Generalization of ErrP-calibration for different error-rates in P300-based BCIs

Aniana Cruz
Institute for Systems and Robotics,
University of Coimbra
3030–290 Coimbra, Portugal
Email: anianabrito@isr.uc.pt

Gabriel Pires
Institute for Systems and Robotics,
University of Coimbra
3030–290 Coimbra, Portugal
and Polytechnic Institute of Tomar
2300-313 Tomar, Portugal
Email: gpires@isr.uc.pt

Urbano J. Nunes
Institute for Systems and Robotics,
and Department of Electrical
and Computer Engineering
University of Coimbra
3030–290 Coimbra, Portugal
Email: urbano@isr.uc.pt

***Abstract*—Automatic recognition of error-related potentials (ErrPs) requires a long calibration time in order to have enough error-samples to train the classifier. In this paper we analyze whether it is possible to reduce the ErrP-calibration time in a P300-based brain-computer interface (BCI), by calibrating the BCI with a high rate of errors (wrong detections of user intent). We analyze if a high error-rate condition still produces a discriminable ErrP and if its classification model generalizes well in sessions of different error-rates. Results show that the classification model built from a high error-rate calibration can be used successfully on sessions with lower error-rates.**

I. INTRODUCTION

Human-machine interaction plays an essential role in many application domains such as industrial, medical, and entertainment [1]. In all these domains, the interaction systems are prone to errors, which may be due to low reliability of the system or due to users' errors. An electroencephalographic signal called error-related potential (ErrP) has been identified in many contexts as a neurophysiological signal associated to error processing. The ErrP is generated naturally in the brain when the user perceives that an error has been made by himself/herself or by the system [2], [3]. It appears within a time window of 500 ms, and thus its automatic detection could be used in myriad ways, in real-time, in human-machine interaction processes. In the context of brain-computer interfaces (BCIs), i.e., when BCI is being used as the primary communication channel, ErrP detection was researched to increase the reliability and the information transfer rate in P300-based BCI spellers [4], [5] and in BCIs based on motor-imagery [6], by eliminating or even correcting errors. ErrPs were also researched in real-life tasks for monitoring decisions/actions of systems not controlled by the user [7], [8]. In [7] ErrPs were used to validate the predictions of a driving assistance system (selection of a driving direction in intersections) in a simulated and a real car. In [8], the selections of a robot were corrected based on ErrPs elicited when a human operator observed wrong actions of the robot. In [9], ErrPs were used in a reinforcement-learning loop to change the behavior of an agent in a simulated environment. Although all these works are proofs-of-concept, they show the wide range of applications that could benefit from the automatic recognition of ErrPs. Yet, this detection is challenging because it has to be obtained from a single trial and ErrPs have a very low signal-to-noise ratio. Although many studies detected single-trial ErrPs successfully, a low recognition accuracy can render this signal useless. For example, in a P300-based BCI, if the classification accuracy of the ErrP is lower than that of the P300, it may lead to a degradation of the performance [10]. In our previous work [4] we tackled this issue by using a double-ErrP detection approach, i.e., a primary ErrP is used to detect a wrong selection, then a correction is proposed, and finally a secondary ErrP is used to validate the correction. This showed on one hand the possibility of increasing the performance of the system mitigating the single trial issue, and on the other hand showed that ErrPs could be used in a closed loop, increasing the level of human-machine interaction. The great variability of single-trial ErrPs poses also difficulties in the generalization of the classification model across sessions. In [11], authors evaluate the capability of the ErrP classifier to generalize across two different recording dates and across different inter-stimulus intervals (pace at which the stimuli are provided to the participants). The results showed that classification performance decrease in both conditions. Although the average waveform of the ErrPs is reported to be stable between sessions of different days, there is still a significant decrease on the classification rate between sessions, which can reach over 10% [11], [4]. In [4] we made a calibration with an error rate similar to that expected during online operation, i.e., with an error rate between 10 and 15%, and then we used this calibration model in a session of a different day. This calibration was quite long (about 2 hours) so that a significant number of error-trials could be gathered to train the classifier. The results of the validation test, conducted some months after the calibration, showed a decrease of the classification accuracy that was more than 10% for some participants, leading to a significant impact on the online performance of the BCI. Other studies reported similar calibration approaches using real or sham errors [10] [12]. Such a long calibration session is suitable if used only once, but is impractical if used every time the user uses the system. On the other hand, the number of error-samples collected during a short calibration may be insufficient to train

2577-1655/18/$31.00 ©2018 IEEE
DOI 10.1109/SMC.2018.00118

the classifier. One approach is to increase the error-rate during calibration to gather more error-trials in less time reducing the calibration time, however it remains uncertain if a calibration obtained with a high error-rate generalizes well for a situation of an error-rate substantially lower.

In this paper we analyse whether the ErrPs elicited by a P300-speller with a high error-rate condition (around 40%) are still discriminable and classifiable, and whether a classification model obtained with this condition could generalize to a scenario in which it is expected an error-rate substantially lower (around 10-15%). This calibration with high error-rate would allow a greater amount of error-trials to be gathered in much less time.

The remainder of this paper is organized as follows: the materials are presented in Section II, including the P300-based BCI paradigm with error detection in Subsection II-A, and data acquisition and detection algorithm in Subsection II-B. The experimental study and methodology are described in Section III which includes three Subsections: participants, phases of experiment and metrics. The experimental results are presented in Section IV, and conclusions are drawn in Section V.

II. Materials

A. P300-ErrP BCI

Figure 1 shows a generic view of the P300-ErrP BCI system. The BCI application is a communication speller called lateral single character (LSC), which was introduced in [13]). Twenty-eight symbols, including all letters of the alphabet and the symbols 'space' and 'del' flash randomly according to an oddball paradigm. A target symbol is expected to elicit a P300 event related potential (ERP). Each symbol flashes individually during 75 ms alternating between left and right sides of the screen, and with no inter-stimulus interval (ISI). The number of rounds of each trial (N_{rep}) was adjusted individually to each user according to the BCI target error-rate we wanted to set for calibration or for online operation. The inter-trial interval (ITI), set to 4 s, accommodated the time for double-error detection and correction and the time for the user to shift his/her attention to the next desired symbol. The overall time for one trial is:

$$TT = N_{rep} \times N_s \times SOA + CT + ITI \tag{1}$$

where $N_s = 28$ is the number of symbols, $SOA = 75ms$ is the stimulus onset asynchrony and $CT = 1s$ is the time associated with the last flash of the trial.

The EEG signals are classified as target or non-target and the detected symbol is shown to the user. It is expected that an ErrP is elicited if the system does not recognize the user intent and a correct ERP is elicited if the system detects it correctly. If the detected symbol shown to the user elicits an ErrP (1^{st} ErrP), the system deletes the symbol and replaces it by the symbol with the second highest P300 classification score. If this correction elicits a 2^{nd} ErrP, the final classification is the first detected symbol, otherwise the corrected symbol is chosen.

B. Data Acquisition and Classification

The electric brain potentials were recorded with a gUSBamp acquisition system (G.Tec, Inc.). EEG signals were acquired with electrodes placed at Fz, Cz, C3, C4, CPz, Pz, P3, P4, PO7, PO8, POz and Oz channels, according to the international extended 10-20 standard system. An electrode positioned on the right ear lobe was used as reference and the AFz electrode was used as ground. The signals were acquired at 256 Hz and filtered using a band-pass filter between 1 and 10 Hz and a notch filter at 50 Hz.

P300 and ErrP classifiers share the same classification framework, which was the same used in [4]. After preprocessing, the EEG signal is segmented into 1 s epochs, $E_{N \times T}$, where N is the number of channels and T is the number of time samples. The features are extracted using a statistical spatial filter based on a Fisher criterion beamformer (SF-FCB) [15]. The EEG epochs $E_{N \times T}$ are projected into $Y = W'E$, where W is the optimal filter obtained from calibration phase, and $'$ represents the transpose operator. Then, the features are classified using a Bayes classifier. The target symbol is the one with the highest classification score and the corrected target is the symbol with the second highest classification score (see a simplified representation of the classification pipeline in Fig. 1).

III. Experimental Study and Methodology

A. Participants

The experiments were carried out by five healthy participants (3 males, 2 females with age ranging 25-33 years old). All participants signed an informed consent to participate in the study. These participants took part also in the study in [4]. Since there is a direct comparison of the results of that study with the present study, we kept the same identification of the participants, namely, S1, S2, S3, S5 and S6. During the experiments, participants sat in front of a computer screen at a distance of around 70 cm. They were asked to focus on the target stimulus and ignore the remaining standard events, while mentally counting the times that the target flashes, helping to increase their attention level. Participants were instructed to be aware of the detected symbol to realize if it was the desired symbol. They were also informed that an automatic correction or re-correction could occur.

B. Calibrations and Online Session

The integration of error detection in the P300 speller requires two calibrations (calibration of the P300 classifier and calibration of the ErrP classifier). Therefore, the experiment consisted of three phases: P300 calibration, error-detector calibration and online session with error detection and correction.

1) Calibration of P300-classifier: In this calibration we gathered the EEG data associated with target and standard events to train the models of the P300 classifier, which was used as the primary communication channel. The P300 calibration phase took about 5 minutes collecting 90 target epochs and 2430 non-target epochs.

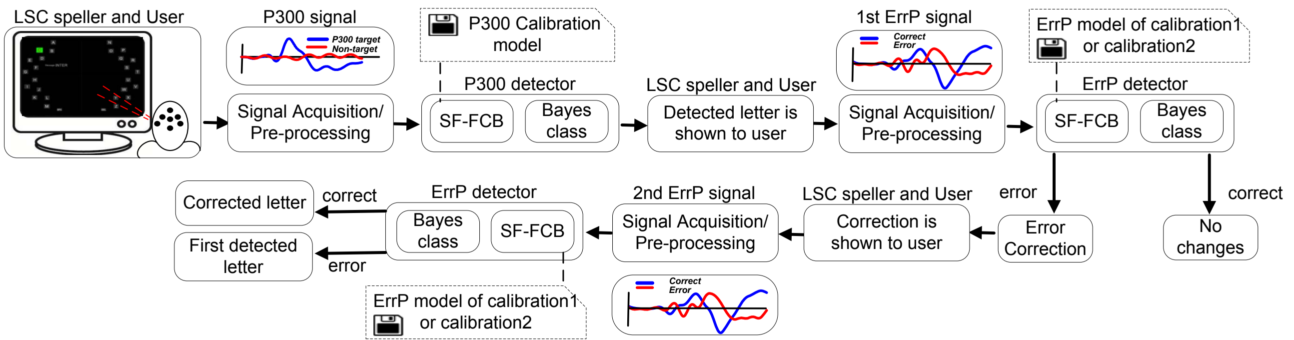

Fig. 1. Schematic representation of the P300-ErrP BCI system. User focuses on the target letter of the P300-LSC speller. The EEG signal is acquired and pre-processed, and the target event is detected with the P300 detector. The detected letter is shown to the user and the ErrP-detector evaluates whether an error has occurred. If an ErrP is detected, the letter with the second highest classification score is shown to the user and the ErrP-detector checks for a possible error again. If the system detects an ErrP, the final classification is the initial detected letter, otherwise the corrected letter is selected (see a demonstrative video in [14]).

TABLE I
ERRP TRAINING AND VALIDATION DATASETS USED IN THE PRESENT STUDY (CALIBRATION1 AND TEST1) AND IN OUR PREVIOUS STUDY [4] (CALIBRATION2 AND TEST2).

	$Calibration1$	$Test1$	$Calibration2$ [4]	$Test2$ [4]
Error-rate	$\sim 40\%$	$\sim 15\%$	$\sim 20\%$	$\sim 15\%$
Duration	10 min	1 hour	2 hours	1 hour
N_{rep}(P300)	2	4.6	4.2	6

2) Calibration of Error-detector: **Calibration1** - In this calibration procedure, participants spelled two 32-letter sentences (64 characters) without correction. It served to gather a dataset associated with positive and negative feedback responses. In order to obtain a high rate of wrong detections, the P300-LSC speller was set-up with a small number of stimuli rounds, namely $N_{rep} = 2$. The calibration took about 10 minutes, which included on average 26.4 wrong samples with a mean error-rate of 41.3%. The number of errors ranged between 21 and 40.

Calibration2 - Refers to the dataset of the error-detector calibration obtained in our previous study [4] that will be used for comparison (see Table I). The calibration lasted about two hours with a number of rounds averaging 4.2, yielding an error-rate of 18.7% (the number of errors ranged between 31 and 75).

3) Online Session: **Test1** - This session used the classification models obtained in the calibration of the P300-classifier and in $Calibration1$. Participants were asked to spell the Portuguese sentence "ESTOU-A-ESCREVER-COM-UMA-INTERFACE-BCI" several times during approximately one hour. At the end of each sentence the participants rested two minutes. The time that each participant took to write the sentence depended on the number of event repetitions. N_{rep} was settled individually for each user to reach $\approx 90\%$ accuracy (based on the outcome of the calibration of the P300-classifier). The BCI system performed the automatic double-error detection and correction as described previously.

Test2 - Refers to the online session of our previous study [4], using the exact same conditions of $Test1$. The classification model was built from $Calibration2$.

C. Metrics

To evaluate the offline and online performance of the P300-ErrP BCI (Fig. 1), we computed the sensitivity ($Sens$), specificity ($Spec$), accuracy (Acc) and balanced accuracy metrics, which are defined as $Sens = \frac{TP}{(TP+FN)}$, $Spec = \frac{TN}{TN+FP}$, $Acc = \frac{TP+TN}{FN+TN+FP+TP}$ and $balancedAcc = \frac{Sens+Spec}{2}$, where TP, TN, FN and FP refer to the number of true positives, true negatives, false negatives and false positives, respectively. The balanced accuracy is used in offline analysis to ensure that the results are not biased by imbalanced classes.

IV. RESULTS AND DISCUSSION

A. Evoked Potentials Morphology

Fig. 2 compares the grand averages of potentials elicited by wrong and correct feedback in channels Fz and Cz for 40% error-rate ($Calibration1$) and for 15% error-rate ($Test1$). The ErrP of $Calibration1$ has different waveforms of the ErrP of $Test1$ (Pearson's correlation coefficient $r = 0.64$ for both channels). The higher error-rate condition produced an ErrP with lower amplitude. This is consistent with other studies [9], which also reported a decrease of amplitude for higher error-rates, but here we also found changes in ErrP morphology, namely, the second positive peak of the ErrP elicited by the 40% error-rate condition had a lower latency (less than 80 and 90 ms for channels Fz and Cz, respectively). This resulted in an ErrP waveform very similar to the correct ERP signal (Pearson's correlation coefficient $r = 0.92$ and $r = 0.87$ for channel Fz and Cz respectively), which makes uncertain the possibility of a successful classification. This result may suggest an habituation effect that might reduce the impact of errors. On the other hand, for the lower error-rate condition the waveforms are clearly discriminable. There is also a consistency of the characteristics of the grand average of the 1^{st} ErrP and correct ERP in regard to those exhibited in $Test2$

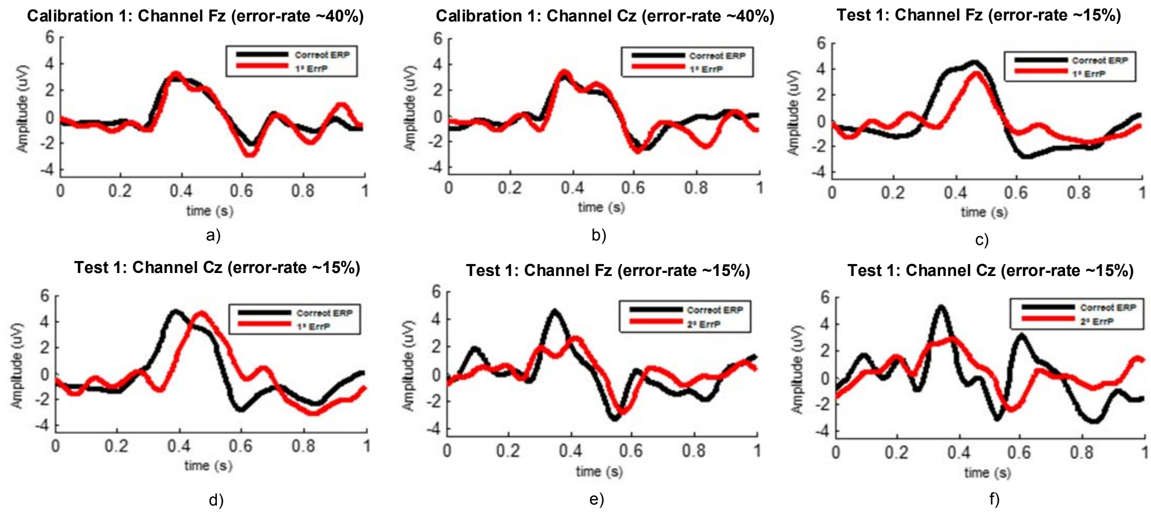

Fig. 2. Grand average of evoked potentials after positive and negative feedback. a) and b) ERPs produced during the calibration of the ErrP detector with an error-rate of 40% for the channels Fz and Cz respectively. 1^{st} ErrP (c and d) and 2^{nd} ErrP (e and f) and ERP after correct feedback during the online session with error-rate of $\sim$ 15% for the channels Fz and Cz respectively.

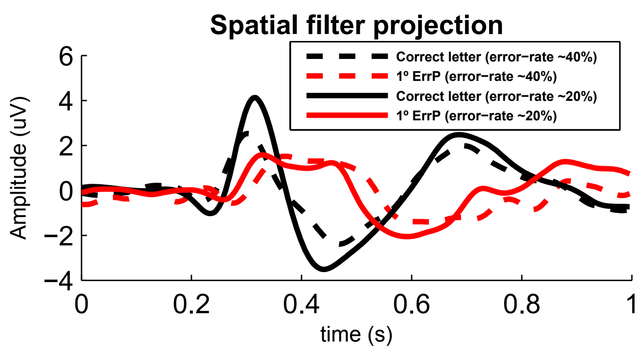

Fig. 3. Grand average of the 1^{st} projection of the FCB spatial filter, using the EEG data of $Calibration1$ (dashed line) and $Test1$ (solid line).

(performed 22 months before). This stability across sessions was also reported in other studies [9], [11]. Nevertheless, this waveform similarity did not avoid a significant performance decay between calibration and test sessions, as reported in [11], [4]. The 2^{nd} ErrP had a morphology different of the one obtained in [4]. We hypothesize that this may be related to the different information we provided to the participants. While in [4] it was not explained what the double-error correction algorithm was doing, in the present study the participants were explained that if the corrected letter elicited a 2^{nd} ErrP, the first detected letter would be chosen as target letter. Therefore, letters corrected wrongly, owing to false positives, may have been less important to participants because they expected the correct letters would be re-selected.

B. Offline Classification Accuracy

Table II shows for each participant the error-rate, the total number of errors and the balanced accuracy of the 1^{st} error-detector obtained in $Calibration1$. For a direct comparison,

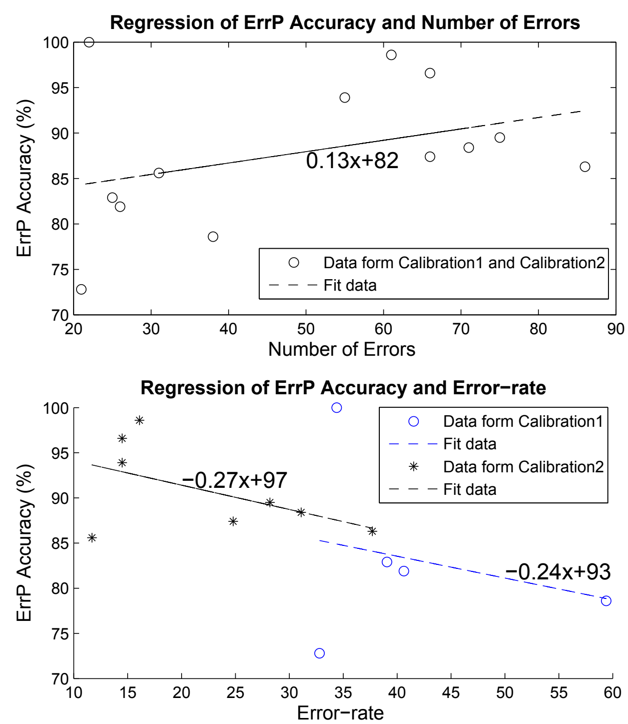

Fig. 4. Linear regression between offline ErrP classification accuracy and number of errors (top) and error-rate (bottom), using datasets of error-detector training collected in the present and previous study [4].

we show the results for the same five subjects who participated in [4] ($Calibration2$). For $Calibration1$, N_{rep} was 2 on average, which yielded an average error-rate of 41.3%. However, participant S6 had a significantly higher error-rate (around 60%), i.e., more wrong samples than correct samples.

TABLE II
OFFLINE RESULTS OBTAINED FROM ERRP-CALIBRATION IN $Calibration1$ ($\sim$ 40% ERROR-RATE) AND IN $Calibration2$ ($\sim$ 20% ERROR-RATE). BALANCED ACCURACY WAS OBTAINED FROM 10-FOLD CROSS VALIDATION.

	ErrP Calibration1 (error rate of $\sim$ 40%)			ErrP Calibration2 (error rate of $\sim$ 20%)			
Subjects	Error-rate (%)	Number of errors	Acc-ErrP1 (%)	Error-rate (%)	Number of errors	Acc-ErrP1 (%)	Acc-ErrP1 (%) (using half of error-samples)
S1	39.1	25.0	82.9	28.2	75.0	89.5	87.2
S2	32.8	21.0	72.8	14.5	55.0	93.9	86.6
S3	34.4	22.0	100.0	14.5	66.0	96.6	90.4
S5	40.6	26.0	81.9	24.8	66.0	87.4	70.8
S6	59.4	38.0	78.6	11.7	31.0	85.6	77.9
Mean	41.3	26.4	83.2	18.7	58.6	90.6	82.6

The mean value of the balanced accuracy of the ErrP detector (obtained using 10-fold cross-validation) was 83.2%. Despite the high similarity between positive and negative feedback for the high error-rate condition, the accuracy of the classification was still high. Plotting the projection of the FCB spatial filter (Fig. 3) we observe that the spatial filter was able to maximize the discrimination of the features of the two classes, approaching the projection obtained with the lower error-rate condition, which explains the good accuracy. For $Calibration2$ the balanced accuracy of the ErrP detector was 90.6%. The average result of the ErrP-detector in $Calibration1$ was 7.4% lower, but the number of error-samples used for training was less than half of that of $Calibration2$, which influenced the classifier training. To verify whether this difference in classification accuracy is related to the high error-rate condition or it is due to the lack of error samples, we computed the results that would be achieved in $Calibration2$ if only half of the calibration dataset would have been used, which yielded an accuracy of 82.6%. Thus, for a similar number of errors, the two conditions have similar performance (around 83%). This suggests that the ErrP classifier is not strongly affected by the error-rate, but mainly by the number of training samples. Additionally, the FCB spatial filter seems to have an important contribution to confer robustness to changes of the ErrP waveforms between the two error-rate conditions. Authors in [11] also highlighted the influence of spatial filtering in the classification accuracy across conditions.

To further analyze whether there is a direct correlation between ErrP accuracy, error-rate and number of error-samples we made a regression analysis, using data from $Calibration1$ and $Calibration2$ (Fig. 4). Joining the datasets of $Calibration1$ and $Calibration2$, we observe a linear relationship between ErrP accuracy and the number of error-samples (regression coefficient is 0.13), however we stress that this result is not statistically significant (t-test, $p = 0.22$). For the regression between ErrP accuracy and error-rate we analyzed the datasets separately. There is a negative relationship (coefficients of -0.24 and -0.27 respectively for $Calibration1$ and $Calibration2$), but the result is not statistically significant (t-test, $p = 0.68$ and $p = 0.19$ respectively). A two-way ANOVA was also performed to evaluate the combined effect of error-rate and number of error-samples in the classification performance. The p-value of 0.28 ($F = 1.32$) showed that the combination of these variables is not statistically significant. Despite the inconclusive analysis, the results point to the need for more error-samples for training, which could be obtained by increasing the duration of the calibration or pushing the error-rate to 50%, achieving a balanced number of samples with correct and wrong responses.

TABLE III
ONLINE CLASSIFICATION RESULTS IN $Test1$ USING THE CLASSIFICATION MODEL FROM *Calibration1* AND ACCURACY OF 1ST ERRP OBTAINED IN $Test2$ [4].

	Pre-Acc ($Test1$)	Post-Acc ($Test1$)	Acc-ErrP1 ($Test1$)	Acc-ErrP2 ($Test1$)	Acc-ErrP1 ($Test2$)	N_{rep}
S1	79,6	90,5	91,4	75,0	90,8	5,0
S2	87,4	90,5	93,2	100,0	96,3	4,0
S3	90,5	94,7	99,5	89,5	89,5	3,0
S5	82,2	77,0	74,3	52,3	92,1	6,0
S6	87,4	90,5	75,7	85,7	76,3	5,0
Mean	85,4	88,7	86,8	80,5	89,0	4,6

C. Online Classification Accuracy

Table III presents the online BCI results obtained in $Test1$ using the classification model from $Calibration1$. The online average classification accuracy of the 1^{st} ErrP was 86.8% (error-rate of 14.6%), i.e., 3.6% higher than the one obtained from cross-validation in $Calibration1$ (error-rate of 41.3%). This result showed that the calibration model obtained for the higher error-rate generalized well for the lower error-rate condition, the main hypothesis that we wanted to verify. Comparing the average accuracy achieved in $Test2$ (89.0%) there was only a small performance loss (2.2%).

Regarding the 2^{nd} ErrP (ErrP evoked when the automatic correction is wrong), the ErrP classification dropped to 80.5%. The classification of the 2^{nd} ErrP uses the model trained with the responses of the 1^{st} ErrP. This shows that the classification model of the 1^{st} ErrP does not generalize well to the 2^{nd} ErrP. This was also identified in [4], which consistently shows that the characteristics of the 1^{st} and 2^{nd} ErrPs are different.

D. Analysis of Automatic Error Correction

The classification accuracy before the automatic error correction was 85.4%, and increased 3.2% after automatic error correction. However, the enhancement is not statistically significant (paired t-test, p = 0.14). Participant S1 had the greatest improvement, about 10%. On the other hand, participant S5 had a worse classification accuracy after correction. The 2^{nd}

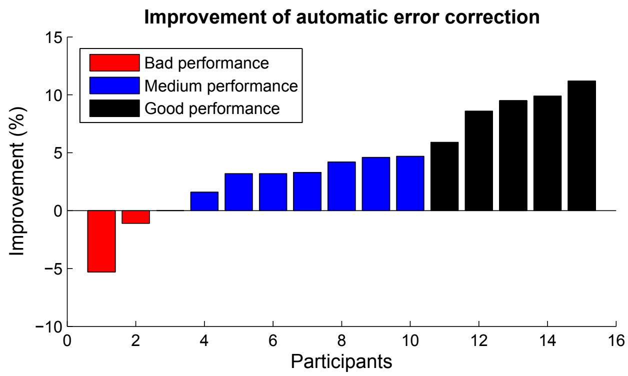

Fig. 5. Improvement after automatic error correction for each participant using datasets of present and previous study [4].

ErrP is effective in reducing the number of false positives by correcting the correct targets that are detected as errors in the 1^{st} error-detector, however when the classification accuracy of the 2^{nd} ErrP-detector is poor, the performance decreases (which was the case of this participant).

To analyze the relationship between the number of errors obtained in the calibration of the error-detector and the online improvement after automatic error correction, we grouped the data in three categories (Fig. 5): poor performance (the final classification accuracy is equal or lower than initial accuracy), medium performance (improvement less than 5%) and good performance (improvement greater than 5%). Averaging the number of errors got in $Calibration1$ and $Calibration2$ for participants of each category, the mean number of errors for poor, medium and high performance was 28.5, 51.3 and 56.8 respectively. This result may suggest that a calibration phase with at least 50 error-samples could always produce an improvement after error correction. However, it should be noted that the improvement also depends on the correction rate, i.e., the ability of replacing a wrong symbol by the correct one.

V. Conclusion

In this paper, we evaluated the generalization of the ErrP classifier over different error-rates. The online ErrP accuracy (trained with high error-rate and tested with low error-rate) was higher than the offline ErrP accuracy (trained and tested with high error-rate). This result is indicative that classification model built from a calibration with high error-rate generalizes to conditions with lower error-rates. Nevertheless, the classification is still affected by the low number of samples gathered during training. The results show that the FCB spatial filter makes the ErrP features almost stable over different error-rates.

Acknowledgment

This work has been financially supported by the Ph.D. Scholarship SFRH/BD/111473/2015 of Aniana Cruz, granted by the Portuguese foundation for science and technology (FCT), by the Project B-RELIABLE: SAICT/30935/2017, with FEDER/FNR/OE funding through programs CENTRO2020 and FCT, and by UID/EEA/00048/2013 with FEDER funding, through programs QREN and COMPETE.

References

[1] M. A. Goodrich and A. C. Schultz, “Human-robot interaction: a survey,” *Foundations and trends in human-computer interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[2] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, “ERP components on reaction errors and their functional significance: a tutorial,” *Biological psychology*, vol. 51, no. 2-3, pp. 87–107, 2000.

[3] P. W. Ferrez and J. d. R. Millán, “Error-related EEG potentials generated during simulated brain–computer interaction,” *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 923–929, 2008.

[4] A. Cruz, G. Pires, and U. J. Nunes, “Double ErrP Detection for Automatic Error Correction in an ERP-Based BCI Speller,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 26–36, 2018.

[5] T. Zeyl, E. Yin, M. Keightley, and T. Chau, “Adding real-time Bayesian ranks to error-related potential scores improves error detection and auto-correction in a P300 speller,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 46–56, 2016.

[6] P. W. Ferrez and J. d. R. Millán, “Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy,” in *Proceedings of the 4th international brain-computer interface workshop and training course*, no. CNBI-CONF-2008-004, 2008, pp. 197–202.

[7] H. Zhang, R. Chavarriaga, Z. Khaliliardali, L. Gheorghe, I. Iturrate, and J. d R Millán, “EEG-based decoding of error-related brain activity in a real-world driving task,” *Journal of neural engineering*, vol. 12, no. 6, p. 066028, 2015.

[8] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus, “Correcting robot mistakes in real time using eeg signals,” in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6570–6577.

[9] R. Chavarriaga and J. d. R. Millán, “Learning from EEG error-related potentials in noninvasive brain-computer interfaces,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 4, pp. 381–388, 2010.

[10] M. Spüler, M. Bensch, S. Kleih, W. Rosenstiel, M. Bogdan, and A. Kübler, “Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI,” *Clinical Neurophysiology*, vol. 123, no. 7, pp. 1328–1337, 2012.

[11] F. Iwane, R. Chavarriaga, I. Iturrate, and J. d. R. Millán, “Spatial filters yield stable features for error-related potentials across conditions,” in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 000 661–000 666.

[12] N. M. Schmidt, B. Blankertz, and M. S. Treder, “Online detection of error-related potentials boosts the performance of mental typewriters,” *BMC neuroscience*, vol. 13, no. 1, p. 19, 2012.

[13] G. Pires, U. Nunes, and M. Castelo-Branco, “Comparison of a row-column speller vs. a novel lateral single-character speller: assessment of BCI for severe motor disabled patients,” *Clinical Neurophysiology*, vol. 123, no. 6, pp. 1168–1181, 2012.

[14] G. Pires, “LSC:P300-ErrP Video,” http://home.isr.uc.pt/~gpires/ErrP/, 2016, [Online; accessed July-2018].

[15] G. Pires, U. Nunes, and M. Castelo-Branco, “Statistical spatial filtering for a P300-based BCI: tests in able-bodied, and patients with cerebral palsy and amyotrophic lateral sclerosis,” *Journal of neuroscience methods*, vol. 195, no. 2, pp. 270–281, 2011.