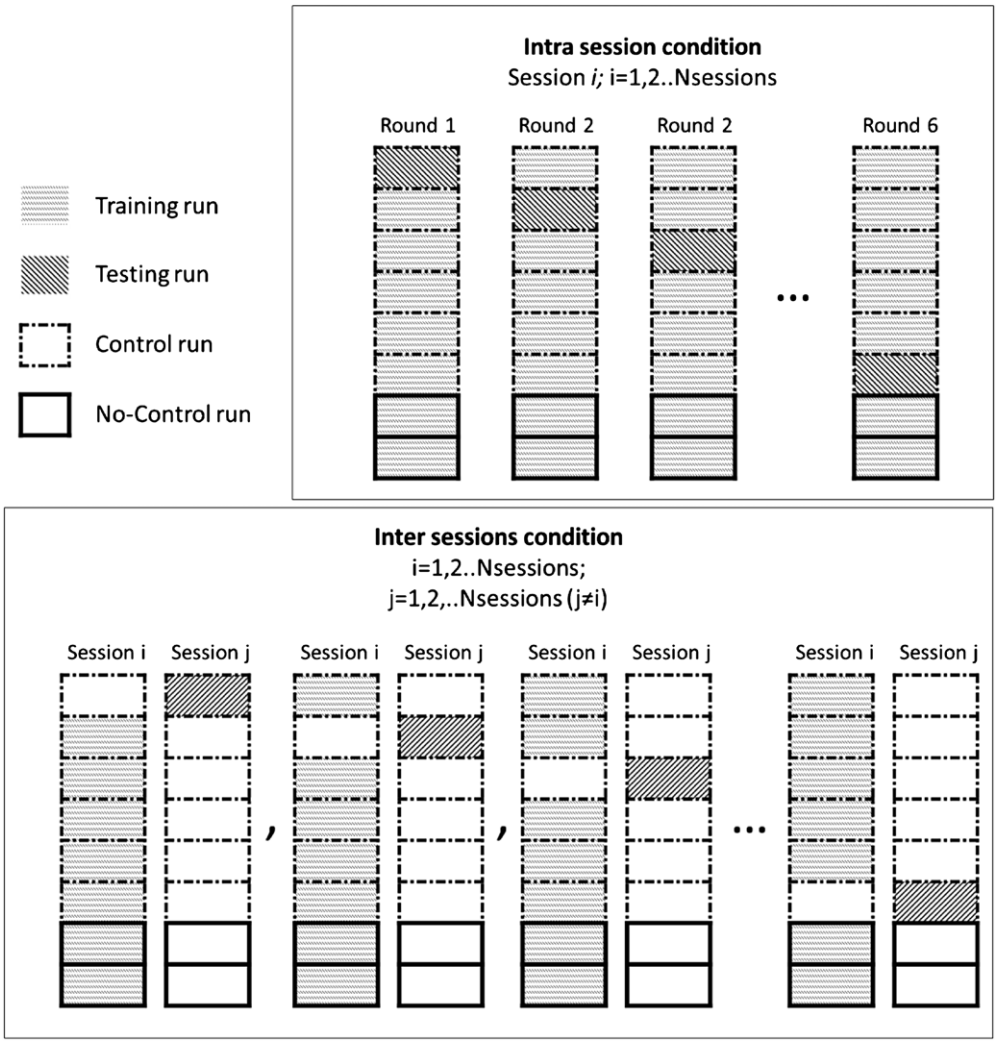


Figure 2. Intra-session and inter-sessions cross-validation. Each block represents a run relating to the spelling of six different characters, except for the two no-control runs, in which the subject was required to divert his attention from the stimulation interface.

and second highest scores were computed for the rows and columns separately. The differences in scores were labeled 1 or 0 if the highest score was a target or a non-target/no-control score, respectively. Thus, it was possible to define a distribution of differences in scores for each stimulation sequence (eight distributions for rows and eight distributions for columns in this case).

The distribution of differences in scores was used to plot 16 ROC curves (8 for rows and 8 for columns). The threshold values corresponded to the point on the ROC curve that ensured a 0% FPR with the maximum possible TPR value on the training data. Consequently, considering the testing data, only trials in which the maximum score differed vastly from the second highest score exceeded the threshold, ensuring a high level of robustness compared with false positives.

2.4. Performance assessment

2.4.1. Intra-session and inter-session validation. We first determined if the accuracy and communication efficiency of an asynchronous P300-based BCI benefited from continuous updating of the classifier's parameters. Two conditions were investigated through offline cross-validation with the asynchronous classifier (i.e. dynamic stopping and abstention features enabled): intra-session and inter-session (figure 2).

In the intra-session condition, the training and testing datasets belonged to the same session. For each cross-validation iteration, the training dataset consisted of five control runs and two no-control runs, and the remaining control run of the same session was used as the testing dataset. Every control run was used as the testing dataset once; thus, six cross-validation iterations were performed for each session.

In the inter-session condition, the training and testing datasets belonged to different sessions. The cross-validation design matched the intra-session's cross-validation design as closely as possible. For each iteration, the training dataset consisted of five control runs and two no-control runs that were extracted from session i, and the testing dataset comprised one control run from session j—i.e. the run with the index that corresponded to session i's run was not included in the training set. Each pair (i, j) of sessions participated in the cross-validation. Performance was assessed in terms of correct classifications, errors and abstentions.

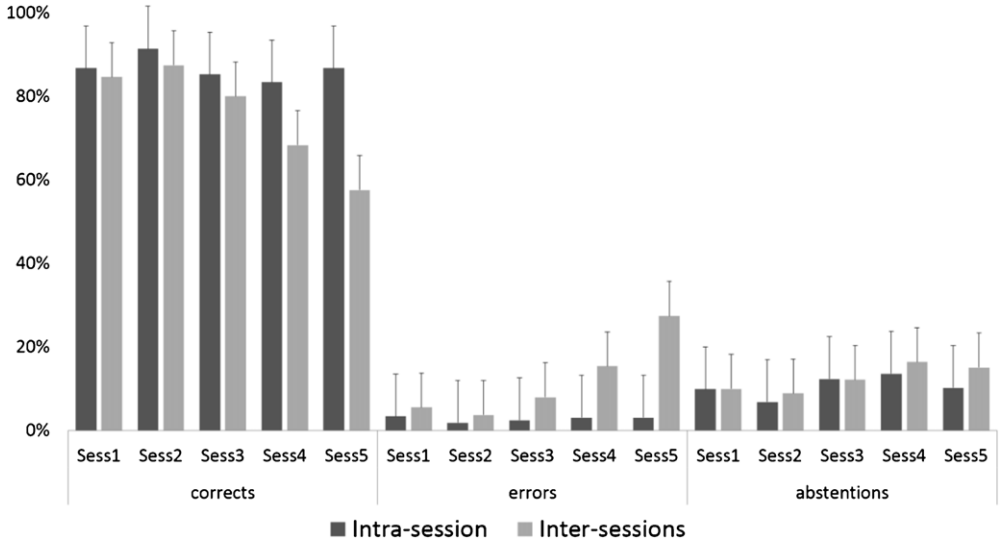


Figure 3. Intra-session and inter-session performance as a function of sessions. Bars denote standard error.

2.4.2. Evaluation of the self-calibration algorithm. We compared the performance of three conditions: no-recalibration, intra-session and self-calibration. In the no-recalibration condition, the classifier's weights and classification threshold values were extracted using data from the first session and applied to the other sessions, because simulates plausible use of the system during the day. The intra-session condition was the same as in section 2.4.1. Even if it was not a realistic condition, we considered the intra-session condition a reference for the best possible performance by continuous supervised calibration of the classifier's parameters.

To determine the performance of the self-calibration algorithm, we applied the following procedure: at the beginning the classifier's weights, the classification thresholds and the labeling thresholds were extracted using data from the first session, and the self-calibration algorithm was run on data from the four sessions that were acquired later. Particularly, beginning with data from the second session, performances were sequentially assessed by runs for all available sessions, updating the database for recalibration accordingly. The classifier's parameters and the thresholds values were updated when 5% of new data (with respect to the dimension of the starting calibration dataset) were stored.

2.4.3. Evaluation of communication efficiency. To summarize the system's performance under the various conditions, we adopted a metric to quantify its efficiency with regard to accuracy, error rate and speed. An asynchronous classifier has three possible classification outputs: (i) correct classification, when the target item is correctly recognized by the system; (ii) error, if the item that is selected differs from the one that is attended by the subject; and (iii) abstention, when no item reaches the classification threshold at the end of the trial.

For this reason, metrics that only consider classification accuracy, such as the written symbol rate (WSR—[25]) and Wolpaw's bit-rate [1], yield an incomplete characterization of asynchronous systems, because they fail to distinguish incorrect selections from abstentions.

In an earlier study [19], the efficiency metric that was introduced by Bianchi *et al* [26, 27] was applied successfully to assess the communication efficiency of a proposed asynchronous system. This metric predicts the extent to which the accuracy of the classification supports communication—i.e. whether the time that is spent in correcting mistakes is less than that needed to generate a correct selection. The efficiency of a system, with regard to the time that is needed to achieve a classification, is expressed as a function of the number of stimulation sequences (NumSeq):

$$\mathrm{Eff} = \frac{1}{\mathrm{NumSeq} * \overline{\mathrm{ESC}}}$$

where ESC is the expected selection cost (ESC), which is the mean number of selections that is needed to generate a correct symbol, taking into account the recovery from errors and abstentions. When the accuracy is lower than 50%, the time that is needed to correct errors is longer than the time that is spent for effective communication; thus, ESC approaches infinity, and Eff is 0.

In this study, we made the following assumptions about costs: we associated a cost of 1 with abstentions (the user only needs to repeat the trial to select the desired character) and a cost of 2 with misclassifications (the user must first delete the incorrect character and then reselect the desired one). Due to the small size of our datasets, we assumed that the probabilities of abstention and misclassification were independent of target character. Additionally, in accordance with the way our target characters were generated, we used uniform probability for all characters on the grid when calculating ESC.

3. Results

3.1. Inter-session and intra-session performance

Results on the average performance in terms of correct classifications, errors and abstentions for test session, for

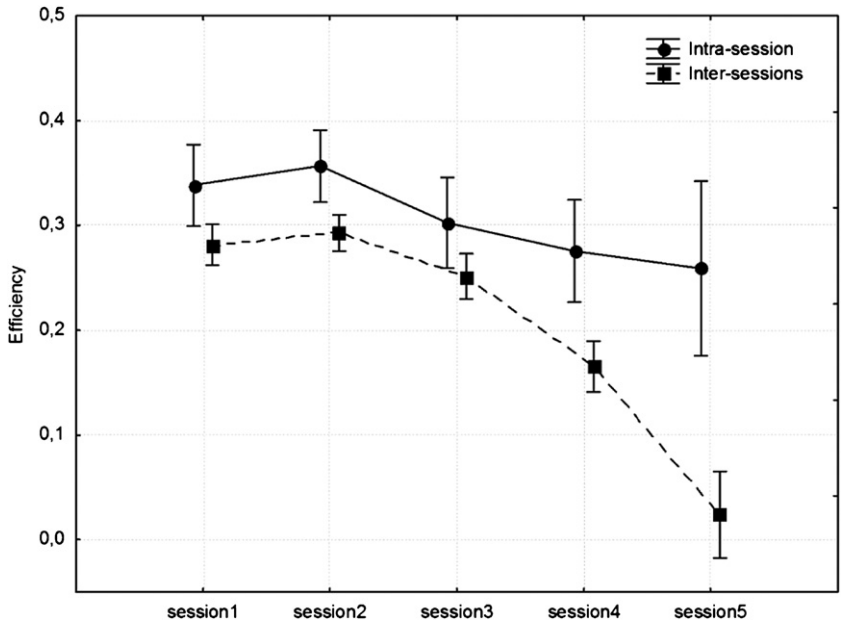


Figure 4. Trend of communication efficiency across sessions for intra-session and inter-session cross-validation conditions.

Table 1. Communication efficiency for each cross-validation condition: values on the main diagonal (in bold) correspond to the intra-session condition, entries outside the main diagonal refer to inter-session condition.

		Testing session				
		Sess1	Sess2	Sess3	Sess4	Sess5
Training session	Sess1	**0.33 ± 0.15**	0.32 ± 0.14	0.26 ± 0.18	0.14 ± 0.14	0.10 ± 0.11
	Sess2	0.29 ± 0.15	**0.35 ± 0.12**	0.27 ± 0.12	0.18 ± 0.14	0.14 ± 0.13
	Sess3	0.29 ± 0.13	0.30 ± 0.13	**0.30 ± 0.15**	0.22 ± 0.12	0.15 ± 0.13
	Sess4	0.28 ± 0.13	0.28 ± 0.14	0.26 ± 0.13	**0.27 ± 0.14**	0.16 ± 0.13
	Sess5	0.26 ± 0.12	0.27 ± 0.08	0.24 ± 0.11	0.21 ± 0.11	**0.26 ± 0.11**

the intra-session and inter-session conditions are reported in figure 3. One-way repeated measures ANOVA was performed three times using the cross-validation conditions as the factor (intra-session versus inter-session) and correct classifications per test session, errors per test session and abstentions per test session as dependent variables. The intra-session condition effected a significantly higher correct classification rate versus the inter-session condition ($F(4, 1192) = 17.232$, $p < 0.01$), a difference that was compensated by a significantly lower error rate in the former ($F(4, 1192) = 15.85$, $p < 0.01$). No significant differences were observed in the percentage of abstentions between conditions ($F(4, 1192) = 1.49$, $p = 0.20$).

Communication efficiency was significantly higher in the intra-session condition than in the inter-session condition, based on repeated measures ANOVA using cross-validation condition as the factor (intra-session and inter-session) and efficiency per test session as the dependent variable ($F(4, 1192) = 10.62$, $p < 0.01$). Figure 4 shows the mean efficiency values for each session over the day in the intra-session and inter-session conditions. Table 1 reports the average efficiency values for each cross-validation condition. Specifically, values on the main diagonal of the matrix correspond to the intra-session condition, and the entries outside of the main diagonal refer to the inter-session conditions.

3.2. Self-calibration algorithm assessment

Figure 5 shows the average performance of the system for all 10 subjects under the three conditions: intra-session, self-calibration and no-recalibration. Because the data that were acquired during the first session were used as the training dataset for all conditions, the performance for session 1 was unavailable. The accuracy fell in the self-calibration and no-recalibration conditions over sessions, whereas the error rate increased. However, performance of the self-calibration algorithm declined versus the no-recalibration condition.

Repeated measures ANOVA was performed three times using the cross-validation conditions as factor (intra-session, self-calibration and no-recalibration) and correct classifications, errors and abstentions per test session as dependent variables. There were significant differences in the correct classification rate between conditions ($F(6, 477) = 6.62$, $p < 0.01$), and by Duncan's post-hoc tests, correct classification in the intra-session cross-validation condition were significantly higher ($p < 0.05$) versus the no-recalibration condition (all sessions) and self-calibration condition sessions 3, 4 and 5. Moreover, in the latter condition, the correct classification rate significantly higher ($p < 0.01$) than in the no-recalibration condition for sessions 3, 4 and 5. None of the other comparisons was significant.

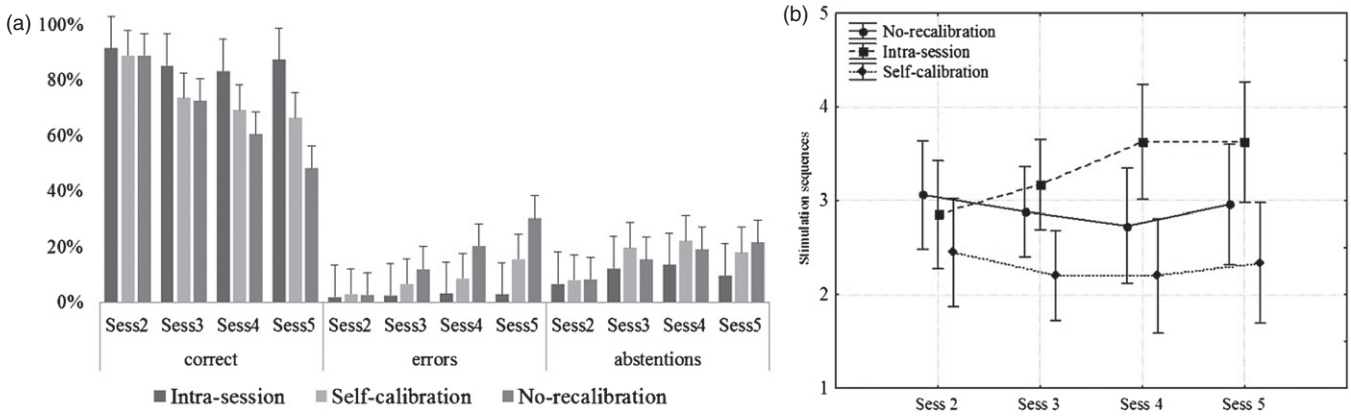


Figure 5. (a) Average performance across all ten subjects of the asynchronous system in the intra-session, self-calibration and no-recalibration cross-validation conditions. Bars denote standard error. (b) Mean of stimulation sequences needed to achieve a correct classification in the intra-session, self-calibration and no-recalibration conditions across all ten subjects.

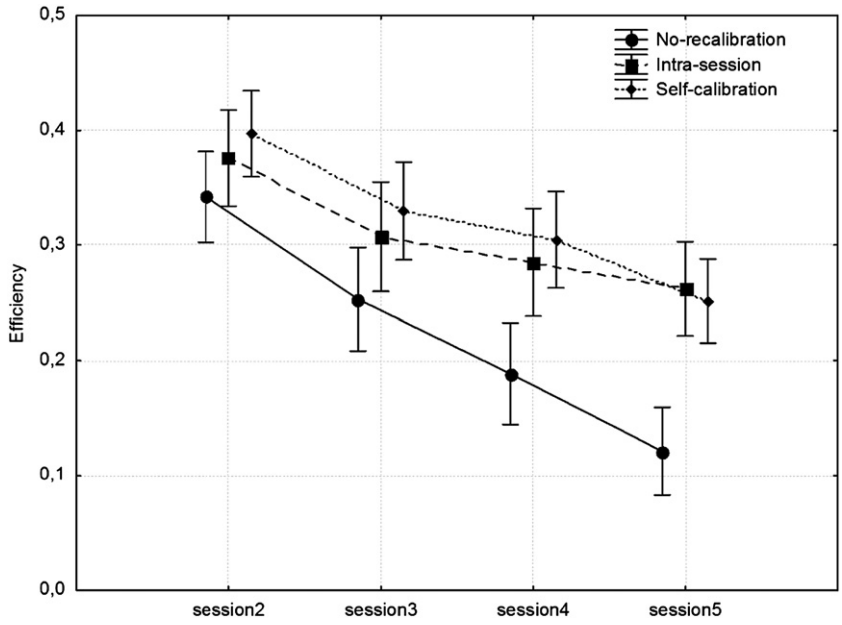


Figure 6. Mean efficiency values for the no-calibration, intra-session, and self-calibration conditions.

By repeated measures ANOVA, the error rate differed between conditions ($F(6, 477) = 5.78, p < 0.01$). By Duncan's post-hoc tests, the error rate in the no-recalibration condition was significantly higher ($p < 0.05$) compared with the intra-session (sessions 3, 4, and 5) and self-calibration conditions (sessions 4 and 5). The error rate in the self-calibration condition was significantly higher ($p < 0.05$) than in the intra-session condition only for session 5. All other comparisons were insignificant, and no significant differences in abstentions were observed between the three conditions ($F(6, 477) = 1.17$, $p = 0.31$).

With regard to the robustness to false positives in the no-control trials, in the self-calibration condition, the classification threshold was erroneously exceeded by 4.1% of the no-control trials versus 19.4% in the no-recalibration condition. Specifically, in the self-calibration conditions, we detected 0.04, 0.09, 0.11 and 0.16 false positives/minute in sessions 2, 3, 4 and 5, respectively. In the no-recalibration condition, there were 0.09, 0.44, 0.53 and 0.88 false positives/minute in sessions 2, 3, 4 and 5, respectively.

Figure 6 reports the efficiency values for the three conditions over all sessions. The self-calibration exhibited efficiency values that approximated those in the intra-session condition, whereas values for the no-recalibration condition were significantly lower than in the other two conditions. By repeated measures ANOVA with the cross-validation conditions as factors (intra-session, self-calibration and no-recalibration) and efficiency as a dependent factor, the differences between groups was not significant ($F(6, 477) = 1.80$, $p = 0.09$). By Duncan's post-hoc test, efficiency was significantly higher for the intra-session (sessions 4 and 5) and self-calibration conditions (sessions 3, 4, and 5) compared with the no-recalibration condition. The efficiency between the intra-session and self-calibration conditions did not differ.

3.3. Unsupervised data labeling

On average, 36 ± 10% of online data exceeded the LT and thus were stored for recalibration; 95.5 ± 3.8% of stored data were correctly labeled as target or non-target. For two of ten subjects, all stored data were labeled with 100% accuracy, whereas the highest percentage of incorrectly labeled data over subjects was 12.5%. However, this rate did not affect performance, because significant differences in efficiency ($p < 0.05$) between the self-calibration and intra-session conditions were seen only for session 4, based on Duncan's post-hoc test on repeated measures ANOVA using cross-validation conditions as factors and efficiency as the dependent variable ($F(6, 45) = 4.67$, $p < 0.01$).

4. Discussion

This report describes and validates an algorithm for the automatic adaptation of a classifier's parameters, designed to be used during online sessions. First, we examined whether updating the parameters increased system accuracy. Contrary to what was reported in [28], recalibration of the system with data that were acquired in the same session ensures greater reliability and efficiency versus recalibration that is performed with data from a different session.

Moreover, in this study, the phenomenon of performance variability was examined only in BCI sessions that were acquired on the same day and with young and healthy subjects. With regard to these aspects, several factors should be considered

(i) overall performance declined in the afternoon sessions compared with morning, even in the intra-session condition, which might be due to fatigue and decreased motivation after repeat sessions;
(ii) this study reports results data that were acquired under controlled experimental conditions, whereas a test that involves end-users in real life contexts would provide more realistic results;
(iii) the authors in [29] demonstrated that potential end-users had wider variability in performance in terms of stimulation sequences that were needed to select the desired item compared with young healthy subjects; thus, the inter-session/intra-session difference might be greater in end-users in the evaluation;
(iv) we are interested in measuring the variability in performance over repeated sessions on different days to determine if the decrease in PM is always present or due primarily to the strict pace of the experimental protocol.

The performance of the self-calibration algorithm was between two conditions: intra-session, which represents the reference condition, and no-recalibration, in which the user was assumed to calibrate the system once and continue to use the same parameters for the entire day. Although the correct classification rate of the self-calibration algorithm was significantly lower than the intra-session condition, it significantly exceeded that of the no-recalibration condition; moreover, the latter had a higher error rate than the other two conditions. The lower error rate, the high correct classification rate and the lower number of stimulation sequences that were needed to exceed the classification in the self-calibration algorithm were reflected in the efficiency values—the communication efficiency that was observed with the self-calibration algorithms did not differ from the intra-session condition.

Finally, the self-calibration algorithm was reliable in labeling data; less than 5% of the data that were stored for recalibration were incorrectly labeled. However, additional tests that involve end users in non-experimental conditions with on-line implementation of the proposed algorithm are needed to confirm the promising results in healthy subjects by offline speculation.

5. Conclusion

In this paper, an algorithm for automatic and continuous adaptation of a classifier's parameters in an asynchronous P300-based BCI system has been described and validated by offline analysis. Continuous recalibration of classifier parameters can enhance the system's performance over several sessions in a single day, and the proposed algorithm can recalibrate the system using unlabeled data from online sessions and ensure performance stability. After an initial supervised calibration session, the entire recalibration procedure becomes hidden to the user, which is a significant property that increases the usability of BCI systems as assistive technology. Although they are not conclusive, these results are promising, and further online tests that involve end users for multiple sessions over several days should be performed to determine the efficacy and reliability of the proposed algorithm in non-experimental conditions.

Acknowledgments

The work is supported in part by the Italian Agency for Research on ALS–ARiSLA, project 'Brindisys.' This paper only reflects the authors' views, and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91
[2] Millán J D R *et al* 2010 Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges *Front. Neurosci.* **4** 161
[3] Pichiorri F, De Vico Fallani F, Cincotti F, Babiloni F, Molinari M, Kleih S C, Neuper C, Kübler A and Mattia D 2011 Sensorimotor rhythm-based brain–computer interface training: the impact on motor cortical responsiveness *J. Neural Eng.* **8** 025020
[4] Blankertz B *et al* 2010 The Berlin brain–computer interface: non-medical uses of BCI technology *Front. Neurosci.* **4** 198
[5] Zander T O and Kothe C 2011 Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general *J. Neural Eng.* **8** 025005

[6] Cipresso P, Carelli L, Solca F, Meazzi D, Meriggi P, Poletti B, Lulé D, Ludolph A C, Silani V and Riva G 2012 The use of P300-based BCIs in amyotrophic lateral sclerosis: from augmentative and alternative communication to cognitive assessment *Brain Behav.* **2** 479–98

[7] Polich J and Kok A 1995 Cognitive and biological determinants of P300: an integrative review *Biol. Psychol.* **41** 103–46

[8] Ravden D and Polich J 1999 On P300 measurement stability: habituation, intra-trial block variation, and ultradian rhythms *Biol. Psychol.* **51** 59–76

[9] Polich J and Bondurant T 1997 P300 sequence effects, probability, and interstimulus interval *Physiol. Behav.* **61** 843–9

[10] Cano M E, Class Q A and Polich J 2009 Affective valence, stimulus attributes, and P300: color vs. black/white and normal versus scrambled images *Int. J. Psychophysiol.* **71** 17–24

[11] Geisler M W and Polich J 1992 P300 and individual differences: morning/evening activity preference, food, and time-of-day *Psychophysiology* **29** 86–94

[12] Polich J 1997 On the relationship between EEG and P300: individual differences, aging, and ultradian rhythms *Int. J. Psychophysiol.* **26** 299–317

[13] Thompson D E, Warschausky S and Huggins J E 2012 Classifier-based latency estimation: a novel way to estimate and predict BCI accuracy *J. Neural Eng.* **10** 016006

[14] Lu S, Guan C and Zhang H 2009 Unsupervised brain computer interface based on intersubject information and online adaptation *IEEE Trans. Neural Syst. Rehabil. Eng.* **17** 135–45

[15] Panicker R, Puthusserypady S and Sun Y 2010 Adaptation in P300 brain–computer interfaces: a two-classifier co-training approach *IEEE Trans. Biomed. Eng.* **57** 2927–35

[16] Kindermans P-J, Verstraeten D and Schrauwen B 2012 A Bayesian model for exploiting application constraints to enable unsupervised training of a P300-based BCI *Plos One* **7** e33758

[17] Kaufmann T, Völker S, Gunesch L and Kübler A 2012 Spelling is just a click away—a user-centered brain–computer interface including auto-calibration and predictive text entry *Front. Neurosci.* **6** 72

[18] Aloise F, Schettini F, Aricò P, Leotta F, Salinari S, Mattia D, Babiloni F and Cincotti F 2011 P300-based brain–computer interface for environmental control: an asynchronous approach *J. Neural Eng.* **8** 025025

[19] Schettini F, Aloise F, Aricò P, Salinari S, Mattia D and Cincotti F 2012 Control or no-control? Reducing the gap between brain-computer interface and classical input devices *Conf. Proc. IEEE Eng. Med. Biol. Soc. 2012* pp 1815–8

[20] Krusienski D J, Sellers E W, McFarland D J, Vaughan T M and Wolpaw J R 2008 Toward enhanced P300 speller performance *J. Neurosci. Methods* **167** 15–21

[21] Farwell L A and Donchin E 1988 Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials *Electroencephalogr. Clin. Neurophysiol.* **70** 510–23

[22] Schalk G, McFarland D J, Hinterberger T, Birbaumer N and Wolpaw J R 2004 BCI2000: a general-purpose brain–computer interface (BCI) system *IEEE Trans. Biomed. Eng.* **51** 1034–43

[23] 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 *J. Neural Eng.* **3** 299–305

[24] Zweig M H and Campbell G 1993 Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine *Clin. Chem.* **39** 561–77

[25] Furdea A, Halder S, Krusienski D J, Bross D, Nijboer F, Birbaumer N and Kübler A 2009 An auditory oddball (P300) spelling system for brain–computer interfaces *Psychophysiology* **46** 617–25

[26] Bianchi L, Quitadamo L R, Garreffa G, Cardarilli G C and Marciani M G 2007 Performances evaluation and optimization of brain computer interface systems in a copy spelling task *IEEE Trans. Neural Syst. Rehabil. Eng.* **15** 207–16

[27] Quitadamo L R, Abbafati M, Cardarilli G C, Mattia D, Cincotti F, Babiloni F, Marciani M G and Bianchi L 2011 Evaluation of the performances of different P300 based brain-computer interfaces by means of the efficiency metric *J. Neurosci. Methods* **203** 361–8

[28] McFarland D J, Sarnacki W A and Wolpaw J R 2011 Should the parameters of a BCI translation algorithm be continually adapted? *J. Neurosci. Methods* **199** 103–7

[29] Aloise F, Schettini F, Aricò P, Salinari S, Guger C, Rinsma J, Aiello M, Mattia D and Cincotti F 2011 Asynchronous P300-based brain–computer interface to control a virtual environment: initial tests on end users *Clin. EEG Neurosci.* **42** 219–24