Online asynchronous detection of error-related potentials in participants with a spinal cord injury using a generic classifier

Catarina Lopes-Dias, Andreea I Sburlea, Katharina Breitegger, Daniela Wyss, Harald Drescher, Renate Wildburger and Gernot R Müller-Putz

1 Institute of Neural Engineering, Graz University of Technology, Graz, Austria
2 AUVA-Rehabilitationsklinik Tobelbad, Tobelbad, Austria
3 BioTechMed, Graz, Austria

E-mail: gernot.mueller@tugraz.at

Keywords: brain–computer interface, error-related potential, generic classifier, spinal cord injury, online experiment, asynchronous classification, end-users

Supplementary material for this article is available online

Abstract

For brain–computer interface (BCI) users, the awareness of an error is associated with a cortical signature known as an error-related potential (ErrP). The incorporation of ErrP detection into BCIs can improve their performance.

Objective. This work has three main aims. First, we investigate whether an ErrP classifier is transferable from able-bodied participants to participants with a spinal cord injury (SCI). Second, we test this generic ErrP classifier with SCI and control participants, in an online experiment without offline calibration. Third, we investigate the morphology of ErrPs in both groups of participants.

Approach. We used previously recorded electroencephalographic data from able-bodied participants to train an ErrP classifier. We tested the classifier asynchronously, in an online experiment with 16 new participants: 8 participants with SCI and 8 able-bodied control participants. The experiment had no offline calibration and participants received feedback regarding the ErrP detections from the start. To increase the fluidity of the experiment, feedback regarding false positive ErrP detections was not presented to the participants, but these detections were taken into account in the evaluation of the classifier. The classifier was not trained with the user’s brain signals. However, its performance was optimized during the online experiment by the use of personalized decision thresholds. The classifier’s performance was evaluated using trial-based metrics, which considered the asynchronous detection of ErrPs during the entire trial’s duration.

Main results. Participants with SCI presented a non-homogenous ErrP morphology, and four of them did not present clear ErrP signals. The generic classifier performed better than chance in participants with clear ErrP signals, independently of the SCI (11 out of 16 participants). Three out of the five participants that obtained chance level results with the generic classifier would not have benefitted from the use of a personalized classifier.

Significance. This work shows the feasibility of transferring an ErrP classifier from able-bodied participants to participants with SCI, for asynchronous detection of ErrPs in an online experiment without offline calibration, which provided immediate feedback to the users.

1. Introduction

Brain–computer interfaces (BCIs) can assist people with severe motor impairments to operate external devices by converting their modulated brain activity into the control of these devices [1–3]. Although a promising technology, most BCIs are still error-prone, and the frequent occurrence of errors can spoil the BCI user’s experience. The user’s awareness of an unintended response from the device that he/she is controlling is associated with a neural signature known as error-related potential (ErrP) [4, 5].

ErrPs are associated with conflict monitoring and error processing [6] and can be measured using non-invasive techniques such as electroencephalography (EEG), that are often used for BCI control. Therefore,
ErrPs can be used to improve BCI performance, either in a corrective manner, by allowing corrective actions, or in an adaptive manner, by reducing the possibility of future errors [7–11]. The real-time detection of ErrPs is pertinent in BCIs used by persons with motor impairments and also in applications targeting healthy users [12–15]. The incorporation of ErrP detection into a BCI promotes a smoother interaction with its user. Nevertheless, this incorporation has not been widely investigated.

The use of ErrPs in discrete BCIs, which are controlled in discrete steps, is well established in healthy participants [4, 10, 16–22] and has also been marginally tested in potential end-users of BCIs [23]. Still, BCIs are developing in the direction of offering users continuous control of an external device; these are known as continuous BCIs [24–28]. The incorporation of ErrPs into such BCIs requires an asynchronous detection of ErrPs, since the user can realise at any moment during the control of the device that an error has occurred. The asynchronous detection of ErrPs has been studied in healthy participants, both in offline scenarios [13, 29–34] and more recently in online scenarios [35].

A possible explanation for the limited use of ErrPs in BCIs can be linked to the reliance of most BCIs on personalised classifiers, which are constructed with the user’s brain signals. Since a considerable amount of data is necessary to reliably train the classifier, personalised classifiers commonly require a long calibration period before the user can receive feedback from their own brain signals. In this manner, combining ErrPs with other controlling signals would imply collecting calibration data for all the different signals, increasing the calibration period further. Alternatively, using an ErrP classifier that does not require calibration with the user, could encourage the integration of ErrPs with other control signals when constructing BCIs. This could be achieved by either transferring an ErrP classifier across different tasks or different participants. Both options have been tested in discrete tasks, in offline conditions [23, 36–43] and online conditions [20]. Recently, the asynchronous detection of ErrPs with a generic classifier has been studied in the context of a continuous task, in offline conditions [44] and pseudo-online conditions [45].

Very few works have addressed the study of ErrPs in potential BCI end-users, and studies to date have mainly been conducted offline. Keyl and colleagues characterized the morphology of the ErrPs of participants with a spinal cord injury (SCI) and compared it with able-bodied control participants [46]. The ErrP morphology was comparable in the two groups, but the ErrPs of the SCI group showed smaller peak amplitudes. Kumar and colleagues studied ErrPs during post-stroke rehabilitation movements [47]. In this work, individual participants did not display very clear ErrP patterns. Spüler and colleagues studied ErrPs in six participants with amyotrophic lateral sclerosis (ALS) in an online experiment and showed that the incorporation of ErrPs improved the BCI performance [23]. This work also included the offline analysis of the transfer of an ErrP classifier across ALS participants.

Our study has three main aims. First, we test the feasibility of transferring an ErrP classifier for asynchronous classification from able-bodied participants to potential end-users of BCIs, in particular, to participants with a high-level SCI. Second, we test the feasibility of asynchronously using a generic ErrP classifier in an online experiment, in which both participants with SCI and control participants took part. Third, we investigate the morphology of ErrPs, both in participants with SCI and in control participants.

In the work presented here, we recorded EEG signals from both participants with SCI and control participants, while asynchronously testing a generic ErrP classifier in a closed-loop online experiment. The generic classifier had been trained with the EEG data of 15 able-bodied participants from a previous study of ours and was not retrained during the experiment [35]. This allowed us to create an online experiment with no offline calibration period, in which participants received immediate feedback for their brain signals from the very beginning of the experiment onwards.

2. Methods

2.1. Participants

Sixteen volunteers participated in the experiment, eight of whom had a spinal cord injury. The mean age of the participants with SCI was 37.5 ± 9.7 years (mean ± std). The remaining participants were able-bodied control participants. Each participant with SCI was matched with a control participant of the same sex and with a maximum age difference of 5 years. The control participants were, on average, 35.9 ± 10.8 years old (mean ± std).

The participants with SCI had an injury between levels C4 and Th2. Table 1 summarizes the demographic and clinical data of the participants with SCI: age, sex, neurological level of injury (NLI) and ASIA impairment score (AIS).

2.1.1. Inclusion and exclusion criteria

All participants had to be aged between 18 and 65 years. Given that the experimental paradigm required a preserved arm function, all participants with SCI had an injury at level C4 or lower. Participants with SCI were excluded if they were artificially ventilated or had major spams, due to possible interference with the EEG measurement. Control participants were required to be able-bodied and with no history of neurological diseases.
2.2. Ethical approval and measurements
This study was approved by the local ethics committee of the Medical University of Graz (ethical approval number 31-501 ex 18/19) and by the Allgemeine Unfallversicherungsanstalt (AUV A) ethical committee. All participants read and signed an informed consent form before the start of the experiment and were paid for their participation. The EEG measurements of the participants with SCI took place at the AUV A Rehabilitation Clinic Tobelbad, and the EEG measurements of the control participants took place at the Graz University of Technology.

2.3. Hardware and electrode layout
We recorded EEG data with a sampling rate of 500 Hz using BrainAmp amplifiers and ActiCap caps (Brain Products, Munich, Germany) with 61 active electrodes positioned in a 10–10 layout, as detailed in figure 1 of the supplementary material (available online at stacks.iop.org/JNE/18/046022/mmedia). The ground electrode was placed on AFz and the reference electrode was placed on the right mastoid. Additionally, we placed three electrodes above the nasion and below the outer canthi of the eyes to record electrooculographic (EOG) activity.

2.4. Experimental setup
Similarly to the experimental setup described in [35], the participants sat in front of a table, on top of which was a wooden 4-sided box, with its open sides facing towards the participant and the tabletop, as depicted in figure 1 (top). On the ceiling of the box was a Leap Motion device (Leap Motion, San Francisco, United States) that tracked each participant’s right or left hand, according to their preferred hand. The position of the Leap Motion on the ceiling of the box was adjusted to the handedness of each participant. Each participant kept their hand inside the wooden box. On the right side of the participants, attached to the table, we placed a robotic arm (Jaco Assistive robotic arm—Kinova Robotics, Bonn, Germany). A screen monitor was placed on the wooden box, centred in relation to the robotic arm. Opposed to the situation described in [35], the monitor lied horizontally on the box, with a slight inclination of 15 degrees, to offer the participants a better view of the screen. This change, relative to [35], was introduced to minimize head and eye movements during the experiment.

2.5. Controlling the robotic arm
During the trials, participants were able to control the robotic arm on a horizontal plane by moving their preferred hand on the tabletop. To reduce the range of the participants’ movements, we implemented the robot’s hand displacement to be three times larger than the participants’ hand displacement. Many participants with SCI had a very closed fist, due to hand spasticity caused by their injury, and this impaired their hand’s recognition by the Leap Motion. When this occurred, we inserted a small object into the participant’s hand to sustain the hand in a more open position and facilitate tracking.

2.6. Experimental overview
Before the experiment, we recorded one block in which the participant performed eye movements [48, 49]. The experiment then consisted of eight blocks of 30 trials each. Thirty percent of the trials of each block were error trials (9 trials). The remaining 70% of the trials were correct trials (21 trials). The sequence of correct and error trials within each block was randomly generated using a uniform distribution. We defined a maximum of two consecutive error trials in each block and repeated the randomization procedure until the sequence of trials satisfied this condition. Similarly, the trials of each block were equally split between the right and the left targets. The sequence of targets within each block was randomly assigned using a uniform distribution. We defined a maximum of three consecutive trials with the same target in each block and repeated the randomization procedure until the targets’ sequence satisfied this requirement.

All the eight blocks were online blocks: we used a generic ErrP classifier in an asynchronous manner to give participants real-time feedback about the ErrP detections during the experiment. To increase the fluidity of the experiment, we decided not to give participants feedback about the false positive ErrP detections, i.e. of the ErrP detections that happened when no error had occurred. This decision ensured that all participants experienced the same number of errors, which aimed to create a comparable expectation regarding the occurrence of errors across participants. False positive ErrP detections can occur in both correct and error trials and were considered when evaluating the classifier. The details regarding the construction of the generic classifier are described in section 2.10.

2.7. Experimental protocol
During the pre-trial period, the monitor displayed two squares on the top part of the screen, both with a side length of 14 cm. As depicted in figure 1, one of the squares was filled in white and the other square had no fill. The filled square represented the target of the forthcoming trial. The centres of the squares were 35 cm apart and their midpoint was located
Figure 1. Top: experimental setup. Participants sat in front of a table that had a robotic arm attached. The participants controlled the robotic arm during the trials using their hand. Bottom: the experimental protocol was displayed on the monitor. During the pre-trial period, the white square represented the target of the forthcoming trial. The small rectangle located centrally, at the bottom of the screen, represented the home position of the participant’s hand. A trial started when the participant moved their hand to its home position. The participants were instructed to move the robot to the target square during the trials. After the trial (post-trial period), the target changed colour, indicating whether it had been reached, and the robot automatically returned to its home position.

Participants could decide when to start a new trial and could rest for as long as they needed in between trials. A trial started when the dot entered the rectangle. This ensured that the participant’s hand was at a similar position at the beginning of each trial. Participants were instructed that when they felt ready to start a new trial, they should position the dot representing their hand below the home position’s rectangle, fixate their gaze on the target and finally enter the rectangle from the bottom. This last step ensured a forward movement of the robot. Participants were also asked to keep their gaze fixed on the target during the entire trial to prevent eye movements.

The aim of each trial was to move the robot’s hand from its home position to the target square. During the trials, only the two squares were displayed on the screen: the white square representing the target and the square with no fill. A trial ended when the robot’s hand was above the target or after 6 s (timeout), in case the target had not been reached. After the end of the trial (post-trial period), the target’s colour changed from white to either green or red, for 1.2 s, indicating whether or not the target was reached, respectively. This feedback was always in line with the robot’s behaviour. Afterwards, the screen then turned black, the robot automatically returned to its home position, and a new pre-trial period started.

2.7.1. Error trials
In each of these trials, an error was triggered during the movement of the robot towards the target. The error consisted in interrupting the participant’s control of the robot and adding a 5 cm upward displacement to the robot’s hand. The participants perceived the error by noticing the robot stopping and raising itself, and by realizing that the control of the robot was lost. The errors occurred randomly, when the robot’s hand was within 6 to 15 cm from its home position, in the forward direction. For every error trial, this distance was drawn from a continuous uniform distribution. In participants with SCI, the error onset occurred, on average, $1.36 \pm 0.14$ s after the start of
the error trial (mean ± std). In control participants, the error onset occurred, on average, 1.30 ± 0.07 s after the start of the error trial (mean ± std).

We used the generic ErrP classifier asynchronously to give participants feedback about the ErrP detections occurring after the error onset. Figure 2 illustrates all the possible interactions between the participants and the robot during error trials, taking into account the generic ErrP classifier feedback. If no ErrP was detected after the error onset, the robot remained still for the rest of the trial. In this situation, the total duration of the trial was 6 s and the target square then turned red. Conversely, if an ErrP was detected by the classifier after the error onset, the robot’s hand was lowered by 5 cm and the participants regained its control. The
downward movement informed the participants of the ErrP detection and their consequent regain of control. Since participants instinctively stopped their hand movement when they noticed the error, they were instructed to reinitiate the movement and move the robot’s hand to the selected target, when regaining control of the robot. To accommodate the extra movement, we added 6 s to the maximal trial duration when the first ErrP detection after the error onset occurred. If the robot reached the target after the error onset, the target square turned green. Participants did not receive feedback about the false positive ErrP detections occurring during the error trials, i.e. of the ErrP detections occurring before the error onset. Before the experiment, participants were informed that errors would occur and were shown the characteristic robot movement associated with error occurrence, i.e. the robot stopping and raising its hand.

2.7.2. Correct trials

In these trials, the paradigm did not trigger any error. Participants did not receive feedback about the false positive ErrP detections occurring during the correct trials. Figure 2 illustrates all the possible interactions between a participant and the robot during the correct trials. The correct trials lasted, on average, 2.11 ± 0.17 s for participants with SCI and 2.05 ± 0.13 s for control participants (mean ± std). All participants reached the target in over 99.4% of the correct trials.

2.8. Data processing

Eye movements and blinks were removed online from the EEG data, using the subspace subtraction algorithm [48, 49] and the eye-movement data recorded right before the start of the experiment. For the online detection of ErrPs with the generic classifier, the EEG data were bandpass filtered between 1 and 10 Hz with a fourth-order causal Butterworth filter. For the offline electrophysiological analysis presented here, the EEG data were bandpass filtered between 1 and 10 Hz with a fourth-order non-causal Butterworth filter.

2.9. Defining events

In the error trials, we defined the error onset as the moment at which the robot started its upward displacement once the participant lost its control. Before the experiment, we calculated the robot’s delay on 100 uncorrected errors, i.e. the time difference between the error marker and the robot’s upward displacement. This resulted in an average delay of 0.225 ± 0.005 s (mean ± std). Since the robot’s delay was rather stable, we added the average delay to each recorded error marker to obtain the error onset.

The correct trials had no clear onset. Therefore, to obtain comparable onsets in the correct and error trials for the electrophysiological analysis, we defined a virtual onset for the correct trials at a time point at which errors could occur in the error trials. For every participant, we defined the virtual onset for his/her correct trials as the average time difference between the error onsets and the start of the corresponding trials. For participants with SCI, the correct onset occurred, on average, 1.36 ± 0.14 s after the start of the correct trials (mean ± std). For the control participants, the correct onset occurred, on average, 1.30 ± 0.07 s after the start of the correct trials (mean ± std).

2.10. Generic ErrP classifier

We built a generic ErrP classifier using the EEG data from 15 able-bodied participants of a previous study of ours [35]. None of these previous participants took part in the experiment described here. The EEG data from those participants were filtered between 1 and 10 Hz using a fourth-order causal Butterworth filter. Eye movements were removed from the data using the subspace subtraction algorithm [48].

For each participant from [35], we used the eight calibration runs of the dataset and extracted an epoch with 450 ms from every trial. In the error trials, the selected epoch started 300 ms after the error onset. In the correct trials, the selected epoch started 300 ms after the virtual onset. Hence, our initial features were the amplitudes of the 61 EEG electrodes at all the time points of the 450 ms of each epoch.

To remove outlier epochs, we first applied principal component analysis (PCA) to the initial features and kept the PCA components that explained 99% of the data variability. We then removed 1% of the correct epochs and 1% of the error epochs as outliers. The rejection criterion was based on a large Mahalanobis distance of the rejected epochs within each class type (error or correct) in the PCA space. After this step, 2475 correct epochs and 1059 error epochs were kept.

Finally, we repeated the PCA step on the initial feature space, after discarding the outlier epochs, and kept the PCA components that preserved 99% of the data variability as features. This step resulted in 412 PCA components. These components were then used as features to train a shrinkage-LDA classifier with two classes: error and correct [50]. The linear scores of the classifier were transformed into probabilities using a softmax function. The PCA components preserved most of the activity of the original space, as depicted in figure 2 of the supplementary material. Figure 3 of the supplementary material depicts the classifier pattern, obtained by applying the discriminant feature analysis (DFA) method to the training matrix with 3534 epochs and 412 features [51]. The generic classifier remained unchanged during the entire experiment. In [45], we showed that the generic ErrP classifier offered a comparable performance to a personalized ErrP classifier for the asynchronous
The metrics TNR and TPR used here address the asynchronous detection of ErrPs. Therefore, we chose not to retrain the classifier with the participants’ own data.

2.11. ErrP detection
Similarly to the classifier developed in [35], the generic classifier developed here was constructed to be used and evaluated asynchronously. In the online experiment, the incoming EEG signals were analysed in real-time by the ErrP classifier, which received an EEG window of 450 ms as input. Consecutive analyzed windows had a leap of 18 ms. The classifier’s evaluation of each window resulted in the probability of the analysed window belonging to either class (correct or error). Hence, the classifier produced a probability output every 18 ms, during the entire duration of each block. We defined an ErrP detection as the occurrence of two consecutive windows with a probability of belonging to the error class above a certain threshold \( \tau \). In [44], we performed an offline evaluation of the asynchronous ErrP detection with the generic classifier and tested the effect of varying the decision threshold. From [44], we concluded that the combination of the generic ErrP classifier with a personalized decision threshold led to the achievement of a better performance. Hence, in this online experiment, we decided to apply this strategy. The procedure for determining the personalized thresholds is described in section 2.14.

2.12. Metrics for evaluating the ErrP classifier
To evaluate the performance of the generic classifier, we considered the trial structure of the experiment and the asynchronous nature of the decoding. The proposed metrics assess a trial as successful or unsuccessful, based on the asynchronous detection of ErrPs over the entire trial’s duration. This strategy has been applied to the study of asynchronous detection of ErrPs and other event-related potentials in several other works [29–35, 44, 45, 52–54]. Figure 2 presents a graphical representation of the metrics proposed here. The correct trials were labelled negative and the error trials were labelled positive.

2.12.1. True negative trials
We defined the true negative trials (TN trials) as the correct trials in which no ErrP detection occurred during the entire trial duration. For the classifier’s evaluation, we considered the true negative rate (TNR): the fraction of correct trials that are TN trials, i.e. that have no ErrP detections.\(^4\)

2.12.2. True positive trials
We defined the true positive trials (TP trials) as the error trials in which no ErrP detection occurred before the error onset and where at least one ErrP detection occurred within 1.5 s after the error onset. For the classifier’s evaluation, we considered the true positive rate (TPR): the fraction of error trials that were TP trials. An additional metric, the ErrP detection rate (EDR), considering only the ErrP detections within 1.5 s after the error onset, is defined in figure 5 of the supplementary material, where its relation with the TPR is detailed.\(^4\)

2.12.3. Chance level
To calculate the chance level for TNR and TPR, we performed several classifications with a classifier in which the training labels were randomly permuted (500 times to evaluate the online detection with the generic classifier and 50 times to evaluate the offline cross-validation with a personalized classifier). Furthermore, we used permutation-based \( p \)-values to present the significance of the classification results obtained with the generic ErrP classifier [55, 56].

2.13. False activation rate
The false activation rate (FAR) is the percentage of 1 s-long intervals that are contaminated with at least one false positive ErrP detection [57]. For this evaluation, we considered the entire duration of the correct trials and the period before the error onset in the error trials. These periods were divided into 1 s-long intervals and these intervals were evaluated for the presence of false positive ErrP detections.

2.14. Tailoring the decision threshold of the generic classifier to each participant
In [44], we evaluated offline the asynchronous detection of ErrPs with a generic classifier similar to the one described here. There, we observed that the decision threshold that maximized the group performance was \( \tau = 0.7 \). Moreover, we also concluded that to optimize individual performance with the generic classifier, participants benefited from the use of a personalized threshold. Therefore, in this experiment, we decided to initiate the generic classifier with \( \tau = 0.7 \) in the first block. This enabled us to skip offline calibration and allowed us to give participants immediate feedback about their ErrP detections. Afterwards, we tailored \( \tau \) to each participant. After each of the first three blocks, we performed an offline asynchronous classification with the generic ErrP classifier on all the available data and tested thresholds between 0 and 1 in steps of 0.025. For each of the 41 thresholds analysed, we calculated the corresponding TPR and TNR. The TNR and TPR curves were further smoothed using a moving average with seven samples. The smoothed curves were called the smooth TPR and the smooth TNR. For every participant, we chose the threshold that maximized the product of the smooth TPR and the smooth TNR. This was considered to be the

\(^4\) The metrics TNR and TPR used here address the asynchronous detection of ErrPs in a trial-based scenario and are not directly comparable with the TPR and TPR definitions commonly used in time-locked classification.
threshold that maximized performance and it was used in the next block. From block four onwards, the generic ErrP classifier was combined with the threshold \( \tau \) obtained after the third block. The generic ErrP classifier was not retrained with the participants’ data and only the decision threshold was updated based on the data.

2.15. Evaluation of the generic ErrP classifier

We stopped tailoring \( \tau \) to each participant after the third block because we wanted to collect a substantial amount of data in unchanged conditions. From blocks four to eight, all participants used the generic classifier with a fixed but personalized threshold. Therefore, we only used the data from blocks four to eight to evaluate the performance of the generic classifier, ensuring comparable conditions across the participants.

2.16. Personalized ErrP classifier

To evaluate offline the performance of a personalized classifier, we performed 10 times a 5-fold cross-validation with the entire dataset of each participant, where a personalized classifier was tested asynchronously in each fold. We also tested all thresholds from 0 to 1 in steps of 0.025. For every participant, we obtained, in each fold, a TPR and a TNR for every threshold tested. For every participant, we averaged the TPR and TNR of the 50 iterations in the cross-validation, obtaining an average TPR and an average TNR per participant. Finally, we selected the threshold that maximized the product of the average TPR and the average TNR, for every participant. The evaluation of the personalized classifier followed the metrics defined in section 2.12.

3. Results

3.1. Neurophysiology

The electrophysiological results presented here comprise the entire recorded dataset. Figure 3 shows the grand average correct and error signals at channel FCz (green and red lines, respectively) for participants with SCI and control participants. The green and red shaded areas depict the 95% confidence interval for the grand average signals. The figure also displays the topographic plots of the grand average correct and error signals at the time points of the peaks in the grand average error signal.

As the morphology of the error signals was not homogeneous across participants, we found it relevant to also present the electrophysiological results of the individual participants. Figure 4 displays the average correct and error signals at channel FCz (green and red lines, respectively) of every participant. The green and red shaded areas depict the 95% confidence interval for the average signals. The grey areas indicate the time points at which the correct and error signals were statistically different (Wilcoxon rank-sum tests,
Figure 4. Average correct and error signals at channel FCz of every participant (green and red lines, respectively). The shaded areas represent the 95% confidence interval of the average signals. The black line at $t = 0$ s represents the error onset of the error trials and the virtual onset of the correct trials. The grey regions indicate the time points at which the correct and error signals were statistically different (Wilcoxon rank-sum tests, Bonferroni corrected, with $\alpha = 0.01$).

3.2. Adaptation of the classifier’s threshold in the first three experimental blocks

This experiment required no offline calibration and the participants received feedback regarding their ErrP detections from the very beginning. This was made possible by combining the generic ErrP classifier with a generic decision threshold ($\tau = 0.7$) for the first experimental block. However, we still used the first three experimental blocks to reach a fixed personalized decision threshold. After each of the first three blocks, we updated the decision threshold $\tau$ to maximize the participant’s performance. Hence, participants used a generic classifier combined with a personalized decision threshold from block two onwards. Figure 5 (top) depicts the initial threshold ($\tau = 0.7$) and the calculated thresholds after each of the first three blocks, for every participant. At the end of block three, the average threshold was $\tau = 0.68$ for participants with SCI and $\tau = 0.59$ for control participants. Figure 5 (bottom) shows the TNR and TPR obtained offline after block three, for all tested thresholds (green and red dashed lines, respectively). It also shows the smooth TNR and smooth TPR obtained with a moving average (green and red solid lines). The black dotted line depicts the product of these smooth curves and the blue vertical line indicates the threshold that maximizes it. This is the decision threshold used for every participant from blocks four to eight.
3.3. Evaluation of the online asynchronous classification using a generic ErrP classifier

To evaluate the asynchronous classification results obtained with the generic ErrP classifier during the experiment, we only considered the data from the last five blocks of the experiment, i.e. from blocks four to eight, since no parameters were changed during these blocks.

Figure 6 (top) depicts the classification results obtained with the generic classifier in terms of the TPR and the TNR. For participants with SCI, we obtained an average TPR of 46.9% and an average TNR of 71.9%. For the control participants, we obtained an average TPR of 56.4% and an average TNR of 77.9%. The circles on the individual bars represent the chance level of the corresponding
Figure 6. Evaluation of the generic ErrP classifier. Top: classification results in terms of the true positive rate (TPR) and the true negative rate (TNR). The circles on the individual bars represent the chance level of the corresponding metrics. Bottom: permutation-based $p$-value for the significance of the results obtained with the classifier.

Figure 7. Online detection of ErrPs with the generic classifier and trials’ offline evaluation for participant C1. Left: error trials, aligned to the error onset (black vertical line). Right: correct trials, aligned to the start of the trial. The dark grey areas represent the trials and the white marks within them represent the ErrP detections. The narrow rectangles colour code the trials’ offline evaluation. In these rectangles, successfully classified trials (true positive trials and true negative trials) are colour coded in white and trials with false positive ErrP detections are colour coded in light grey. Error trials with no ErrP detection are colour coded in black.

metrics. The chance level results for each participant were obtained by averaging the classification results of 500 classifiers in which the training labels were randomly permuted, and by considering the final participant-specific threshold, as depicted in figure 4 of the supplementary material. Figure 6 (bottom) presents the permutation-based $p$-values for the significance of the classification results [55, 56]. Figure 5 of the supplementary material depicts a comparison between the TPR and EDR metrics. Table 1 of the supplementary material presents the FAR in correct and error trials.

Figure 7 illustrates the online asynchronous detection of ErrPs with the generic classifier and the trials’ offline evaluation for participant C1. The dark grey areas represent the trials and the white marks within them represent the ErrP detections. The narrow rectangles colour code the trials’ offline evaluation. In these rectangles, successfully classified trials (true positive trials and true negative trials) are colour coded in white and trials with false positive ErrP detections are colour coded in light grey. Error trials with no ErrP detection are colour coded in black.
3.4. Offline evaluation of the asynchronous ErrP classification with a personalized classifier

To evaluate offline the asynchronous classification results with a personalized classifier, we considered the eight experimental blocks and performed ten times a five-fold cross-validation. As this evaluation was done offline, we tested thresholds from 0 to 1 with a leap of 0.025 and the results obtained are shown as a function of the threshold, $\tau$.

Figure 8 depicts the grand average TNR and TPR (green and red solid lines, respectively) as well as the grand average chance level for TNR and TPR (green and red dashed lines, respectively) as a function of the threshold. The shaded areas represent the 95% confidence intervals of the grand average curves. The chance level curves were obtained by performing 10 times a 5-fold cross-validation with 50 classifiers, in which the labels of the training trials were randomly permuted.

Figure 9 depicts, for every participant, the average TNR and TPR (green and red solid lines, respectively) and the chance levels of the TNR and TPR (green and red dashed lines, respectively). The blue vertical line indicates the threshold that maximizes individual performance with the personalized ErrP classifier.

Figure 10 depicts the average TNR and TPR obtained by cross-validation when using the optimal personalized decision threshold for every participant (green and red bars, respectively). The small circles on the bars indicate the chance level obtained for every participant with the considered threshold. For participants with SCI, the grand average TNR was 77.9% and the grand average TPR was 55.0%. For the control participants, the grand average TNR was 86.1% and the grand average TPR was 71.5%.

4. Discussion

In this work, we investigated the transfer of a generic ErrP classifier from able-bodied participants to participants with SCI. The classifier was developed using data from able-bodied participants from a previous experiment of ours [35] and was tested asynchronously in a closed-loop online experiment in which participants with SCI and able-bodied control participants took part. Using the classifier asynchronously, entire trials were evaluated and not only a time-locked window. The online experiment required no offline calibration period and the participants received feedback about the ErrP detections immediately, from the start of the experiment onwards. Additionally, we also analysed the morphology of ErrPs in participants with SCI and in able-bodied control participants.

The grand average correct signal displayed, as expected, no particular potential, in both participants with SCI and in control participants. The correct epochs correspond to the period in which the participants were continuously controlling the robot and are not associated with any specific event. The grand average error signal was associated with a fronto-central activity, in both participants with SCI and control participants. The peaks of the grand average error signal were less pronounced in participants with SCI than in control participants, as visible in figure 3. This matches the results described in [46]. Nevertheless, the electrophysiological patterns of participants with SCI were rather heterogeneous and half of the participants with SCI did not display the characteristic error-related activity (participants P4, P5, P7 and P8). The remaining participants with SCI revealed
Figure 9. Evaluation of the personalized ErrP classifier. Single-subject average TNR and TPR (green and red solid lines, respectively) and chance levels of the TNR and TPR as a function of the threshold (green and red dashed lines, respectively). The shaded areas indicate the 95% confidence interval for the average curves. The blue vertical line indicates the threshold that maximizes individual performance.

Figure 10. Evaluation of the personalized ErrP classifier. Average TNR and TPR were calculated by the cross-validation procedure, with an optimal personalized threshold, for every participant and their average. The small dots on each bar indicate the chance level for the considered threshold, for every participant.
patterns comparable to those of the control participants. Therefore, we believe that in our study, the decrease in peak amplitudes observed in the grand average error signals of participants with SCI was not directly related to the injury, but was rather a consequence of the heterogeneity of the signals in the population with SCI. Several studies have reported the effects of psychological factors, such as depression and anxiety, on ErrPs [58, 59]. The population with SCI is particularly vulnerable to emotional disorders and higher levels of distress [60–62]. Nevertheless, the individual differences were large [62]. The possibility of a psychiatric evaluation and medication of the participants would have been valuable for the current work and should be considered in future studies involving a population with SCI [63]. Interestingly, the error signals of the control participants were also less homogeneous than in our previous studies with a similar experimental protocol [34, 35]. Several studies showed that ageing affects error processing and consequently ErrPs; hence, we hypothesise that the higher variability observed in the signals of the control participants in this study is related to the wider age range of the participants, in comparison with our previous studies [64–66].

To interpret the classification results, we focus on the TPR. This metric considers an interval after the error onset and the period before the error onset. Hence, it translates not only the classifier’s ability to detect ErrPs after the occurrence of an error but also its ability to avoid detecting ErrPs when no error occurs. The TNR only captures the classifier’s ability to avoid detecting ErrPs when no error occurs. It is still a meaningful metric, but the TNR’s outcome can be artificially increased by the use of a high decision threshold, as depicted in figure 4 of the supplementary material. The classification results for the generic classifier were, on average, lower in participants with SCI than in the control participants. Only half of the participants with SCI obtained a TPR above chance level. These participants were the ones that displayed clear error patterns. In the control participants, seven out of eight participants obtained a TPR above chance level. The remaining participant (participant C5) did not obtain a TPR above chance level and did not display a very clear error signal. To summarise, all participants that displayed clear ErrP patterns in the electrophysiological analysis obtained better than chance results with the generic classifier, independently of the group (SCI or control). It would be rather interesting to further investigate the factors that affect the error patterns, independently of the SCI. These results support the view that using a generic ErrP classifier is a valuable option for giving immediate feedback to participants. Moreover, it indicates that ErrPs are transferable across participants and that the transfer can be applied to distinct populations, such as participants with SCI.

With the generic classifier developed, participants received real-time feedback about the ErrP detections from the beginning of the experiment. However, the first three blocks of the experiment were still used to update the threshold applied to the generic classifier. We made this choice because we had previously shown that some participants strongly benefit from combining the generic ErrP classifier with a personalized decision threshold [44]. For most participants, the threshold was relatively stable after the first block. This supports the use of a personalized threshold with the generic classifier, as suggested in [44]. In a real-world online application, the occurrence of errors can not be easily assessed, since it is determined by the subjective perception of the BCI user. Such a constraint hinders an objective evaluation of any ErrP classifier unless the participants can use a motor-based strategy to report the occurrence of errors. Still, our approach could be applied to a real-world asynchronous situation in which the occurrence of errors is unknown. Nevertheless, to establish a personalized decision threshold, our approach would need a short online application beforehand, in which the occurrence of errors was known. Such an application could be the equivalent of one of our experimental blocks, which contained nine errors and lasted less than 5 min.

In our experiment, we only gave the participants feedback about ErrPs detected after the error onset. This aimed to ensure that participants experienced the same number of errors and had comparable expectations regarding the occurrence of errors. Providing participants with feedback about the false positive ErrP detections would have brought our experiment closer to a real-world application, at the cost of putting participants in dissimilar circumstances, given that false positive ErrP detections could have affected their behaviour and the generation of ErrPs. For instance, participants with many false positive ErrP detections would certainly have been negatively affected by the feedback, either losing engagement or disregarding the feedback. Such participants would no longer have perceived the errors as meaningful and relevant, and this could have altered their ErrPs.

When testing offline the asynchronous classification with a personalized classifier, two participants with SCI (participants P4 and P5) and one control participant (participant C5) obtained chance level TPR results. This indicates that the signals of these participants were not sufficiently different to build a personalized classifier and that these participants obtained chance level results with both generic and personalized classifiers.

The classification results obtained with the personalized classifier are not directly comparable with the results obtained with the generic classifier, because the classifiers were evaluated using different datasets. In a real-world scenario, we could provide...
participants with immediate feedback about their brain signals using a generic classifier, while collecting data to train a personalized classifier. Simultaneously, we could compare the performance of the personalized and generic classifiers at regular intervals, and swap the generic classifier for a personalized classifier, once the latter would offer significantly better performance.

5. Conclusions

Our work shows that a generic ErrP classifier can be used, asynchronously and online, by participants with SCI and able-bodied participants. Moreover, the generic ErrP classifier is transferable from an able-bodied population to a population with SCI. The developed classifier required no previous calibration with the participant and granted immediate feedback about the ErrP detections. Therefore, our findings can help to promote the incorporation of ErrPs in BCIs for different types of users.

Acknowledgment

This work was supported by Horizon 2020 ERC Consolidator Grant 681231 ‘Feel Your Reach’. The authors would like to thank Reinmar J Kobler, Lea Heinenberger and Joana Pereira for the fruitful discussions. The authors would also like to thank the staff and patients of the AUVA-Rehabilitationklinik Tobelbad, for the warm welcome.

ORCID iDs

Catarina Lopes-Dias  @ https://orcid.org/0000-0001-9775-3743
Andreea I Sburlea  @ https://orcid.org/0000-0001-6766-3464
Gernot R Müller-Putz  @ https://orcid.org/0000-0002-0608-3720

References

[1] Brunner C et al 2015 BNCI Horizon 2020: towards a roadmap for the BCI community Brain-Comput. Interfaces 21–10
[2] Millán J D R et al 2010 Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges Front. Neurosci. 4 161
[3] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control Clin. Neurophysiol. 113 767–91
[4] Schalk G, Wolpaw J R, McFarland D J and Pfurtscheller G 2000 EEG-based communication: presence of an error potential Clin. Neurophysiol. 111 2138–44
[5] Ferré P W and Millán J D R 2005 You are wrong: automatic detection of interaction errors from brain waves Proc. 19th Int. Conf. on Artificial Intelligence (IJCAI’05) pp 1413–18 (https://www.ijcai.org/Proceedings/2005)
[6] Yeung N, Botvinick M and Cohen J D 11 2004 The neural basis of error detection: conflict monitoring and the error-related negativity Psychol. Rev. 111 931–59
[7] Chavarriaga R, Sobolewski A and Millán J D R 2014 Errare machinae est: the use of error-related potentials in brain-machine interfaces Front. Neurosci. 8 208
[8] Llera A, van Gerven M A, Gómez V, Jensen O and Kappen H J 2011 On the use of interaction error potentials for adaptive brain computer interfaces Neural Netw. 24 1120–7
[9] Llera A, Gómez V and Kappen H J 2012 Adaptive classification on brain–computer interfaces using reinforcement signals Neural Comput. 24 2900–23
[10] Yousefi R, Sereshkhe A R and Chau T 2019 Online detection of error-related potentials in multi-class cognitive task-based BCIs Brain-Comput. Interfaces 6 1–12
[11] Artusi X, Niazi I K, Lucas M F and Farina D 2011 Accuracy of a BCI based on movement-related and error potentials 2011 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society pp 5688–91
[12] Zhang H, Chavarriaga R, Khaliliardali Z, Gheorghe L, Iturrate I and Millán J D R 2015 EEG-based decoding of error-related brain activity in a real-world driving task J. Neural Eng. 12 066028
[13] Spüler M and Niethammer C 2015 Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity Front. Hum. Neurosci. 9 155
[14] Kreilinger A, Bielag H and Müller-Putz G R 2016 Single versus multiple events error potential detection in a BCI-controlled car game with continuous and discrete feedback IEEE Trans. Biomed. Eng. 63 519–29
[15] Si-Mohammed H, Lopes-Dias C, Duarte M, Argelaguet F, Jeunet C, Casiez G, Müller-Putz G R, Lécuyer A and Scherer R 2020 Detecting system errors in virtual reality using EEG through error-related potentials 2020 Conf. on Virtual Reality and 3D User Interfaces (VR) pp 653–61
[16] Parra L C, Spence C D, Gerson A D and Saajda P 2003 Response error correction—a demonstration of improved human-machine performance using real-time EEG monitoring IEEE Trans. Neural Syst. Rehabil. Eng. 11 173–7
[17] Kreilinger A, Neuper C and Müller-Putz G R 2012 Error potential detection during continuous movement of an artificial arm controlled by brain–computer interface Med. Biol. Eng. Comput. 50 223–30
[18] Zander T, Krol L, Birbaumer N and Gramann K 2016 Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity Proc. of the National Academy of Sciences of the United States of America vol 113 pp 14898–903
[19] Salazar-Gomez A F, DelPreto J, Gil S, Guenther F H and Rus D 2017 Correcting robot mistakes in real time using EEG signals 2017 IEEE Int. Conf. on Robotics and Automation (ICRA) pp 6570–7
[20] Kim S K, Kirchner F A, Stefes A and Kirchner F 2017 Intrinsic interactive reinforcement learning—using error-related potentials for real world human–robot interaction Sci. Rep. 7 17562
[21] Mousavi M, Koerner A S, Zhang Q, Noh E and de Sa V R 2017 Improving motor imagery BCI with user response to feedback Brain-Comput. Interfaces 4 74–86
[22] Ehrlich S K and Cheng G 2018 Human-agent co-adaptation using error-related potentials J. Neural Eng. 15 066014
[23] Spüler M, Bensch M, Kleih S, Rosenstiel W, Bogdan M and Kübler A 2012 Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI Clin. Neurophysiol. 123 1328–37
[24] Kobler R J, Sburlea A I and Müller-Putz G R 2018 Tuning characteristics of low-frequency EEG to positions and velocities in visuomotor and oculomotor tracking tasks Sci. Rep. 8 17713
[25] Ofner P and Müller-Putz G R 2012 Decoding of velocities and positions of 3D arm movement from EEG 2012 Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society pp 6406–9
[62] Post M W M and van Leeuwen C M C 2012 Psychosocial issues in spinal cord injury: a review Spinal Cord 50 382–9
[63] van Leeuwen C M C, van der Woude L H V and Post M W M 2012 Validity of the mental health subscale of the sf-36 in persons with spinal cord injury Spinal Cord 50 707–10
[64] Falkenstein M, Hoormann J and Hohnsbein J 2001 Changes of error-related ERPs with age Exp. Brain Res. 138 258–62
[65] Hoffmann S and Falkenstein M 2011 Aging and error processing: age related increase in the variability of the error-negativity is not accompanied by increase in response variability PloS One 6 e17482–e17482
[66] Nieuwenhuis S, Riddervikof K R, Talsma D, Coles M G H, Holroyd C B, Kok A and van der Molen M W 2002 A computational account of altered error processing in older age: dopamine and the error-related negativity Cogn. Affective Behav. Neurosci. 2 19–36