Latency Correction of Event-Related Potentials Between Different Experimental Protocols

I Iturrate[1,2], R Chavarriaga[3], L Montesano[1,2], J Minguez[1,2,4] and JdR Millán[3]

1 Instituto de Investigación en Ingeniería de Aragón (I3A), Edificio I+D+i, Mariano Esquillor, 50018 Zaragoza, Spain

2 Departamento de Informática e Ingeniería de Sistemas (DIIS), Universidad de Zaragoza, Maria de Luna 1, 50018 Zaragoza, Spain

3 Chair in Non-Invasive Brain-Machine Interface (CNBI), Center for Neuroprosthetics and Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Switzerland

4 Bit&Brain Technologies SL, 50018 Zaragoza, Spain

E-mail: iturrate@unizar.es, ricardo.chavarriaga@epfl.ch, montesano@unizar.es, jminguez@unizar.es, jose.millan@epfl.ch

Abstract. Objective: A fundamental issue in EEG event-related potentials (ERPs) studies is the amount of data required to have an accurate ERP model. This also impacts the time required to train a classifier for a brain-computer interface (BCI). This issue is mainly due to the poor signal-to-noise ratio, and to the large fluctuations of the EEG caused by several sources of variability. One of these sources is directly related to the experimental protocol or application designed, and may affect to amplitude or latency variations. This usually prevents BCI classifiers to generalize among different experimental protocols. In this work, we analyze the effect of the amplitude and the latency variations among different experimental protocols based on the same type of ERP. **Approach:** We present a method to analyze and compensate for the latency variations in BCI applications. The algorithm has been tested on two widely used ERPs (P300 and observation error potentials), in three experimental protocols in each case. We report the ERP analysis and single-trial classification. **Results and significance:** The results obtained show that the designed experimental protocols significantly affect the latency of the recorded potentials but not the amplitudes; and how the use of latency-corrected data can be used to generalize the BCIs, reducing this way the calibration time when facing a new experimental protocol.

Submitted to: *J. Neural Eng.*

1. Introduction

Event-related potentials (ERPs) reflect brain responses to external events [1], and are modeled as the average of multiple trials of time-locked scalp EEG signals characterized by its polarity, latency, and spatial localization [2]. These characteristics have been used to assess psychiatric and neurological conditions [2, 3], or even for the understanding of brain processes such as attention, or error processing [4, 5, 6, 7]. Furthermore, ERPs have also been used for brain-computer interfacing (BCIs, see [8] for a review), where the ERP model is trained and used to translate the EEG signals into control commands to operate different devices such as text spellers, mobile robots, or wheelchairs [9, 10].

Characterization of ERPs require the acquisition of enough trials to build a reliable model represented by their grand averages [1]. This is due to the poor signal-to-noise ratio of the EEG as well as several sources of variability that may affect the amplitude or the latency of the ERP components. For instance, the early ERP components (appearing within 200 ms from the stimulus presentation, e.g. visually-evoked potentials, VEP) are affected by application-specific factors such as the spatial attention [11] or the stimuli contrast [1]; as well as user-specific factors such as arousal or valence [4]. In turn, late ERP components (occurring later than 200 ms) are affected by application-specific factors such as the probability of occurrence of the expected stimulus [1] or the inter-stimulus interval [12]; user-specific factors such as the age and the cognitive capabilities [6, 7]; or application- and user-specific variability such as the stimulus evaluation time (i.e., the amount of time required to perceive and categorize a stimulus) [13, 1].

Typically, experiments are designed in a well-controlled manner to reduce the ERP variability. In consequence, it is not clear whether the obtained model also reflects the same neural phenomena under different conditions. This is of particular importance for practical BCI applications where decoding algorithms are expected to keep their performance level irrespective of external factors. Moreover, BCIs often exploit the same brain processes in different applications with different associated stimuli, feedback modality or controlled device (e.g., see [14, 15, 16, 17] for different applications based on error-related processing). In the ideal case, these systems should be able to generalize across different operating BCIs independently of the device that is being controlled. In practice, however, there is a need for training a model for each new experimental protocol or session, which is a time-consuming operation and a major issue when deploying BCIs out of the lab. To address this issue, previous researches have tried to reduce this calibration time either by using adaptive classifiers [18, 15], or by initializing the model with data from a pool of subjects [19, 20].

Although previous studies have described the effect of variations in the ERP amplitudes [16] and latencies [21] within the same BCI experimental protocol, the effect of these variations among different protocols remains unclear. We hypothesize that it could be possible to build or re-adjust models that compensate for these variations by using information from previous experimental protocols, thus enabling generalization of existing BCI decoders to different protocols or applications. The main idea is depicted

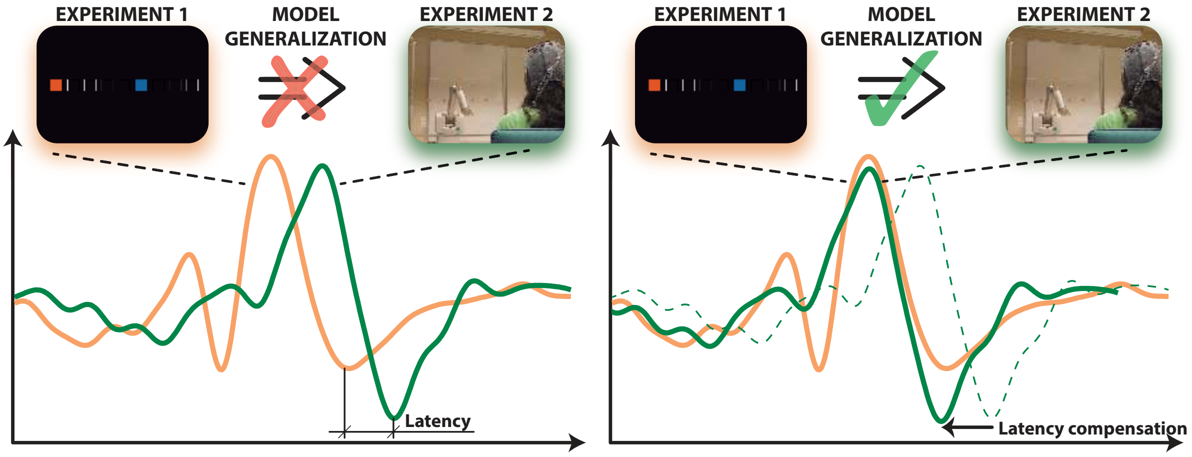

Figure 1. (Left) Example of the latency between two grand averaged event-related potentials elicited from different experimental protocols. Such difference prevents from having classifiers that generalize among protocols. (Right) By estimating and removing the latency of the two ERPs, the classifier would be able to work under different experimental protocols.

in Figure 1 (Left), where two experimental protocols elicit the same ERP with similar waveforms and amplitudes but different latencies. If we could estimate the latency variations between the two experimental protocols, the model of one experimental protocol could be used in the new protocol after compensating for the latency shift (see Figure 1 Right).

In this paper, we analyze the effect of ERP amplitude and latency variations among different experimental protocols based on the same cognitive process. We also present a method to analyze and compensate for the latency variations in BCI applications. Two widely used signals were analyzed: the P300 evoked potentials [9, 1, 10] and the observation error-related potentials (ErrP) [5, 14, 16]. For each kind of ERP, three different experimental protocols with different levels of difficulty were designed. The latencies between protocols were studied from two points of view: the characteristics of the ERPs and the single-trial classification. The results illustrate (*i*) how the designed experimental protocols significantly affect the latency of the recorded potentials but not the amplitudes, and (*ii*) how the use of latency-corrected data allows for the generalization of BCI decoders, reducing in this way the calibration time when facing a new experimental protocol. This work extends our previous work [22] with a more robust technique to compensate the latencies and shows its application to ERPs of different nature.

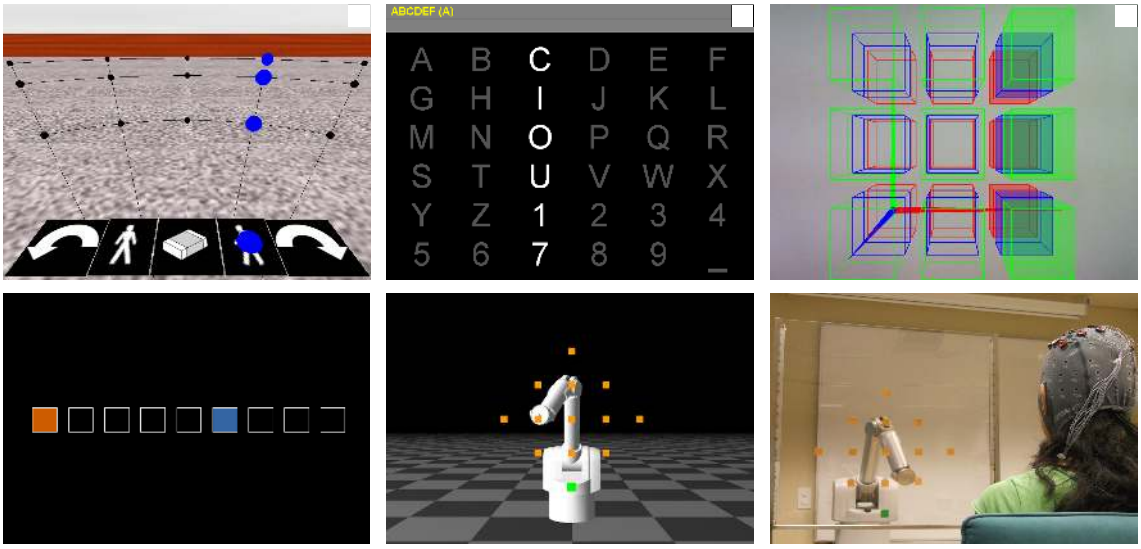

Figure 2. Experiments performed for the (Top) P300 potentials and (Bottom) observation error potentials (from left to right: experiments 1 to 3).

2. Experimental methods

We focus on two types of ERPs: the P300 evoked potentials and the observation error-related potentials (ErrP). For each of these signals, three types of experimental protocols were designed (i.e., three different ways of evoking the P300 and the ErrPs).

2.1. Data recording and experimental setup

The recordings and signal processing were made following previous studies [10, 23]. EEG was recorded by means of a gUSBAmp amplifier (*gTec medical engineering*, Schiedelberg, Austria) with 16 active electrodes, with the ground and reference placed on the forehead and left earlobe. Different montages were made for the P300 and ErrP protocols (see details below). EEG was digitized at 256 Hz, power-line notch filtered at 50 Hz, and zero-phase Butterworth band-pass filtered at $[1, 10]$ Hz. Participants were seated on a comfortable chair facing the visual displays of the protocols approximately one meter away. During all experiments participants were asked to restrict eye movements and blinks to specific resting periods.

2.1.1. P300 experimental protocols For these protocols we recorded EEG signals with the BCI2000 framework [24] from 16 active electrodes located at Fp1, Fp2, Fz, FC1, FCz, FC2, Cz, CP1, CPz, CP2, P3, Pz, P4, O1, Oz and O2 according to the 10-10 system and following previous studies [10]. Five participants (one female, mean age 27.80±2.49 years) took part in the study. We synchronized the onset of the visual stimuli with the EEG by means of an optical trigger placed on the monitor [25]. This removed latencies introduced by the protocol implementation and thus the latency variations across experiments were restricted to the user side [13].

Three experimental protocols were used to evoke the P300 potentials (Figure 2, Top), with different types of stimuli (with overall workloads of 39.10 ± 11.11, 42.45 ± 9.98, and 64.83 ± 18.23, estimated from six subjects using the NASA TLX). The stimulation process followed the oddball paradigm [9], where subsets of potential targets (e.g. an entire row or column) are sequentially highlighted in random order. The stimulus (row or column) remained highlighted for 125 ms on the screen, and the inter-stimulus interval was randomly set within the range $[1.7, 3.0]$ s. The participants were instructed to observe the stimulation process fixing their attention to a given target, and to count the times the target was highlighted while ignoring the other stimuli. All participants executed the experiments in the same order, each experiment lasting ≈ 1.5 hours and with a time between experiments of 1.10 ± 0.81 days.

Experiment 1, 2D Simulated Wheelchair (Figure 2 Left, Top) [10] The visual display showed a virtual environment with 20 possible targets to drive a wheelchair, located in 2D in a 4x5 matrix. For the stimulation process, the rows and columns were highlighted showing a blue dot over each possible target position. The probability of target appearance was 22%. For each subject, all possible target positions were recorded, obtaining 144 target (P300) and 720 non-target responses respectively.

Experiment 2, 2D Speller (Figure 2 Middle, Top) [9] The visual display showed a matrix of 36 possible letters to spell represented in 2D as a 6x6 matrix. The stimulation was made by highlighting the corresponding row or column. The probability of target appearance was 17%. For each subject, all possible target positions were recorded, obtaining 200 and 700 target and non-target responses respectively.

Experiment 3, 3D Augmented Reality Protocol (Figure 2 Right, Top) The display showed a gray background and 27 possible targets located in 3D in a 3x3x3 matrix. The stimulation was made by illuminating rows, columns, and depths. To facilitate the user's distinction among the three depths, each depth was illuminated with a different colour (green, blue or red). The probability of target appearance was 33%. For each subject, all possible target positions were recorded, obtaining 273 and 610 target and non-target responses respectively.

2.1.2. Error potentials experimental protocols We recorded ErrPs with a custom C++ framework using 16 active electrodes located at Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, and CP4 according to the 10-10 system and following previous studies suggesting that these signals are generated in fronto-central areas [23]. Six participants (one female, mean age 27.33 ± 2.73 years) took part in the study. In these experiments, the use of an optical trigger was not possible since one experiment involved a real robotic device instead of visual stimuli on the screen (see Experiment 3). Thus, latency variations could be originated by both the subject and the implementation (i.e. the amount of time of receiving and executing the delivered command).

The three experimental protocols designed to elicit error potentials (Figure 2, Bottom) had different setups and devices (with overall workloads of 35.50 ± 11.53, 53.50 ± 19.88, and 58.11 ± 16.47, estimated from six subjects using the NASA TLX), where in all cases the goal of the device was to reach a target from different starting points. The device executed random movements with approximately 30% probability of performing an erroneous movement. The time between two movements was randomly set within the range $[1.7, 4.0]$ s. The target position was randomly changed after 100 actions. The participants were instructed to observe the device movements and evaluate them as correct when there was progress towards the target position, and as incorrect otherwise. Each participant executed the experiments in the same order, each experiment lasting ≈ 2.5 hours and with a time between experiments of 17.58 ± 10.09 days.

Experiment 1, Virtual Moving Square (Figure 2 Left, Bottom) [16] The visual display showed a one-dimensional space with 9 possible positions (marked by a horizontal grid), a blue square (device) and a red square (target). The device could execute two discrete actions: move one position to the left or to the right. For each subject, the left- and right- most positions were tested as targets, and around 250 and 600 error and non-error potentials were recorded.

Experiment 2, Simulated Robotic Arm (Figure 2 Middle, Bottom) The display showed a simulated robotic arm (Barrett WAM) with 7 degrees of freedom (device) [26] moving within a two-dimensional space with 13 possible positions (marked in orange), and a target location (green square). The robot was situated behind the squares pointing at one position, and could perform four possible actions: moving one position to the left, right, up, or down. The robot actions were continuous, with each displacement lasting ≈ 500 ms. For each subject, the left-, right-, up- and down-most positions were tested as targets, and around 300 and 700 error and non-error potentials were recorded.

Experiment 3, Real Robotic Arm (Figure 2 Right, Bottom) This experiment followed the same configuration of Experiment 2 but using a real Barret WAM robotic arm (*Barret Technology Inc.*). The user was seated two meters away from the robot, and between them there was a transparent panel to mark the positions (the distance between two neighbor positions was 15 cm). For each subject, the left-, right-, up- and down-most positions were tested as targets, and around 300 and 700 error and non-error potentials were recorded.

2.2. Analysis of Event-Related Potentials

We assessed protocol-dependent variations in the latency and amplitude of the ERPs of each experimental protocol. First, the grand averaged signals were computed for each condition (target and non-target trials for the P300; error and correct trials for the ErrP), for the time window $[-200, 1000]$ ms, being 0 ms the stimulus/action onset. Following previous studies, we analyzed the activity over parietal areas from the target average [1] for the P300, and over fronto-central areas from the difference average (error minus correct averages) for the ErrPs [16]. A one-way within-subjects ANOVA was performed separately for each type of signal (P300 or ErrP), where the factor was the experiment (three levels corresponding to each experiment), and two dependent variables were tested: the peak amplitudes and the peak latencies. For the P300 experiments, the peak amplitudes and latencies were measured from the P3 component (most prominent positive peak) of the target average from the parieto-occipital channels. For the ErrP experiments, amplitudes and latencies were measured from the P3 and N4 components (most prominent positive and negative peaks) of the difference average from the fronto-central channels. When needed, the Geisser-Greenhouse correction was applied to data to assure sphericity [1]. Pairwise post-hoc tests (Bonferroni-corrected t-tests) were computed to determine the differences between pairs of experiments.

2.3. Estimation and evaluation of latencies among different protocols

The first goal is to estimate the temporal variations between two experimental protocols, which can be achieved using cross-correlation. Cross-correlation has been used in the past for the detection and analysis of brain signals with successful results [21, 27, 28]. In order to assure the best estimations, the input to the cross correlation (for each channel) were the grand averages of the condition of interest, with the time window narrowing to the event-related potential elicitation. For the P300 experimental protocols, the average ERP for target stimuli within the time window $[50, 400]$ ms was used; for the ErrP experimental protocols, the error average within the time window $[0, 500]$ ms was used. These windows were chosen following two premises: (*i*) the ERP components of interest were within the windows; and (*ii*) an r^2 discrimination analysis between conditions (i.e. targets vs. non-target, error vs. non-error) showed that the most significant differences were present in those windows. The cross-correlation outputs were the maximum correlation value of the two grand averages and the latency variation between them (i.e. the shift that yields the maximum cross-correlation).

We then assessed whether the main ERP change was due to the latency variation and whether this variation could be compensated for. To do so, the latency variation across two protocols was estimated as described above using all the available data. Let D_i and D_j be the datasets from two experiments E_i and E_j, we compensate for the variation by shifting the trials in D_i by the estimated latency shift between them, $d_{D_iD_j}$. Then, we computed the same ANOVA test for the peak latencies performed in subsection 2.2.

We performed a further analysis on how sensitive the latency estimation was with respect to (*i*) the number of trials used to compute the grand average for experiment j, and (*ii*) the channel used to perform the estimation. Assuming that data D_i from a

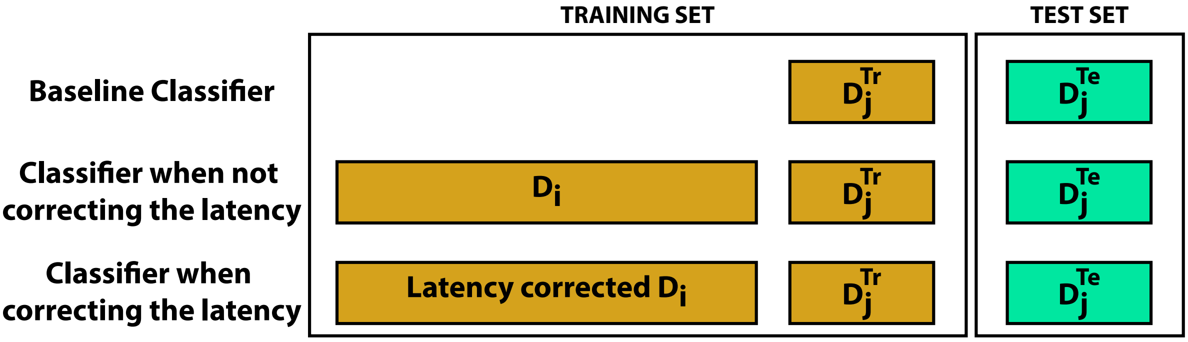

Figure 3. Training and testing datasets used for each classifier. For the baseline classifier, the classifier is trained with a subset D_j^{Tr} from experiment E_j. When reusing a previous experiment E_i, the whole dataset D_i is added to D_j^{Tr}. In the third case, the latency between D_i and D_j^{Tr} was estimated, and then D_i was corrected accordingly.

previous experiment E_i is available, we computed the latency variation using a training dataset D_j^{Tr} from the new experiment E_j ($D_j^{Tr} \subseteq D_j$). We assessed the estimation using different sizes of the training dataset (ranging from 10 to 200 trials with increments of 10). For each size, we perform 10 repetitions and report the average of the maximum cross-correlation value, $\max(C_{D_i}^{D_j^{Tr}})$, and the average latency variation, $d_{D_i D_j^{Tr}}$. In each repetition the training subset D_j^{Tr} was randomly drawn from D_j, keeping the proportion of target/non-target and error/correct trial. The analysis was performed independently for each recorded channel.

The latency variations were computed in a pair-wise manner among the three experiments for each of the signals of interest. The combinations of experiments tested were E_1E_2, E_1E_3, and E_2E_3 for both the P300 and the ErrPs. For each pair of experiments a within-subjects two-way ANOVA (factors: number of trials and repetitions) was performed on the latency estimations. The ANOVA results served to study the latency variations by the number-of-trials main effect, to determine whether the amount of trials used from experiment E_j led to different latency estimations; and by the number of trials x repetitions interaction, to determine whether different data from a fixed number of trials affected the latency estimations.

As a sanity check, we also evaluated the method by computing the latency variation among datasets from the same experiment ($d_{D_i^1 D_i^2}$), with the two datasets D_i^1 and D_i^2 mutually exclusive. Therefore, this baseline latency computation should give correlations close to one for latencies near to 0 ms.

2.4. Single-trial classification of latency-corrected ERPs

The objective of the single-trial classification study was to determine whether it is possible to reduce the calibration time of a new experiment by re-using latency-corrected data from a previous experiment. The study used the same combination of experiments (E_iE_j) detailed in the previous section. To evaluate the benefit of reusing data and correcting the latency, three classifiers were learned each with a different training dataset (see Figure 3).

The first case, denoted baseline classifier, followed the standard calibration approach of current BCIs, where the classifier for experiment E_j was trained using only a subset D_j^{Tr} of the data. For the second classifier, the training data was formed by the whole dataset D_i from a previous experiment E_i and the training data from the new experiment D_j^{Tr}. The third classifier utilized the same training sets as the second one, but used the latency estimated between D_i and D_j^{Tr} to compensate the delay between experiments E_i and E_j. Recall that the latency is estimated and corrected for each channel separately as described in subsection 2.3. This correction was performed by shifting all the trials from D_i accordingly. All the single-trial analysis (latency estimation, feature extraction and training the classifier) was done using only the corresponding training data. Results were obtained using the same test data for all cases. As in the previous subsection, we performed ten repetitions of this process randomly drawing D_j^{Tr} from D_j.

2.4.1. Feature extraction and classification Feature extraction was based on a spatio-temporal filter [29]. The filter input was a dataset with labeled trials and worked as follows: Firstly, the EEG data were common-average-reference (CAR) filtered and downsampled to 64 Hz. For each trial, the features were extracted using a combination of channels and time points. For the P300, eight centro-parietal and occipital channels (Cz, CPz, P3, Pz, P4, O1, Oz, and O2) were used within a time window of $[100, 700]$ ms. For the ErrP, eight fronto-central channels (Fz, FC1, FCz, FC2, C1, Cz, C2, and CPz) were used within a time window of $[200, 800]$ ms. For both cases, this resulted in a feature vector of 312 features per trial. Then, the features were normalized, and decorrelated using PCA retaining 95% of the explained variance, leading to an average of 45 ± 10 features. Single-trial classification was carried out using a linear discriminant (LDA) [30].

2.4.2. Analysis of the single-trial classification We compared the accuracies of the three different classifiers for a fixed dataset of the new experiment, namely D_j^{Te}, composed of 400 trials (see Figure 3). As in the delay estimation analysis, the size of the training data D_j^{Tr} was varied to assess the accuracy of the classifier for different calibration times. Additionally, the performance of these three classifiers was compared with the ten-fold cross-validation (CV) performance obtained using all the data D_j from E_j.

For each pair of experiments (E_1E_2, E_1E_3, and E_2E_3), the receiver operating characteristic (ROC) curve was computed [31], and we compared the area under the curve (AUC) obtained for each case and classifier. To assess statistical differences among the classification results, two-tailed paired t-tests were computed with the p-values adjusted with the false discovery rate (FDR) procedure [32].

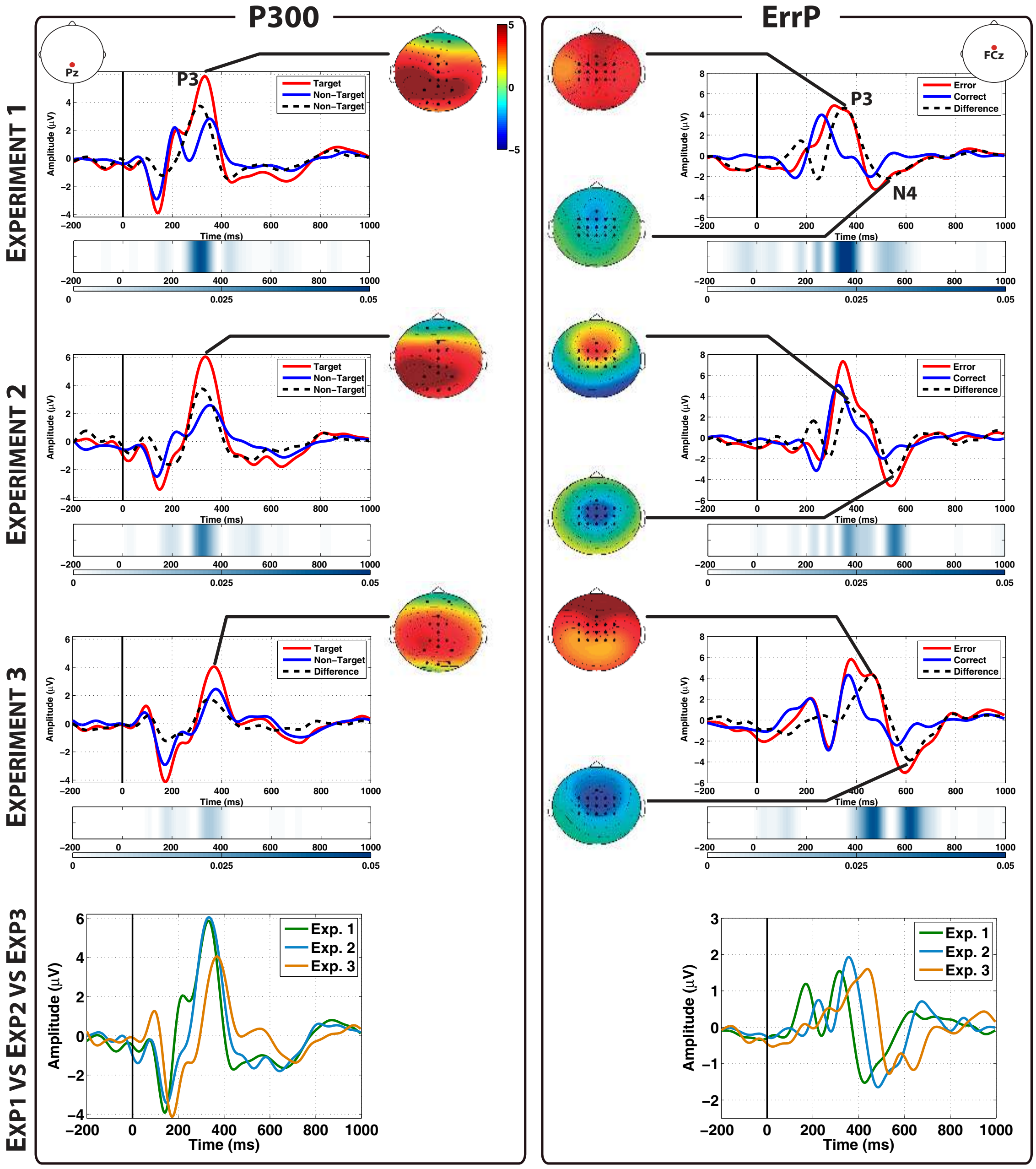

Figure 4. Grand averages and r^2 (lower part) of each experiment for the (Left) P300 potentials at channel Pz, and (Right) error potentials at channel FCz. Time 0 ms indicates when the stimulus was presented on the screen (P300), or when the device started the action (ErrPs). For the P300, the topographic interpolation of the most prominent positive peak of the target average is shown. For the ErrPs, the topographic interpolation of the most prominent positive and negative peaks of the difference average are shown. The bottom plot shows the GA of the three experiments for the target (P300) and difference (ErrP) conditions.

3. Results

3.1. Analysis of Event-Related Potentials

Figure 4 shows the ERP grand averages of all experiments. In the P300 experiments, as in previous studies [9, 1, 10], a clear sharp positive peak (P3) appears on parietal channels after presentation of the target stimuli. For the ErrP experiments, the difference grand averages (error minus correct) are also consistent with the literature [16], with two early positive and negative peaks in fronto-central sites, followed by two larger positive and negative peaks (P3 and N4).

Regarding the P300 experimental protocols, the amplitude of the P3 component showed no statistical differences among the three experiments ($p = 0.123$). In contrast, its latency does exhibit statistical differences ($F_{(2,8)} = 22.924, p = 0.0005$). Post-hoc tests revealed significant differences between experiments 2 and 3 ($p = 0.032$), and between experiments 1 and 3 ($p = 0.01$), but not between experiments 1 and 2 ($p = 1.0$).

Similarly, no differences were found for the P3 and N4 amplitudes of the ErrPs ($p = 0.510$ and $p = 0.391$ respectively). Interestingly, significant differences were found on the latencies of both the P3 ($F_{(2,10)} = 29.422, p = 0.00006$) and the N4 component ($F_{(2,10)} = 6.979, p = 0.013$). For the former, post-hoc tests showed significant differences between experiments 1 and 2 ($p = 0.018$), and between experiments 1 and 3 ($p = 0.003$), and nearly significant differences between experiments 2 and 3 ($p = 0.053$). For the N4 component, there were significant differences between experiments 1 and 3 ($p = 0.006$), but not between experiments 1 and 2 ($p = 0.472$) nor between experiments 2 and 3 ($p = 0.492$). Thus, the main differences on the elicited ERPs across the experiments were due to latency variations of the components, while the amplitudes remained similar.

3.2. Analysis of latency estimations

The ANOVA analysis yielded no significant differences in latency after performing the correction for the P300 ($p = 0.12$ for the P3 component), nor for the ErrP experiments ($p = 0.67$ and $p = 0.17$ for the P3 and N4 components, respectively). Thus, the latency correction algorithm successfully removed the latency variations among experiments.

Figures 5 and 6 (Top) show the maximum correlation (see section 2.3) for all electrodes when different numbers of trials from E_j are used. Unsurprisingly, correlation values increase until they converge to an upper value as more trials are used to compute the grand average. ERPs elicited in the P300 experiments (Figure 5, Top) show high correlation (≥ 0.8) in parieto-occipital channels when more than 50 trials are used. In turn, the ErrPs (Figure 6, Top) required at least 100 trials to yield correlation values higher than 0.8, always over fronto-central channels. These locations, as for the P300, agree with the locations reported as more discriminant for these phenomena.

When we computed the correlation using data from the same experiment, we obtained correlations above 0.8 when more than 40 and 70 trials were used (P300 and ErrP respectively). Thus, both cases needed a number of trials to reach high correlations

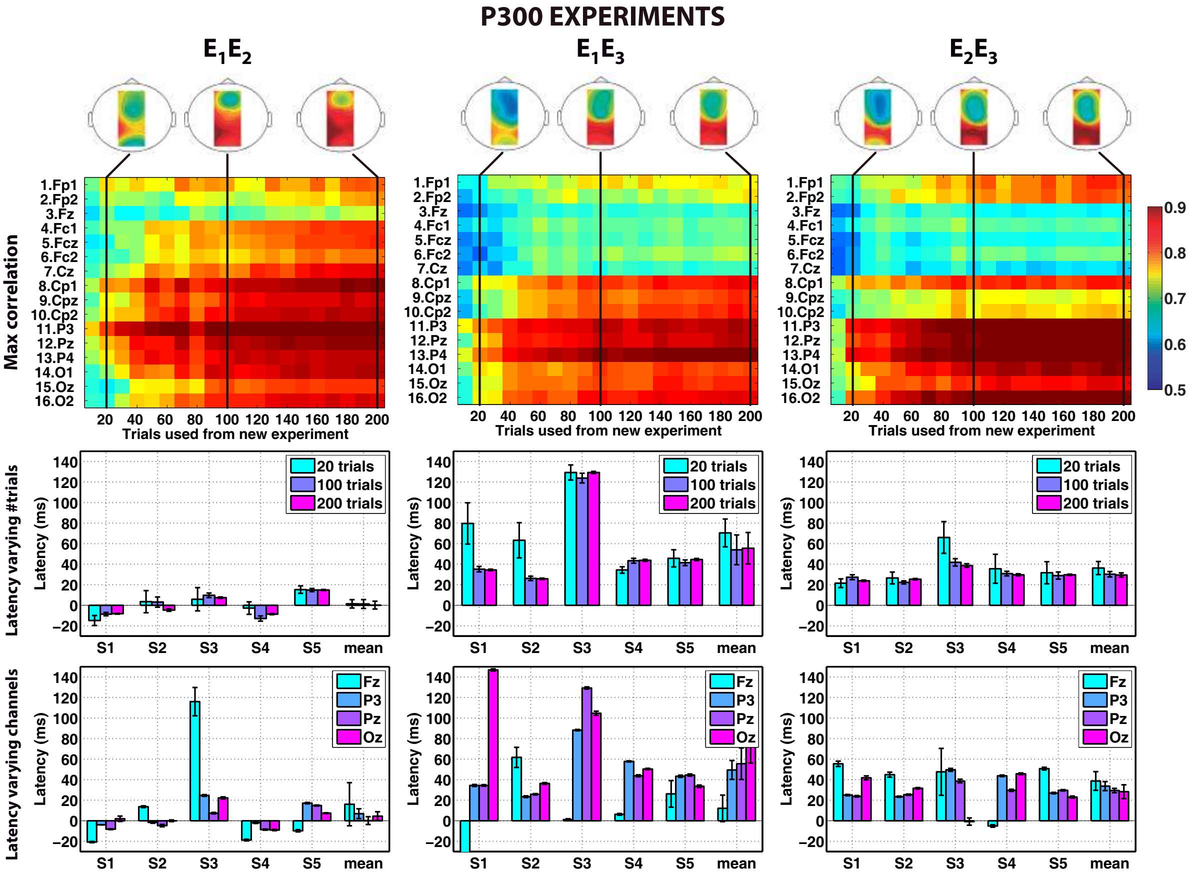

Figure 5. Latency results computed for each pair of P300 experiments E_iE_j (from left to right, E_1E_2, E_1E_3, and E_2E_3). For each pair of experiments, the results represent: (Top) Colour encoded image of the maximum correlation values (averaged for all subjects), when varying the number of trials used from D_j^{Tr} (x-axis) and the channel used for the latency computation (y-axis). The topographic interpolation of the correlation values is shown when using 20, 100, and 200 trials from D_j^{Tr} (for the sake of simplicity, the topographic plot is shown only within the field of the recorded channels). (Middle) Mean ± SEM latency estimations (in ms) of each subject, and subject-wise average latency for channel Pz while varying the number of trials used (20, 100 and 200 trials), and (Bottom) Mean ± SEM latency estimations (in ms) of each subject, and subject-wise average latency for 200 trials while varying the channels (Fz, P3, Pz and Oz). Figure is best viewed in colour.

comparable to the generalization cases.

Figures 5 and 6 (Middle) show the latency values of each subject computed for different number of trials in D_j^{Tr} (20, 100 and 200 trials). We show the latency calculation for channels Pz and FCz for the P300 and ErrP experiments respectively, since they had high correlation values and are commonly used for studying these signals [1, 16]. For the P300 experiments (Figure 5, Middle), the baseline latencies (i.e. computed on the same experiment) after 200 trials from D_i^{Tr} were -1.17 ± 14.01 ms, -3.90 ± 7.69 ms, and -3.13 ± 13.87 ms for experiments 1 to 3 respectively. The latency between E_1 and E_2 using 200 trials was of 0.16 ± 9.36 ms. This agrees with

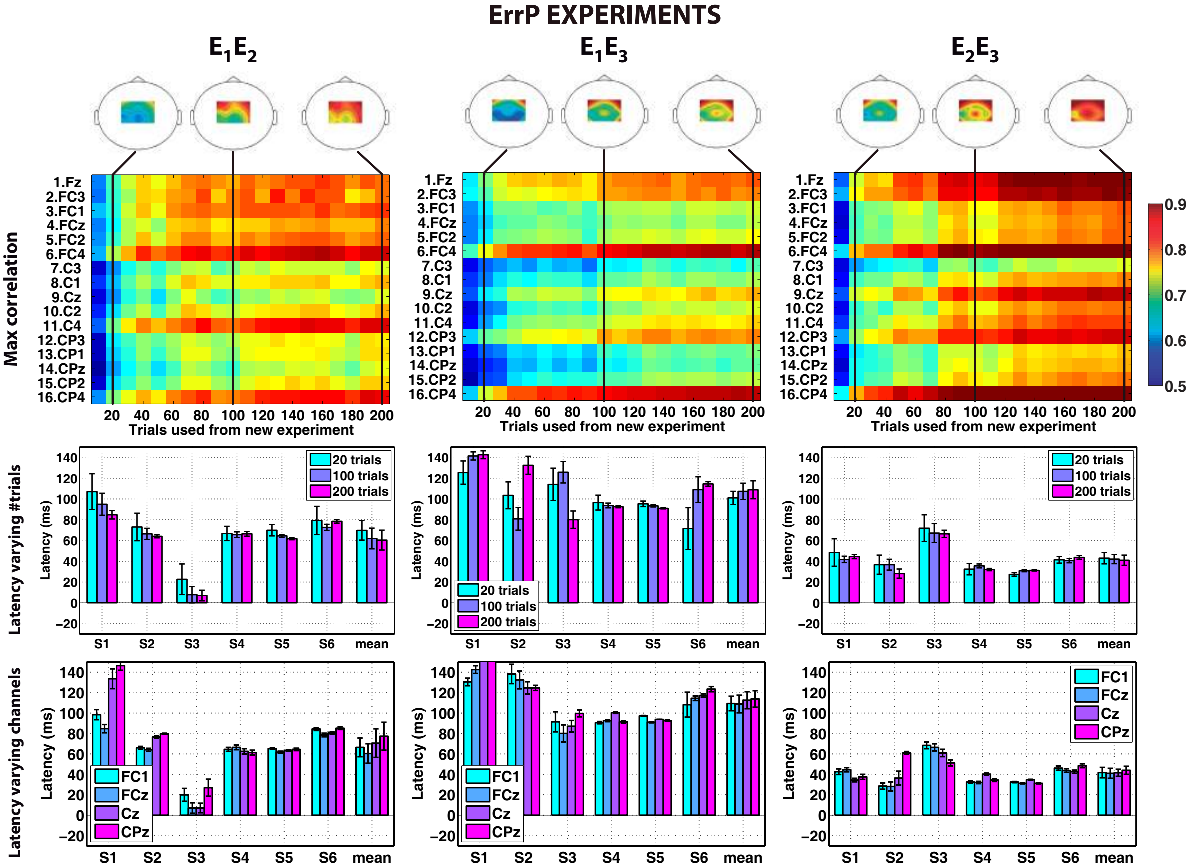

Figure 6. Latency results computed for each pair of ErrP experiments E_iE_j. (Top) Maximum correlation values. (Middle) Latency estimations for channel FCz while varying the number of trials (20, 100 and 200 trials) and (Bottom) Latency estimations for 200 trials while varying the channels (FC1, FCz, Cz, and CPz).

the previous results, where no statistical differences in the latencies were found between these experiments (c.f. Section 3.1). For the E_1E_3 and E_2E_3 cases, larger latencies were estimated (on average 55.54 ± 37.51 and 29.45 ± 5.16 ms, respectively). No statistical differences were found in the computed latencies as the number of trials varied ($p > 0.05$ for the three combinations of experiments). Similarly, no significant interactions between the number of trials and repetitions was found ($p > 0.05$). These results suggest that the latency estimation is rather robust to the number of trials used for their computation, and that the specific trials used (i.e. repetition) did not affect the latencies obtained.

For ErrPs (Figure 6, Middle), the baseline latencies after 200 trials were 5.40 ± 5.62, 12.96 ± 21.77, and 2.02 ± 4.80 ms for experiments 1 to 3. On the other hand, the latency variations across experiments were larger than those obtained for the P300: 60.42 ± 25.24, 108.85 ± 22.86 and 41.02 ± 12.95 ms for the E_1E_2, E_1E_3, and E_2E_3 pair of experiments. Larger inter-subject variability was also observed. There were statistical differences in the latency computation as the number of trials increased for the E_1E_2 and E_2E_3 cases ($F_{(19,95)} = 3.329, p = 0.0001$, and $F_{(19,95)} = 2.249, p = 0.005$, respectively), but not for the E_1E_3 ($p > 0.4$). On the other hand, no significant interactions between number of trials and repetitions were found for any case ($p > 0.05$). This indicates that the latency estimation was robust to the trials used. However, the latency estimation was affected by the number of trials used from E_j.

Figures 5 and 6 (Bottom) show the latency values of each subject computed for different channels. The number of trials remained fixed to 200. For the P300 experiments (Figure 5, Bottom), using frontal channels (e.g. Fz in the plot) for the latency calculation led to different results and higher standard deviations than using parietal channels (e.g. P3 and Pz). Regarding the ErrP experiments (Figure 6, Bottom), the latency estimations were more uniform across channels. Nonetheless, higher standard deviations and lower correlation values were obtained when using parietal channels, except for the E_2E_3 case, where similar results were obtained.

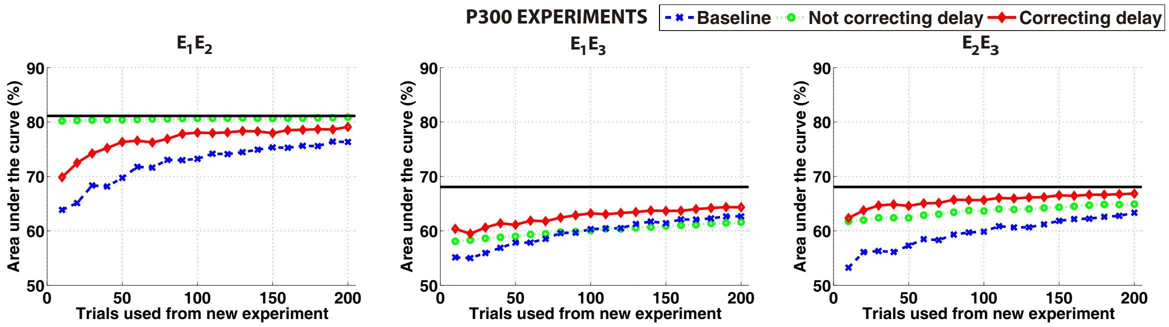

Figure 7. Mean values of the area under the curve (AUC) when correcting the latency from E_1E_2, E_1E_3 and E_2E_3 for the P300 experiments. The x-axis represents the number of trials of the training dataset D_j^{Tr}. Blue-dashed, green-dotted and red-solid lines represent, respectively, the results for the baseline classifier, the classifier trained when not correcting the latency, and the classifier trained when correcting the latency. Horizontal black lines mark the ten-fold cross-validation AUC of the E_j experiment.

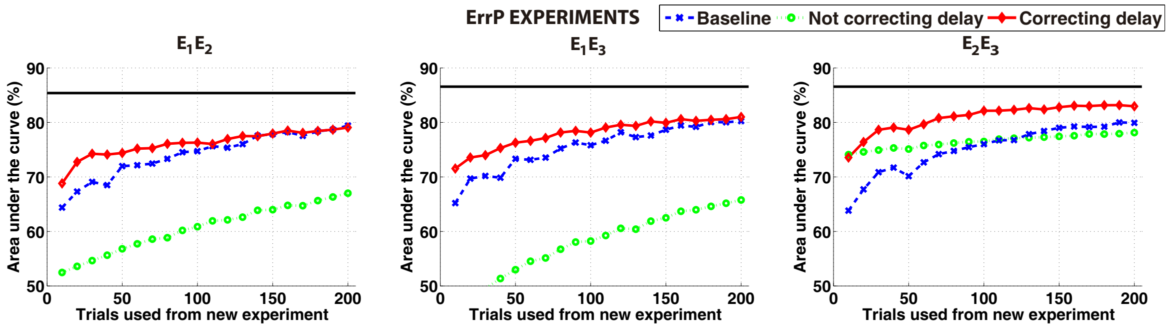

Figure 8. Mean AUC when correcting the latency from E_1E_2, E_1E_3 and E_2E_3 for the ErrP experiments.

3.3. Single-Trial classification of latency-corrected ERPs

3.3.1. P300 potentials Figure 7 shows the mean area under the curve (AUC) for all experiments and tested conditions (see Figure 3). In the E_1E_2 case the AUC of the baseline classifier (i.e. trained only with data from the new experiment) increased as more examples were added, reaching 76.33% after 200 trials. In contrast, using data from the previous experiment (E_1) significantly improved (two-tailed paired t-test, $p < 0.001$) the AUC, both when correcting the latency (reaching 79.10% after 200 trials) and when not correcting the latency (80.87% after 200 trials). In these cases, only 10 trials from E_2 were enough to improve the AUCs with respect to the baseline classifier. Additionally, these two classifiers had better AUC than the ten-fold CV with more than 50 trials from E_2. Thus, re-using data from a previous experiment allowed for an improvement both in the classifier AUC and calibration time. However, at least 110 trials were required for the latency correction method to perform similarly to the no correction approach, seemingly due to errors in the latency estimation.

Compared to the previous case, going from E_1 to E_3 resulted in lower AUCs for all types of classifiers (c.f. Figure 7 central column), always lower than the CV AUCs. After 200 trials the AUCs were of 62.67%, 61.60% and 64.34% for the baseline, not correcting latency and correcting latency classifiers, respectively. The AUC when correcting the latency was significantly better than the baseline ($p < 0.05$). On the other hand, re-using data where the latency was not corrected did not significantly improve the baseline AUC ($p > 0.1$).

In the last case (E_2E_3, c.f. Figure 7 right), the latency correction mechanism yielded significantly higher AUCs ($p < 0.05$) than the baseline or no correction approaches (66.84%, 63.34%, and 64.91% after 200 trials, respectively), converging to the AUC of the 10-fold CV. These differences appeared even when a small number of trials were available. Thus, the use of latency-corrected data allowed for a significant improvement in the AUCs.

3.3.2. Error potentials Results for the ErrP protocols are shown in Figure 8. In the first case, E_1E_2, the AUC of the baseline classifier reached 79.42% after using 200 trials for training. The latency-corrected classifier showed a peak performance of 79.06%, a 6% lower than the ten-fold CV AUC. Notably, the latency-corrected classifier performed significantly better than the baseline for less than 90 trials (two-tailed paired t-test, $p < 0.05$). In contrast, the use of previous data without correcting the latency always led to significantly lower AUCs than the other classifiers ($p < 0.05$), reaching 67.01% after 200 trials.

In the second case, E_1E_3, the latency-corrected classifier significantly outperformed the baseline when less than 150 trials were used ($p < 0.05$), obtaining similar AUCs after 200 trials (81.01% and 80.25% respectively) without reaching the AUC of the ten-fold CV. Again, the classifier using not corrected data always performed significantly worse than the others ($p < 0.001$), with an AUC of 65.77% after 200 trials.

In the last case (E_2E_3, c.f. Figure 8 right), the baseline classifier was always significantly worse than the latency-corrected one ($p < 0.0001$). After 200 trials, the baseline classifier reached a 79.93% of mean AUC versus a 82.97% when correcting the latency, close to the ten-fold CV classifier. The latter classifier was also significantly better than the non-corrected classifier with 30 trials or more ($p < 0.0001$).

To summarize, apart from one case –generalization from the P300 experiments E_1 to E_2– the latency-correction mechanism improved the classification performance in all cases. It allows to obtain significantly better classifiers than the baseline ones when a small number of trials from the new experiment are available.

4. Discussion

A practical issue in the study of event-related brain activity and its use for BCI applications is the time required to acquire sufficient data to have a reliable model or a usable classifier of the EEG signals. In general, each protocol is addressed as a completely new experiment even if they are tapping into similar cognitive processes. Besides the increase in the required time and resources, this also provides little information about how similar responses are across experimental conditions. Using several protocols on two well-studied signals, we showed that the experimental design mainly affected the ERP latencies. Moreover, we proposed a simple, yet powerful mechanism to compensate for these changes allowing to generalize BCI classifiers across experiments using a reduced amount of new data.

Variations in our protocols did not result in statistically significant amplitude differences across experiments. As stated in the introduction, however, there are several factors that could affect the amplitude of the ERP components. Indeed, it is well known that the P300 amplitude can vary depending on the target-to-target interval, a measure encoded by the target stimulus probability and the inter-stimulus interval among others [1, 33]. Similarly, previous studies have reported modulations on the ErrP amplitude depending on the error probability [16] or the error magnitude [34].

In contrast, ERP latencies were found to be different across several experiments (c.f. Section 3.1). In the P300 experiments, significant variations appeared between the pairs E_1E_3 and E_2E_3. Interestingly, experiment 3 had the most complex visual stimuli (a three-dimensional grid), seemingly requiring the subject longer time to evaluate the stimulus. This was supported by the NASA TLX questionnaire showing that experiment 3 induced a significantly higher workload than the other two; and by neurophysiological studies suggesting that the P300 is related to the stimulus evaluation time [13, 35].

Regarding the error potentials experiments, the latency changes were larger for both peaks (P3 and N4) than for P300 when changing experimental conditions. In this case, the NASA TLX revealed that the workload increased significantly from experiment 1 to experiment 2, and increased on average but not significantly from experiments 2 to 3. The selected protocols were designed so as to have an increased level of complexity both in the number and type of possible actions: changing from two to four possible actions at each state (from E_1 to E_2 and E_3); changing from 1D to 2D (from E_1 to E_2 and E_3); and changing from a simulated to a real device (from E_2 to E_3). Accordingly, increasingly longer latencies were found from protocols E_1 to E_3. It should be noticed that for the ErrP protocols part of the measured latencies may be also due to differences between the virtual and the real robot such as the time it takes the robot to start the movement after the control command has been issued or the velocity of its actions. Nonetheless, the use of a simple technique such as cross-correlation significantly allowed removing these latency jitters among experiments, as presented by the ANOVA results after correcting the latencies.

Despite one of the reasons behind these latency variations might be the overall workload of the performed task, there could be additional factors that affect the ERP latencies. For instance, different system implementations are a common source of latency jitter obtained in different experimental protocols [25]. More generally, the stimulus evaluation time is a well-known factor that affects the ERP latencies [13, 1]. This way, similar workloads of two experiments could have different ERP latencies. Similarly, there are other aspects that could affect direct or indirectly the stimulus evaluation time such as perceptibility [1], fatigue [36], target-to-target interval [33], recognition performance [37] or even cognitive capabilities [6]. Nonetheless, the latency estimation method should in principle be independent of the reason behind the latency differences, and thus should estimate these differences irrespectively of their nature. Studying these variations and how the latency estimation works under these circumstances is an interesting issue to address in future work.

Focusing on applications of brain-computer interfacing, we propose a simple latency correction mechanism to re-use data from previous experiments when building classifiers for new experiments on a related phenomenon. This yields a reduction in the calibration time as a smaller number of trials is required to achieve similar classification performance than if a new classifier is built from scratch (c.f. Figures 7 and 8). In those cases where there is no latency between protocols (e.g. moving from P300 experiment E_1 to E_2), the latency-corrected classifier was still better than the baseline one. In a similar way, Thompson et al. [21] also found that the latency variations among trials but within the same experimental protocol were one of the main problems for the classification performance, and proposed the use of within-experiment latency variations as a predictor of online BCI accuracies. The authors also argued that a brute-force method (i.e. testing a classifier for each possible latency and taking the classifier with maximum accuracy) could be used to estimate these latencies. Similarly, Aricò et al. found that larger within-experiment latency jitters present in covert-attention P300 spellers could be the reason behind a lower system performance [38]. It would be thus interesting to test the proposed approach as a way of correcting this jitter during online control and improve this way the system performance.

The latency variations have been assessed on two different ERPs (P300 and observation error potentials), showing their effect on the single-trial classification. In the future, the generalization of BCI decoders across protocols can be assessed for other ERPs, such as those generated during rapid visual processing [39], the N2 evoked component [40], or more generally the visually-evoked potentials (VEPs) [1], also present in the experiments performed in this work. Here, we correct the latency for each separate channel, and thus the latency estimation may be different for each channel depending on its most relevant components. However, we have not specifically addressed the fact that different components such as the VEPs could have different latency variations across protocols. An algorithm estimating the latency on each separate component could thus be useful to further improve the classifier generalization capabilities. Moreover, additional studies of event-related potentials in controlled and non-controlled applications may yield new findings. The proposed method could be used to elucidate common patterns across conditions, not only in BCI applications but also in neurophysiological studies, e.g. comparing latency variations between error-related activity in choice reaction tasks [5], and in feedback tasks [41].

Furthermore, more sophisticated techniques could be tested to cope with the latency variations such as dynamic time warping [42, 43]. Finally, one disadvantage of the proposed approach is that it relies on the assumption that there are only temporal changes in the ERPs, whereas the spatial contributions remain fixed among experiments. However, this assumption may be wrong. Thus, a more complete approach could be designed by performing a spatio-temporal compensation of the ERP variations.

Acknowledgments

This work has been partially supported by Spanish Ministry of Science projects HYPER-CSD2009-00067 and DPI2009-14732-C02-01 (I.I. and J.M.), DPI2011-25892 (L.M.), CAI Programa Europa (I.I.), DGA-FSE, grupo T04 (I.I., L.M. and J.M.), European ICT Programme Project FP7-224631 (J.d.R.M.), and Swiss NCCR "Robotics" (R.C. and J.d.R.M.).

References

[1] Luck SJ. *An Introduction to the Event-Related Potential Technique.* The MIT Press, 2005.

[2] Duncan CC, Barry RJ, Connolly JF, Fischer C, Michie PT, Ntnen R, Polich J, Reinvang I, and Van Petten C. Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clin Neurophysiol*, 120:1883–1908, Sep 2009.

[3] Pfefferbaum A, Wenegrat BG, Ford JM, Roth WT, and Kopell BS. Clinical application of the P3 component of event-related potentials. II. Dementia, depression and schizophrenia. *Electroencephalogr Clin Neurophysiol*, 59(2):104–124, 1984.

[4] Olofsson JK, Nordin S, Sequeira H, and Polich J. Affective picture processing: An integrative review of ERP findings. *Biol Psychol*, 77(3):247–65, March 2008.

[5] Falkenstein M, Hoormann J, Christ S, and Hohnsbein J. ERP components on reaction errors and their functional significance: A tutorial. *Biol Psychol*, 51:87–107, 2000.

[6] Polich J. On the relationship between EEG and P300: Individual differences, aging, and ultradian rhythms. *Int J Psychophysiol*, 26(1-3):299–317, 1997.

[7] Davies PL, Segalowitz SJ, and Gavin WJ. Development of error-monitoring event-related potentials in adolescents. *Ann N Y Acad Sci*, 1021:324–328, Jun 2004.

[8] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, and Vaughan TM. Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113(6):767–91, June 2002.

[9] Farwell LA and Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr Clin Neurophysiol*, 70(6):510–523, 1988.

[10] Iturrate I, Antelis J, Kübler K, and Minguez J. Non-invasive brain-actuated wheelchair based on a P300 neurophysiological protocol and automated navigation. *IEEE Trans Robot*, 25(3):614–627, 2009.

[11] Li L, Yao D, and Yin G. Spatio-temporal dynamics of visual selective attention identified by a common spatial pattern decomposition method. *Brain Res*, 1282:84–94, 2009.

[12] Sellers E, Krusienski D, McFarland D, Vaughan T, and Wolpaw J. A P300 event-related potential brain-computer interface (BCI): The effects of matrix size and inter stimulus interval on performance. *Biol Psychol*, 73(3):242–52, October 2006.

[13] Kutas M, McCarthy G, and Donchin E. Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science*, 197(4305):792, 1977.

[14] Schalk G, Wolpaw JR, McFarland DJ, and Pfurtscheller G. EEG-based communication: Presence of an error potential. *Clin Neurophysiol*, 111(12):2138–2144, 2000.

[15] Iturrate I, Montesano L, and Minguez J. Task-dependent signal variations in EEG error-related potentials for brain-computer interfaces. *J Neural Eng*, 10(2):026024, 2013.

[16] Chavarriaga R and Millán JdR. Learning from EEG error-related potentials in noninvasive brain-computer interfaces. *IEEE Trans Neural Syst Rehabil Eng*, 18(4):381–388, 2010.

[17] Chavarriaga R, Perrin X, Siegwart R, and Millán JdR. Anticipation- and error-related EEG signals during realistic human-machine interaction: A study on visual and tactile feedback. In *Conf Proc IEEE Eng Med Biol Soc*, 2012.

[18] Vidaurre C, Kawanabe M, von Bünau P, Blankertz B, and Müller KR. Toward unsupervised adaptation of LDA for brain-computer interfaces. *IEEE Trans Biomed Eng*, 58(3):587 –597, 2011.

[19] Lotte F and Guan C. Learning from other subjects helps reducing brain-computer interface calibration time. *Int Conf on Audio Speech and Signal Processing (ICASSP)*, 1:614–617, 2010.

[20] Kindermans PJ and Verschore H. A P300 BCI for the masses: Prior information enables instant unsupervised spelling. In *Neural Information Processing Systems (NIPS)*, pages 1–9, 2012.

[21] Thompson DE, Warschausky S, and Huggins JE. Classifier-based latency estimation: A novel way to estimate and predict BCI accuracy. *J Neural Eng*, 10(1):016006, December 2012.

[22] Iturrate I, Chavarriaga R, Montesano L, Minguez J, and Millán JdR. Latency correction of error potentials between different experiments reduces calibration time for single-trial classification. In *Conf Proc IEEE Eng Med Biol Soc*, 2012.

[23] Ferrez PW and Millán JdR. Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Transactions on Biomedical Engineering*, 55(3):923–929, 2008.

[24] Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, and Wolpaw JR. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Trans Biomed Eng*, 51(6), May 2004.

[25] Wilson JA, Mellinger J, Schalk G, and Williams J. A procedure for measuring latencies in brain-computer interfaces. *IEEE Trans Biomed Eng*, 57(7):1785–97, July 2010.

[26] Sauser E. Robottoolkit. Available online: lasa.epfl.ch/RobotToolKit.

[27] Woody CD and Nahvi MJ. Application of optimum linear filter theory to the detection of cortical signals preceding facial movement in cat. *Exp Brain Res*, 16:455–465, 1973.

[28] Levine SP, Huggins JE, BeMent SL, Kushwaha RK, Schuh LA, Rohde MM, Passaro EA, Ross DA, Elisevich KV, and Smith BJ. A direct brain interface based on event-related potentials. *IEEE Trans Rehabil Eng*, 8(2):180, 2000.

[29] Iturrate I, Montesano L, Chavarriaga R, Millán JdR, and Minguez J. Spatio-temporal filtering for EEG error related potentials. In *5th Int Brain-Computer Interface Conference*, 2011.

[30] Blankertz B, Lemm S, Treder M, Haufe S, and Müller KR. Single-trial analysis and classification of ERP components-a tutorial. *Neuroimage*, 2010.

[31] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[32] Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, pages 1165–1188, 2001.

[33] Gonsalvez CJ and Polich J. P300 amplitude is determined by target-to-target interval. *Psychophysiology*, 39(3):388–396, 2002.

[34] Luo Q and Qu C. Comparison enhances size sensitivity: Neural correlates of outcome magnitude processing. *PloS one*, 8(8):e71186, January 2013.

[35] Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, and Sejnowski TJ. Analysis and visualization of single-trial event-related potentials. *Hum Brain Mapp*, 14(3):166–185, 2001.

[36] Uetake AQ and Murata A. Assessment of mental fatigue during VDT task using event-related potential (P300). In *Proc of the Int Workshop of Robot and Human Interactive Communication*, pages 235–240. IEEE, 2000.

[37] Johnson R, Pfefferbaum A, and Kopell BS. P300 and long-term memory: Latency predicts recognition performance. *Psychophysiology*, 22(5):497–507, 1985.

[38] Aricò P, Aloise F, Schettini F, Salinari S, Mattia D, and Cincotti F. Evaluation of the latency jitter of P300 evoked potentials during (c)overt attention BCI. In *Proc of TOBI Workshop IV*, 2013.

[39] Gerson AD, Parra LC, and Sajda P. Cortically-coupled computer vision for rapid image search. *IEEE Trans Neural Syst Rehabil Eng*, 14(2):174–179, Jun 2006.

[40] Hong B, Guo F, Liu T, Gao X, and Gao S. N200-speller using motion-onset visual response. *Clin Neurophysiol*, 120(9):1658–1666, 2009.

[41] Nieuwenhuis S, Holroyd CB, Mola N, and Coles MGH. Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neurosci Biobehav Rev*, 28:441–448, 2004.

[42] Berndt D and Clifford J. Using dynamic time warping to find patterns in time series. In *AAAI Workshop on Knowledge Discovery in Databases*, volume 398, pages 229–248, 1994.

[43] Casarotto S, Bianchi AM, Cerutti S, and Chiarenza GA. Dynamic time warping in the analysis of event-related potentials. *IEEE Eng Med Biol Mag*, 24(1):68–77, 2005.