

Minimizing Calibration Time of Single-Trial Recognition of Error Potentials in Brain-Computer Interfaces

Iñaki Iturrate, Luis Montesano, Ricardo Chavarriaga, José del R. Millán, Javier Minguez

Abstract—One of the main problems of both synchronous and asynchronous EEG-based BCIs is the need of an initial calibration phase before the system can be used. This phase is necessary due to the high non-stationarity of the EEG, since it changes between sessions and users. The calibration limits the BCI systems to scenarios where the outputs are very controlled, and makes these systems non-friendly and exhausting for the users. Although it has been studied how to reduce calibration time for asynchronous signals, it is still an open issue for event-related potentials. Here, we analyze the differences between users for single-trial error-related potentials, and propose the design of classifiers based on inter-subject features to either remove or minimize the calibration time. The results show that it is possible to have a classifier with a high performance from the beginning of the experiment, which is able to adapt itself without the user noticing.

I. INTRODUCTION

Non-invasive brain-computer interfaces (BCI) are systems that decode the brain signals (usually EEG) into actions to control robotic devices or virtual agents, and have been successfully tested in past approaches [1], [2]. The design of an EEG-based BCI implies the choice of a specific brain signal or mental process to use. These signals are usually classified as asynchronous or synchronous. The asynchronous signals mainly rely on changes in the power spectra, such as $\mu - \beta$ rhythms [3] or cognitive mental tasks [4]. The synchronous ones are usually elicited from external events, the most common being the event-related potentials (ERPs) [3].

One of the main problems of both synchronous and asynchronous EEG-based BCIs is the need of an initial calibration phase before being able to use the system. This phase is necessary due to the high non-stationarity of the EEG. These non-stationarities can be related to session-dependent aspects such as the motivation or fatigue of the user, or the slight displacement of electrodes [5]. Additionally, they are user-dependent, meaning that the signals vary between different users due to aspects such as the age or cognitive capabilities [6]. Thus, in a given experiment, the calibration phase is session- and user-dependent. This phase can be boring and exhausting for the user, since it can last for more than 30 minutes, and the user has no control over the device being

used. Furthermore, this phase limits the BCI systems to scenarios where the outputs are very controlled.

In this context, several asynchronous approaches have resulted in supervised and non-supervised learning techniques to cope with the changes in the distributions due to session- and user- related non-stationarities [5], [7], [8]. In a similar way, past studies with synchronous signals have found user-related differences in the averaged P300 ERP component from the neurophysiological point of view [6]. However, there are a few works that deal with the design of robust classifiers for ERPs. Chavarriaga et al. have shown that error-related ERPs remain very stable among different sessions, even with two years of separation between the recordings [9]. However, the changes between different subjects is still a key issue. To our knowledge, there is only a recent paper where the authors designed a non-supervised technique to reduce the P300 calibration time using inter-subject information [10]. However, the P300 was not used in a single-trial manner, but averaged using several repetitions of the stimuli. Thus, it is still unclear how these user-specific changes affect single trial ERPs, and other ERP components.

In this paper, we analyze the problems of non-stationarities between different users with single-trial ERPs. The ERPs chosen were error potentials. In a given task, these potentials encode information about whether a user has considered a specific action erroneous or not. Their presence has been demonstrated in very different situations, such as user's own errors or simply monitoring errors (see [11], [9], [12] among others). This work presents an analysis with respect to the features by using a novel spatio-temporal filter. Additionally, we propose the design of classifiers based on inter-subject features to detect single-trial error potentials and either remove the calibration time or make it more enjoyable.

II. METHODS

A. Experimental setup

Four subjects (one female, mean age 28 ± 7.68) participated in the experiments. The users were comfortably seated in a chair facing a computer screen at approximately one meter of distance. In the experiment performed, the screen displayed a blue dot on the center and three rectangles located at the same distance from the blue dot in a triangle-like shape (Figure 1 left). The blue dot performed discrete actions (Figure 1 right) to one of the squares, staying one second over the square, and one second on the center position. One condition of the experiment was represented by coloring one of the squares green. The users were asked to just monitor and evaluate the actions performed by the blue

Iñaki Iturrate, Luis Montesano and Javier Minguez are with the Instituto de Investigación en Ingeniería de Aragón (I3A) and Dpto. de Informática e Ingeniería de Sistemas (DIIS), Universidad de Zaragoza, Spain. E-mail: {iturrate, montesano, jminguez}@unizar.es. Ricardo Chavarriaga and José del R. Millán are with EPFL, Chair in Non-Invasive Brain-Computer Interface (CNBI), CH-1015 Lausanne, Switzerland. E-mail: name.surname@epfl.ch. This work has been partially supported by the Spanish Government through projects HYPER-CSD2009-00067 and DPI2009-14732-C02-01.

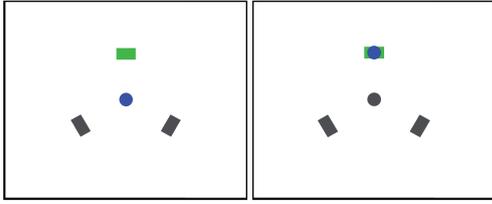


Fig. 1. Design of the experiment.

dot in the following way: a movement to the green square was a correct action, and movements to the other squares were incorrect. Each block of the experiment consisted of 50 random movements with the green square fixed, with a 0.2 probability of error occurrence. Once the block had finished another block started, with the green square being in a different position. 12 runs of each condition were executed, obtaining a total of 1140 correct responses and 360 error responses. The protocol was based on [9].

The instrumentation used to record the EEG brain activity was a gTec system using 32 active electrodes. The location of the electrodes were FP1, FP2, F7, F8, F3, F4, T7, T8, C3, C4, P7, P8, P3, P4, O1, O2, AF3, AF4, FC5, FC6, FC1, FC2, CP5, CP6, CP1, CP2, Fz, FCz, Cz, CPz, Pz and Oz (according to the international 10/20 modified system). The ground electrode was positioned on FPz and the reference electrode was placed on the right earlobe. The EEG was amplified, digitized with a sampling frequency of 256Hz, and power-line notch-filtered. A common average reference (CAR) filter was applied to the signal, and finally a $[0.5 - 10]$ Hz band-pass filter was also applied. The ERP responses were extracted using the time window $[0 - 1000]$ ms, where the action started at $t = 0$ ms. The EEG acquisition and experimental interface were developed under the BCI2000 platform [13].

B. Spatio-temporal filtering

Here, we present a spatio-temporal filter that allowed us to obtain the most discriminative spatio-temporal patterns between the error and correct responses [14]. This technique can either be used to perform feature selection for classification, or to perform characterization analysis. Previous to perform the filtering, it is necessary to choose the n desired channels and the time window of m samples. Then, the filter executes the following steps:

- 1) For each ERP response the input matrix of size $n \times m$ is reshaped as a single feature vector of size $1 \times (n \cdot m)$.
- 2) Each feature is normalized between $[0 - 1]$.
- 3) The normalized features are decorrelated using principal component analysis (PCA) without performing any dimensionality reduction.
- 4) The r^2 statistical coefficient is computed to determine those features that best separate the data, and only the best features are retained.

C. Single-trial classification

For the classification, a classical linear discriminant analysis (LDA) was chosen, since linear classifiers have demon-

strated their feasibility for BCI applications [3], [7]. The LDA discriminant function $D(x)$ is the hyperplane that maximally separates the distribution of the two classes:

$$D(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} + b \quad (1)$$

where \mathbf{w} and b are the normal vector to the hyperplane and the corresponding bias. LDA assumes that both classes have the same covariance Σ , estimating also the mean μ_i of each class i , and the global mean, μ . Then, the parameters of the hyperplane are computed as follows:

$$\mathbf{w} = \Sigma^{-1}(\mu_2 - \mu_1) \quad (2)$$

$$b = -\mathbf{w}^T \mu \quad (3)$$

Next, we present three classification strategies (CS) proposed for minimizing the calibration time.

1) *CS I, Classical Training*: This strategy represents the classical approach usually used in BCIs, where the training phase represents a certain amount of time in the entire experiment for each different user. Here, we tried to reduce as much as possible this training time. We did not use the knowledge of previous subjects to train the classifier. For each subject, we used the labeled examples as they occurred in the experiment. With each new example coming at time t , we trained a classifier with all the examples acquired in the time $[1 - t]$. The test set remained fixed, representing the last 600 examples of the experiment.

Since we wanted an LDA classifier to learn from the examples rapidly, we only used the channels Fz and FCz within the time window $[200 - 800]$ ms. Then, the data was sub-sampled by a factor of 4, having a feature vector of dimension 78. No further feature processing was performed.

2) *CS II, Inter-Subject Calibration*: This strategy was performed (i) on each user separately using a ten-fold cross validation strategy, and (ii) with a classifier trained with three users and tested with the remaining one (without performing any recalibration on the test subject). All the combinations were tested.

Due to the fronto-central nature of the error potentials [11], the following eight fronto-central channels were chosen: Fz, FC1, FC2, FCz, Cz, CP1, CP2 and CPz. The time window selected was $[200 - 800]$ ms. Again, the data was sub-sampled by a factor of 4, having a feature vector of 312 features. Then, the filtering method presented in II-B was applied to select the best f features, which were fed to the classifier. To study the performance as a function of the number of features used, f varied from 1 to 312 features.

3) *CS III, Adaptive Calibration*: In the adaptive strategy, we first used the knowledge obtained from trials of the previous subjects to train an initial classifier (as in CS II). Then, each of the labeled examples of the new subject was used to adapt the initial classifier. The test set and the features used were the same as the ones used in the CS I to perform further comparisons.

The adaptation performed was based on LDA, and was defined in [8]. For each new ERP belonging to class i , the

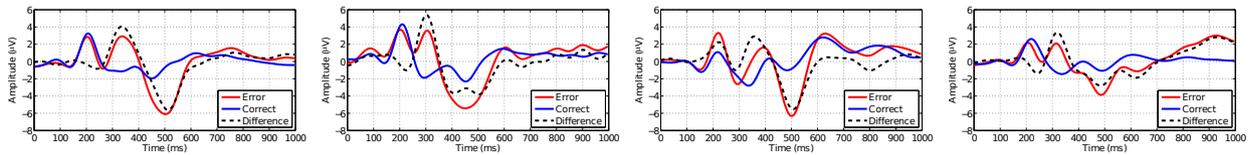


Fig. 2. Grand Averages of error, correct and error minus correct responses for each subject (from left to right, subjects 1-4) at channel FCz.

feature vector \mathbf{x}_t was created, and the corresponding μ_i was updated using an exponential moving average as follows:

$$\mu_i(t) = (1 - \alpha) \cdot \mu_i(t - 1) + \alpha \cdot \mathbf{x}_t, \quad (4)$$

where α represents the update coefficient. To improve the convergence of the method, α was initially fixed to 0.1, decreasing its value by a factor of 0.99 with each new example acquired. Finally, the parameters μ , \mathbf{w} and b were updated.

III. RESULTS

The results are divided into two categories. First, we analyze the similarity of the signals obtained across the different subjects. Then, we present the single-trial classification accuracies obtained from each of the strategies proposed.

A. Similarity analysis

1) *Grand averages*: Figure 2 shows the grand averages obtained for each of the four participants at channel FCz. The averages showed the typical time evolution of these potentials present in previous studies [9], [12]. Concretely, the difference between error and correct responses was represented by a negative peak at approximately 200ms, followed by a positive peak at 300ms, and finally a second negative peak. In general, correct responses seemed to have a roughly stable shape between different users, whereas error responses presented different shapes between the users. To corroborate these conclusions, we computed the Bhattacharyya distance of the distributions for each pair of subjects, separately on error and correct single-trial responses. This metric measures the similarity in terms of distance between two probability distributions. The results are shown on Table I. For all the users, the correct responses always showed a lower distance than the error responses. This indicates that correct responses had smaller changes between users than error responses.

TABLE I
BHATTACHARYYA DISTANCES

	Correct Responses				Error Responses				
	S1	S2	S3	S4	S1	S2	S3	S4	
S1	—	11.36	5.19	2.86	S1	—	12.39	6.23	6.74
S2	11.36	—	6.03	4.37	S2	12.39	—	8.51	4.87
S3	5.19	6.03	—	2.23	S3	6.23	8.51	—	6.04
S4	2.86	4.37	2.23	—	S4	6.74	4.87	6.04	—

2) *Features*: As suggested in [10], there could be user-independent features able to separate the two classes, but the use of other features specific to each user could improve the separability. Here, we corroborated these assumptions analyzing the similarity of the best feature obtained from

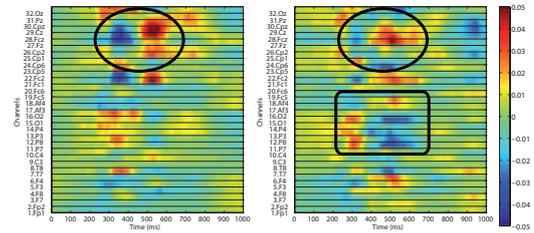


Fig. 3. Weights of the best feature for three concatenated subjects (left) and the remaining user (right).

three subjects altogether, compared with the best feature obtained with the remaining subject. For this analysis, we used the method explained on II-B. All the channels and time window were chosen, and the weights of the best feature were reshaped and plotted as a color-encoded image. Figure 3 (left) shows the weights from the best feature, obtained from subjects 1-3. Figure 3 (right) shows the corresponding weights for subject 4. The other combinations presented similar results, and thus are not shown. The two features showed similar time-channel combinations. For instance, during the time [200 – 500]ms on fronto-central channels (circled area in Fig.3), there was a negative activity followed by a positive one. However, the user-specific coefficients had different activations in terms of absolute value. Furthermore, the user-specific weights showed a higher number of time-channel combinations in other areas (for instance, squared zone of Fig.3 (right)). Thus, the performance obtained with a purely user-independent classifier could be not optimal.

B. Single-trial classification

1) *CS I*: Figure 4 shows the results obtained with the classical training approach averaged for all the subjects. As can be seen, the classical approach almost converged to the maximum accuracy (roughly 76%) with 600 examples. This is equivalent to approximately 40 minutes of EEG recording in the experimental protocol designed. This time is similar to the duration of the training phase on previous EEG-based BCIs [2].

2) *CS II*: Figure 4 shows the results obtained averaged for all the subjects. The results showed that a classifier trained separately for each user achieved higher accuracies than one trained with three users, with a decrease of approximately 8%. This could be due to the existence of user-specific features that helped in the separation of the two classes, as stated previously. However, we obtained high performances with the user-independent classifier, of roughly 70%. Furthermore, with this strategy we completely removed the calibration phase. One interesting fact is that the LDA classifier suffered from an overfitting effect: the performance decreased after

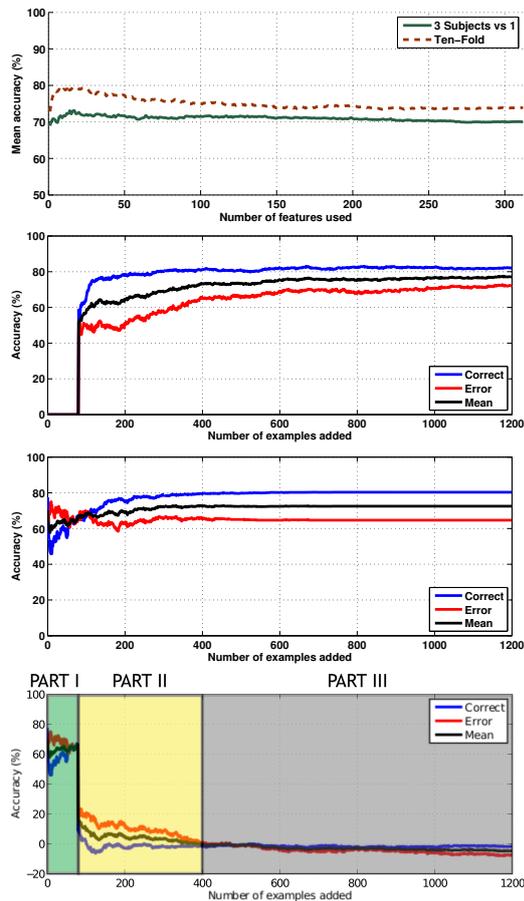


Fig. 4. From top to bottom: accuracy of the inter-subject training approach (CS II) as a function of the number of features; accuracies of the classical training (CS I), adaptive (CS III), and adaptive minus classical approaches as a function of examples added.

adding more than approximately 20 features. Thus, it is advisable to use a relatively low number of features.

3) *CS III*: Figure 4 shows the results obtained with the adaptive calibration approach averaged for all the subjects. Contrary to the results of the classical training approach, the adaptive approach was able to obtain high accuracies from the beginning of the experiment. To analyze the effect of the adaptive approach compared with the classical one, the adaptive minus classical approach was computed, and is also shown on Figure 4. This plot can be separated in three Parts: during Part I, the classical approach was not able to estimate the parameters due to the low number of examples, whereas the adaptive approach obtained high accuracies; during Part II, the classical approach was able to train the classifier. However, the performance of the adaptive classifier was still better. Finally, during Part III the classical approach still increased the performance, while the adaptive remained with the same accuracy. Thus, the adaptive approach can help on the first two Parts. In our case, the duration of these two Parts corresponded to 18 ± 10 min.

As a conclusion, the adaptive approach can help in alleviating the training phase, usually represented by Parts I and II. Furthermore, instead of having a training phase as such, the experiment could start from the beginning without

an explicit training phase, while learning the classifier in a transparent manner without the user noticing it.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the analysis of error potentials with respect to the features, and the inter-user variability existent in the signals. Additionally, two techniques to reduce the calibration time have been presented.

The feature analysis suggested that user-independent features exist, and they could be used to detect signals from other users with a high accuracy (70%). Additionally, the use of an adaptive classifier can improve this accuracy. Furthermore, the adaptive approach can make the calibration phase shorter and transparent to the user. Thus, the user would be able to start the control of the device without a specific training phase.

As future work, we are investigating the use of non-supervised adaptive classifiers to remove the label information when a new user comes. Additionally, the use of a large number of subjects could help to improve the inter-subject classifier, since the features found could be better estimated.

REFERENCES

- [1] J.d.R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Noninvasive brain-actuated control of a mobile robot by human EEG," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1026–1033, June 2004.
- [2] I. Iturrate, J. Antelis, A. Kübler, and J. Minguez, "Non-invasive brain-actuated wheelchair based on a P300 neurophysiological protocol and automated navigation," *IEEE Transactions on Robotics*, vol. 25, no. 3, pp. 614–627, 2009.
- [3] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–91, June 2002.
- [4] J.d.R. Millán, J. Mourino, M. Franze, F. Cincotti, M. Varsta, J. Heikkonen, and F. Babiloni, "A local neural classifier for the recognition of eeg patterns associated to mental tasks," *IEEE Transactions on Neural Networks*, vol. 13, pp. 678–686, May 2002.
- [5] J.d.R. Millán, "On the need for on-line learning in brain-computer interfaces," in *International Joint Conference on Neural Networks*, 2004, vol. 4, pp. 2877–2882.
- [6] J. Polich, "On the relationship between EEG and P300: Individual differences, aging, and ultradian rhythms," *International journal of psychophysiology*, vol. 26, no. 1-3, pp. 299–317, 1997.
- [7] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K.R. Müller, "Toward unsupervised adaptation of LDA for brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2011.
- [8] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain Topography*, vol. 23, no. 2, pp. 194–198, 2010.
- [9] R. Chavarriaga and J.d.R. Millán, "Learning from EEG error-related potentials in noninvasive brain-computer interfaces," *IEEE Trans. on Neural Syst. and Rehab. Eng.*, vol. 18, no. 4, pp. 381–388, 2010.
- [10] S. Lu, C. Guan, and H. Zhang, "Unsupervised brain computer interface based on intersubject information and online adaptation," *IEEE Trans. on Neural Syst. and Rehab. Eng.*, vol. 17, no. 2, pp. 135–145, 2009.
- [11] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: A tutorial," *Biological Psychology*, vol. 51, pp. 87–107, 2000.
- [12] I. Iturrate, L. Montesano, and J. Minguez, "Single trial recognition of error-related potentials during observation of robot operation," in *Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010.
- [13] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, May 2004.
- [14] I. Iturrate, L. Montesano, R. Chavarriaga, J.d.R. Millán, and J. Minguez, "Spatio-temporal filtering for EEG error related potentials," in *5th Int. Brain-Computer Interface Conference*, 2011.