Perturbation-evoked potentials can be classified from single-trial EEG

Jonas C Ditz, Andreas Schwarz and Gernot R Müller-Putz

1 Institute of Neural Engineering, Graz University of Technology, Graz, Austria
2 Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Tübingen, Germany
3 Author to whom any correspondence should be addressed
E-mail: gernot.mueller@tugraz.at

Abstract

Objective. Loss of balance control can have serious consequences on interaction between humans and machines as well as the general well-being of humans. Perceived balance perturbations are always accompanied by a specific cortical activation, the so-called perturbation-evoked potential (PEP). In this study, we investigate the possibility to classify PEPs from ongoing EEG. Approach. Fifteen healthy subjects were exposed to seated whole-body perturbations. Each participant performed 120 trials; they were rapidly tilted to the right and left, 60 times respectively. Main results. We achieved classification accuracies of more than 85% between PEPs and rest EEG using a window-based classification approach. Different window lengths and electrode layouts were compared. We were able to achieve excellent classification performance (87.6 ± 8.0% accuracy) by using a short window length of 200 ms and a minimal electrode layout consisting of only the Cz electrode. The peak classification accuracy coincides in time with the strongest component of PEPs, called N1. Significance. We showed that PEPs can be discriminated against ongoing EEG with high accuracy. These findings can contribute to the development of a system that can detect balance perturbations online.

1. Introduction

The sense of balance is crucial for humans in their everyday routine. Standing and walking are not possible without it. Every human being learns balance control during early childhood and the loss of balance control always leads to uncomfortable, often potentially dangerous situations. The possibility to compensate for loss of balance control can alleviate harmful consequences and, therefore, vastly improve human experience. In gait rehabilitation, exoskeletons are used to support patients during rehabilitation sessions (Veneman et al 2007). However, with the ability to compensate for loss of balance control, the support provided by an exoskeleton can be limited to an on demand state. Thereby, the independence of patients can be increased for better rehabilitation. The system takes control from the patient if it is needed to prevent falling.

In virtual reality (VR), the conflict between sensory and vestibular information can lead to different physiological effects, inter alia, postural instability (Cobb et al 1999). If the system is able to detect postural instability, i.e. perceived balance perturbation, alternative visualization protocols as well as emergency shut downs of the visual environment can be put into action as soon as they are needed. Nevertheless, a reliable and potent detection method for the loss of balance has to be found.

Electroencephalography (EEG) studies found a specific activity pattern that was elicit as a response to a balance perturbation (Dietz et al 1984, Ackermann et al 1986, Duckrow et al 1999, Staines et al 2001, Adkin et al 2006). This cortical activity, called perturbation-evoked potential (PEP), consists of four distinguishable parts. After an initial small positive wave (P1), a large negative deflection (N1) follows. The third and fourth part is again a positive wave (P2) followed by a negative wave (N2). The last two parts are often collectively referred to as late perturbation-evoked response (PER). Timings of the different parts usually reported
by researchers are 30–90 ms after perturbation onset for P1, 80–160 ms after perturbation onset for N1, and 200–400 ms after perturbation onset for late PEPs (Varghese et al 2017). PEPs have been thoroughly investigated on the neurophysiological level, linking them to cortico-cortical transfer processes (Dimitrov et al 1996), error-potentials (mismatch between actual and expected position) (Adkin et al 2006), and compensatory motor planning processes (Marlin et al 2014). Regardless of the underlying neural processes, loss of balance control is always accompanied by a PEP (especially the N1 component is always reproducible) independent of the mode of perturbation (Varghese et al 2017). Therefore, a system that can reliably detect changes in state of mind (i.e. occurrences of specific neural activation patterns) can be used for the detection of PEPs.

Control over a computer or machine by solely using one’s mind is the main goal of research in the field of Brain–computer interfaces (BCIs) (Wolpaw et al 2002, Millán et al 2010, Wolpaw and Wolpaw 2012). While there has been a strong focus on using BCIs for controlling assistive devices (Bhagat et al 2016, Crea et al 2018), BCIs can improve the interaction between humans and machines in the context of human–machine interaction (HMI). These so-called passive BCIs (pBCI) do not provide active control to users; moreover, they monitor their state of mind and detect changes in the state of mind of users (Zander et al 2009, Zander and Kothe 2011). Studies have shown that implicit information about the state of mind of a user can be found in distinct brain patterns. Working memory load can be detected by monitoring oscillatory power in theta band over frontal-midline electrodes (Gerjets et al 2014). Error-related potentials are specific activity patterns that are elicited when a user makes or perceives an error and are used to compensate for erroneous interactions (Scheffers and Coles 2000, Parra et al 2003, Lopes Dias et al 2018). Another potential that is used in pBCIs is the Bereitschaftspotential (BP), which precedes spontaneous movements (Shibasaki and Hallett 2006, Schultzze-Kraft et al 2016). Due to their ability to detect changes in user’s mental state, pBCIs are a promising tool for detecting perceived balance perturbation.

In this study, we investigate whether PEPs can be autonomously discriminated from ongoing EEG. Fifteen healthy participants were exposed to seated whole-body perturbations. Each participant performed 120 trials; they were 60 times rapidly tilted to the right and 60 times rapidly tilted to the left.

We developed a method that can decode PEPs from ongoing EEG recordings. Additionally, we evaluated parameters imperative for boosting PEP classification for existing pBCI systems such as different window lengths and smaller electrode layouts. Finally, we constructed an offline scenario to test our PEP detection method.

2. Materials and methods

2.1. Participants

Fifteen healthy participants (six female, nine male) took part in the study. Participants were between 19 and 57 years old with an average age of 26.7 ± 9.4 years. All participants had normal or corrected-to-normal vision and were without any known medical condition. The study was approved by the ethics committee of the Medical University of Graz. All participants gave written informed consent and received monetary compensation for their efforts.

2.2. Hardware and data acquisition

We recorded EEG from 29 active electrodes using a LiveAmp system (BrainProducts, Gilching, Germany). Electrodes were distributed over frontal, central, and parietal areas at positions F1, Fz, F2, FFC1h, FFC2h, FC1, FC3, FCz, FC2, FC4, FCC1h, FCC2h, C5, C3, C2, Cz, C4, C6, CCP1h, CCP2h, CP3, CP1, CPz, CP2, CP4, P1, Pz, and P2. Additionally, we used three electrodes to record electro-ocular activity (EOG) and placed them above the nasion and below the outer canthi of the eyes. The ground electrode was placed at position Afz and all electrodes were referenced to the left mastoid. EEG was sampled at 500 Hz. All signals were pre-filtered using a 3rd order sinc low-pass at 13 Hz (cutoff). We also recorded the participant-specific electrode positions using an ultrasound-based ELPOS system manufactured by Zebris (Zebris Medical GmbH, Isny im Allgäu, Germany).

2.3. Experimental task

All measurements took place in the BCI-Lab of the Institute of Neural Engineering at Graz University of Technology.

Participants were equipped with an electrode cap and seated in a custom-built tilting chair. Using a mechanical tilting system, we were able to tilt the chair 5° to the left or to the right. The amplifier was fixed to the back of the chair and a fixation cross was put on the wall in front of the chair (see figure 1). Participants were asked to take a comfortable position in the chair and rest their arms on their legs in order to reduce muscle tension in the arms and shoulders. We further instructed them to stay relaxed and fixate the cross in front of them during the whole experiment. In the beginning of the experiment, two perturbations (one perturbation to the right, one to the left) were performed to familiarize them with the task at hand. Subsequently, each trial had the following structure: Within the first four seconds of each trial at a randomly chosen time point, participants were rapidly tilted either to the left or the right (based on a random generator). The chair was tilted manually...
using a mechanical lever. The chair stayed tilted until 8 s after trial start. Thereafter, the chair was put back into the neutral position for an inter-trial interval of 4 s. In this way, each participant experienced 60 perturbations to the left and 60 to the right in random order (120 perturbations in total).

Each event during the experiment was indicated by a marker. The synchronization of marker data and amplifier was realized using LabStreamingLayer (LSL) (Kothe 2020).

2.4. Perturbation onset detection

We recorded the perturbation onset for each trial using the intrinsic accelerometer (3 axes) of the LiveAmp, which was fixed on the backside (upper left corner) of the tilting chair (see figure 1). We applied thresholding on the first derivative of the abscissa (x axis in figure 1) to acquire trial based perturbation onsets.

2.5. Data preprocessing

In order to detect artefact-contaminated trials and exclude them from further analysis, we performed statistical tests. For this purpose, data was band-pass filtered between 0.3 and 35 Hz (zero-phase Butterworth filter, 4th order) and three different statistical parameters were considered for the rejection of artifact-tainted trials. First, we performed an amplitude threshold rejection removing all trials with an amplitude that exceeded ±125 µV. Afterwards, we tested trials for an abnormal joint probability and an abnormal kurtosis. The rejection threshold was four times the standard deviation (STD) for both tests. 13.8% of the trials were rejected on average. The approach used for outlier rejection does not need additionally recorded channels and is well tested in different BCI scenarios (Faller et al. 2012, Schwarz et al. 2015).

After artifact rejection, we band-pass filtered the raw EEG between 0.3 and 10 Hz using a causal 4th order Butterworth filter.

2.6. Perturbation-evoked potential

We epoched the EEG from −0.5 s to 1.5 s with respect to the perturbation onset acquired from the accelerometer. Additionally, we acquired rest trials with a length of 2 s 4.5 s prior to the perturbation onset from −4.5 s to −2.5 s. Therefore, our time region of interest (tROI) had a duration of 2 s. For each participant, we calculated the average over all trials for each condition (perturbation, rest) as well as the 95% confidence interval using nonparametric t-percentile bootstrap statistics (α = 0.05) (Pérez and Benjamin Knapp 2007). Additionally, to account for statistically significant differences between conditions, we performed the nonparametric Wilcoxon Rank Sum test (α = 0.01) on each time point. We corrected for
multiple comparisons (with $n = 1250$ comparisons) using the Bonferroni correction. 3

2.7. Binary single-trial classification
We used a two-phase window-based approach comparable to the method used in (Schwarz et al 2018, 2019) for the classification of PEPs. A schematic of the classification pipeline can be found in the supplementary material (figure S1) (available online at stacks.iop.org/JNE/17/036008/mmedia). In the beginning, we resampled our data to 25 Hz in order to save computational effort and split our data into two sets: calibration set (containing two thirds of the data) and test set (containing the remaining third of the data). In order to simulate online behavior, we took the trials that were recorded first for the calibration set and the trials that were recorded last for the test set.

In the first phase, called the calibration phase, we trained shrinkage linear discriminant analysis (sLDA) classifiers (Frieman 1989, Blankertz et al 2011) using only trials in our calibration set. In order to train the classifiers, we performed a 10 times 5-fold cross-validation. In each fold, we moved a window over our tROI in steps of 40 ms, i.e. we trained a classifier every 40 ms (2000 ms/40 ms + 1 = 51 classifiers for the whole tROI). For each time point, features were extracted by taking amplitude values of each electrode in steps of 40 ms, i.e. the number of features varies depending on the used window length (see table 1 for an overview of tested window sizes). The number of features of, for example, a 200 ms window is 200 ms/40 ms + 1 = 6, where we add one since the first sample in the window is included as a feature. We separated the calibration set into training and validation set in each cross-validation fold and trained the classifiers using only trials from the training set. Afterwards, we evaluated each classifier using the validation set. After cross-validation, we calculated the mean validation accuracy for each time point and used the best performing classifier of the time point with the highest mean validation accuracy for the second phase.

In the second phase, called the test phase, we extracted features from the unseen test set similar as in the calibration phase. We applied the trained classification model to the features of the test set. Accuracies stated in this work are always test accuracies, i.e. the classification performance of our classifier on the test data.

2.8. Classification parameter optimization
We tested four different window lengths. The parameter for each of the used window lengths can be found in table 1. Additionally, we tested four different electrode layouts:

3A bonferroni correction for 1250 comparisons using a significance level of $\alpha = 0.01$ leads to significant p values at $p < 0.000008$.

### Table 1: Different window lengths used for binary classification of PEPs with the corresponding number of features. For each window the subject-specific offset with the highest discriminability between rest trials and PEP trials was deduced in a calibration phase.

| Window size | Number of features per channel | Number of features per trial | Trial-to-feature ratio |
|-------------|-------------------------------|-----------------------------|-----------------------|
| 1 sample    | 1                             | 29                          | 8.28                  |
| 200 ms      | 6                             | 174                         | 1.38                  |
| 400 ms      | 11                            | 319                         | 0.75                  |
| 600 ms      | 16                            | 464                         | 0.52                  |

- minimal electrode layout (1 channel): Cz
- small electrode layout (5 channels): FCz, C1, Cz, C2, and CPz
- medium electrode layout (15 channels): FC1, FCz, FC2, FCCh1, FCCh2, C3, C1, Cz, C2, C4, CCPh1, CCPh2, CP1, CPz, and CP2
- full electrode layout (29 channels)

For each setup (combination of window length and electrode layout), the classification accuracy was calculated as described above. A two-way anova was conducted to test for a significant effect of the two independent variables (window length, electrode layout) on the classification performance. We tested the effect on peak accuracy as well as peak latency. Each independent variable had four levels (1 sample, 200 ms, 400 ms, and 600 ms for window length; minimal, small, medium, and full for electrode layout). Afterwards, we compared the classification performance of different window lengths as well as the classification performance of different electrode layouts.

3. Results

3.1. Perturbation-evoked potential
Figure 2 shows the grand average PEP for channel Cz (solid blue line) as well as individual PEPs of each participant (dashed blue lines). In perturbation trials, the grand average shows a strong negative shift (N1 component of PEPs) starting shortly after perturbation onset and peaking 144 ± 9 ms after perturbation onset with an average amplitude of $-28.3 ± 14.5$ $\mu$V. The negative peak is followed by a strong positive rebound, the P2 component of PEPs, that peaked on average 330 ± 30 ms after perturbation onset with an amplitude of 12.2 ± 4.1 $\mu$V. Neither the P1 nor the N2 component were clearly detectable in the grand average. We also show the grand average on a topographical level for timepoints $t = 0$ ms (perturbation onset), $t = 146$ ms (N1 peak) and $t = 330$ ms (P2 peak): At perturbation onset (0 ms) on average, no visible derivation from the baseline is found. While the peak of the N1 component was centered around Cz and distributed over frontal, central, and parietal
Figure 2. Grand average of perturbation trials at channel Cz. Dashed lines show the average for each of the participants. The vertical line at time point 0 ms indicates the perturbation onset. Topographical plots show the scalp distribution of the grand average at $t = 0$ ms, $t = 146$ ms, and $t = 320$ ms. The N1 component (topographical plot at $t = 146$ ms) is distributed over frontal, central, and parietal areas with the center at Cz. The P2 component (topographical plot at $t = 320$ ms) is distributed over parietal areas with the center at Pz.

Figure 3. Grand average of perturbation (blue) and rest (grey) trials. Furthermore, the grand average of 'left perturbation' trials is shown in red and the grand average of 'right perturbation' trials is shown in green. Shaded areas show the 95% confidence interval (CI) over all subjects. Significantly different time points (Wilcoxon test, $\alpha < 0.01$, with Bonferroni correction for $n = 1250$ comparisons) between perturbation and rest are marked beneath the plot. Similarly, significantly different time points between left and right trials are also marked beneath the plot.

areas, the P2 peak was mainly distributed over parietal areas and centered around Pz.

Figure 3 shows the participant based grand average for channels positioned at FCz, Cz, CPz and Pz as well as positions C1 and C2. The left side compares the perturbation condition versus the rest condition. The black bars beneath each plot indicates significantly different time intervals between rest condition and perturbation condition (Wilcoxon test, $\alpha < 0.01$, with Bonferroni correction for $n = 1250$ comparisons). The right side of figure 3 compares trials where the participant was tilted to the left with trials where the participant was tilted to the right. Similar to the left side, significantly different time intervals between both conditions are indicated beneath each plot.
Grand average of both perturbation and rest conditions show significant differences in morphology starting around 40 ms after perturbation onset. At C1, Cz, C2, CPz and Pz, both N1 and P2 components are significantly different to the rest condition. At FCz, only the N1 component is significantly different compared with the rest condition. Between 500 and 2000 ms after perturbation onset, a small negative peak followed by a positive rebound can be observed for all four channels. The difference of this behavior compared to the rest condition is significant for all channels displayed on the left side of figure 3. The N1 component showed no significant difference between right and left perturbations, as can be seen in figure 3. The P2 component had a slightly higher amplitude for left tilting trials (16.7 ± 5.3 µV at Pz) compared to trials where the subject was tilted to the right (10.3 ± 5.6 µV at Pz). To test the effect of tilting direction on the amplitude of the P2 component, a one-way between subjects ANOVA was conducted using the peak amplitude and latency at electrode Pz. A significant effect of tilting direction on P2 amplitude at the p < 0.05 level was found for the two conditions (F(1, 28) = 11.83, p = 0.002). Additionally, the effect of tilting direction on P2 latency was tested using a one-way between subjects ANOVA. In this case, no significant effect at the p < 0.05 level was found for the two conditions (F(1, 28) = 0.25, p = 0.62).

3.2. Binary single-trial classification

We performed binary classification of perturbation against rest trials. The participant-specific chance level for binary classification as well as the chance level for the grand average were calculated using an adjusted Wald interval (α = 0.05) (Müller-Putz et al 2008, Billinger et al 2012). The result was corrected for multiple comparisons (with n = 76 comparisons) using a Bonferroni correction. The participant-specific chance level is 61.05% and the grand average chance level is 52.84%. All classification accuracies reported in this section were achieved on test data.

The calibration phase was used to investigate the time window of maximal discriminability between rest and perturbation conditions. Figure 4 shows the participant-specific onsets of training windows relative to perturbation onset in the box plot. The average position of the different sized training windows is shown above the box plot. For a training window consisting of a single sample, the window onset was 184.7 ± 80.4 ms (mean ± STD) after perturbation onset. The average window onset for a training
window with a length of 200 ms was 82 ± 74.2 ms (mean ± STD) after perturbation onset. When a window length of 400 ms was chosen for training, the window onset was on average 58.7 ± 74.2 ms (mean ± STD) after perturbation onset. An average window onset of 17.3 ± 152.2 ms (mean ± STD) before perturbation onset was calculated for a training window with a size of 600 ms.

Both independent variables (window length and electrode layout) had a statistically significant effect on the peak accuracy at the 0.05 level. The result of the two-way ANOVA for the window length was (F(3, 224) = 1.96, p = 0.121). For the electrode layout the main effect yielded (F(3, 224) = 15.04, p < 0.001). The interaction of both independent variables did not have a statistically significant effect on peak accuracy: (F(9, 224) = 0.06, p = 0.999). We performed Tukey’s honest significant difference criterion to compensate for multiple comparisons (Tukey 1949). The test showed that the smallest (1 sample) and largest (600 ms) window lengths were significantly different at the 0.05 level. For the second independent variable,
Table 2. Results of binary classification for different window sizes. Peak accuracy and latency with respect to perturbation onset are shown as mean and standard deviation of the individual participant accuracies and for the grand average.

| Window size | Peak accuracy (%) | Peak accuracy (%) | Latency (ms, relative to perturbation onset) | Latency (ms, relative to perturbation onset) |
|-------------|-------------------|-------------------|---------------------------------------------|---------------------------------------------|
|             | (participant-specific) | (grand average)   | participant-specific)                       | grand average)                              |
| 1 sample    | 93.4 ± 5.4        | 81.6              | 221.3 ± 108.9                               | 120                                         |
| 200 ms      | 95.1 ± 5.8        | 83.5              | 277.3 ± 76.3                                | 280                                         |
| 400 ms      | 95.3 ± 5.5        | 78.4              | 434.7 ± 157.8                               | 440                                         |
| 600 ms      | 95.7 ± 4.4        | 76.2              | 536.0 ± 165.5                               | 640                                         |

Table 3. Results of binary classification for different layouts using a window size of 200 ms. Peak accuracy and latency with respect to perturbation onset are shown as mean and standard deviation of the individual participant accuracies and for the grand average.

| Layout                | Peak accuracy (%) | Peak accuracy (%) | Latency (ms, relative to perturbation onset) | Latency (ms, relative to perturbation onset) |
|-----------------------|-------------------|-------------------|---------------------------------------------|---------------------------------------------|
|                       | (participant-specific) | (grand average)   | participant-specific)                       | grand average)                              |
| Minimal (1 channel)   | 87.6 ± 8.0        | 71.5              | 240.0 ± 78.6                                | 280                                         |
| Small (5 channels)    | 93.1 ± 5.9        | 81.2              | 282.7 ± 76.3                                | 320                                         |
| Medium (15 channels)  | 94.4 ± 5.6        | 82.7              | 282.7 ± 73.2                                | 280                                         |
| Full (29 channels)    | 95.1 ± 5.8        | 83.5              | 277.3 ± 76.3                                | 280                                         |

the test showed a significant difference at the 0.05 level between the minimal layout and all other layouts. We then took a more detailed look at the classification differences between the tested window lengths. The full electrode layout (29 channels) was used for the comparison. In figure 5(a) the classification accuracy of all four window lengths for each time point of the tROI are plotted on the left side while the right side shows boxplots of the participant-specific peak accuracies for each condition. Above chance level classification was possible for all four window lengths. However, peak accuracy shifted away from perturbation onset with increasing size of the window. Table 2 summarizes the results for each window. The high inter-participant variability of peak latency led to the difference between grand average peak accuracy and the mean of participant-specific peak accuracy.

Furthermore, we investigated how a reduction of EEG channels impacts classification performance. We tested four different layouts: the full layout using all 29 recorded channels, a medium sized layout with 15 channels, a small layout with 5 channels, and a minimal layout with one channel. Based on our previous results, we used a window length of 200 ms for the comparison. Figure 5(b) shows the grand average classification performance for each time point of tROI on the left side. The right side displays a box plot of peak accuracies of all subjects for each layout. All electrode layouts achieved above chance level performance. The results for each electrode layout are summarized in table 3. Differences between participant-specific and grand average peak accuracy occurred due to variations of peak latency between participants.

Based on our previous results, we decided to use the minimal layout with a window length of 200 ms for further analysis. Figure 6 shows classification results for each of the subjects and the grand average on the left side. Shaded areas indicate the standard deviation (STD) of the grand average. The dashed red line marks the grand average chance level. The confusion matrices shown on the bottom of the left side were calculated for the perturbation onset at 0 ms and for the time point of the maximal accuracy peak at 280 ms. The peak accuracy for each of the participants is shown on the right side of figure 6. Participant-specific chance level is indicated by a dashed red line. The grand average classification accuracy exceeded chance level performance between 40 ms and 780 ms after perturbation onset with a peak accuracy of 71.5% correctly classified trials. Since we evaluated the performance of a classifier every 40 ms, we were able to calculate confusion matrices for different time points of the perturbation. In figure 6, we show the confusion matrix calculated at perturbation onset (0 ms) and at the maximal accuracy peak (280 ms). At perturbation onset, the classifier achieved a true positive rate (TPR) of 10.8% and a true negative rate (TNR) of 93.3%. The Cohen-α coefficient for this time point was 0.364. At the accuracy peak 280 ms after perturbation onset, the classifier achieved a TPR of 49.7% and a TNR of 92.5%. Here, the Cohen-α coefficient was 0.593. TPR, TNR, and Cohen-α coefficient were calculated using the confusion matrices of the grand average classification result.

Each participant exceeded participant-specific chance level classification performance. The maximum accuracy was achieved 240.0 ± 78.6 ms (mean ± STD) after perturbation onset. On average, the participant-specific classification accuracy peaked at 87.6 ± 8.0% (mean ± STD). The difference between grand average peak accuracy
Figure 6. Classification performance of binary classification using a window length of 200 ms and the minimal electrode layout. The upper left plot shows the classification performance of individual subjects over the course of the tROI in blue. The black curve displays the grand average result with the standard deviation depicted by grey areas. The dashed red line indicates the grand average chance level. The calculation time points for the subjacent confusion matrices are indicated by colored marker. The peak accuracy achieved by each of the participants is shown in the upper right plot. Participant-specific chance level is indicated by a dashed red line.

and the participant-specific results occurred due to inter-participant variability of peak time which can be seen on the left side of figure 6. For this reason, we calculated TPR, TNR and Cohen-α coefficient for each participant at the participant-specific peak times. Participants achieved a TPR of 81.5 ± 12.2% (mean ± STD) and a TNR of 93.5 ± 6.4% (mean ± STD) on average. The average Cohen-α coefficient was 0.804 ± 0.112 (mean ± STD).

In order to get a better understanding of our classifier’s behaviour, we calculated receiver operating characteristic (ROC) curves for three different time points in our tROI: At perturbation onset (0 ms), at the time point of grand average peak accuracy (280 ms), and by combining time points of participant-specific peak accuracy. The curves are shown in figure 7. Furthermore, we calculated the area under the ROC curve (AUROC) for all three curves. AUROC was 0.54 at perturbation onset while it reached 0.73 at grand average peak accuracy. When combining the best classification performances of all participants, AUROC reached 0.93.

4. Discussion

In this study, we successfully decoded perturbation evoked potentials on a single trial basis in a controlled laboratory environment. We further identified tuning parameters such as the length of the feature window and the number of EEG channels used to provide configurations for out-of-the-lab use. Post hoc offline analysis showed that even when using only one EEG channel and a feature window of 200 ms, participant-specific performances consistently exceeded 70% accuracy, on average peaking at more than 85%. Underlying EEG correlates show significant differences between the perturbation condition and the rest condition which correspond to the time window used for feature extraction of the best performing classifiers.

4.1. Perturbation-evoked potential

We were able to elicit PEPs using the paradigm and experimental setup described in the method section of this work. Our findings were in agreement with typical morphology of PEPs reported in Dietz et al (1984), Dimitrov et al (1996), Duckrow et al (1999), Staines et al (2001), and Mochizuki et al (2008).

We observed an average amplitude of −24.8 µV for the N1 component of PEPs elicited during our experiment. On average, the N1 component peaked with a latency of 148.5 ms relative to perturbation onset. The N1 was spread around frontal, fronto-central, and central areas with a maximal peak at Cz. These findings are in accordance with results published in previous studies (Dietz et al 1985a, 1985b, Mochizuki et al 2008, Solis-Escalante et al 2019).

The late components of PEPs, called P2 and N2, can be mainly found between 200 and 400 ms after perturbation onset (Varghese et al 2017). We were able to identify one of those components (P2) while the other component (N2) was not identifiable. On average, the P2 peaked 328 ms after perturbation onset with an amplitude of 12.1 µV. This
finding agrees with the results of previously conducted studies involving seated perturbations. Mochizuki and colleagues reported similar latencies for identified P2 components in participants that were exposed to postural perturbations while seated (Mochizuki et al 2009). However, they found higher amplitudes (23.66 ± 6.21 µV). This difference could be explained by the difference in task design between our study and Mochizuki et al: while participants in their experiment had a vertical pole that they were holding during perturbations which induced a compensatory arm reaction, no movement reaction was induced in our experiment. A decreasing P2 amplitude due to lack of balance reaction would however disagree with findings by Quant and colleagues. They reported that the P2 amplitude is higher in passive trials, i.e. trials that do not evoke a compensatory balance reaction (Quant et al 2004). This disagreement between their and our results could occur due to different types of perturbation. In our experiment, participants were exposed to whole-body perturbations while the PEPs in Quant’s experiment were elicited using dislocation of the feet.

The P1 component of PEPs was not distinguishable from background EEG for all of the recorded participants. This finding is consistent with P1 responses observed in previously conducted studies involving seated perturbation tasks (Staines et al 2001, Quant et al 2004, Mochizuki et al 2009). The P1 has a small amplitude which can result in difficulties separating this early PEP component from background EEG. Since the perturbation was induced by hand, there were small mechanical differences in the perturbation movement between the trials. Together with the small amplitude, these variations could have led to the absence of a P1 response in PEPs of single participants due to the averaging of trials (figure 2).

The comparison between PEPs elicited during right perturbation trials and left perturbation trials showed that the P2 component had a higher amplitude during left trials. Evaluating the accelerometer data showed a difference in the perturbation speed/acceleration between left and right trials (see figure S3 in the supplementary material). Furthermore, we surmise that the amplitude difference is partly caused due to our referencing with only one electrode at the left mastoid. To confirm this theory, we performed a subsequent reanalysis of the comparison between left and right trials and applied a common average reference (CAR) (Offner 1950, Osselton 1965). The difference in P2 amplitude at Pz between the two executed perturbation directions was not detectable after re-referencing our recorded data using this spatial filtering approach (P2 amplitude at Pz left trials: 11.3 ± 3.7 µV; P2 amplitude at Pz right trials: 11.1 ± 3.2 µV; see figure S2 in the supplementary material). The result of our reanalysis together with the accelerometer data supports the assumption of an artificial difference between left and right trials due to several confounders. This agrees with findings published in literature, where no effect of motor or sensory information (Quant et al 2005) or

Figure 7. ROC curves of the binary classifier at different time points. Curves were calculated at perturbation onset (0 ms), the grand average peak accuracy (280 ms), and the combined time points of participant-specific peak accuracy.
psychological factors (Sibley et al 2010) on late PEP components was found. Since there is a high probability that the difference between PEPs elicited during left trials and PEPs elicited during right trials is not founded in neural activity, we did not attempt to train a classifier for the discrimination of left and right trials.

4.2. Binary single-trial classification

The result of binary classification shows that a PEP can be discriminated from ongoing resting EEG with an accuracy of above 80%. To the best of our knowledge, there are no other studies with the goal of classifying PEPs. We used a calibration approach to detect the time interval in which the discriminability of perturbation and rest trials is maximal. Our results show that this time interval is congruent with the occurrence of the N1 component around 140 ms after the perturbation onset. All of the four tested window lengths are centered around this EEG correlate. Since the PEP has a participant-specific latency, this centering around the EEG correlate explains the window onset variability shown in figure 5. This finding supports our hypothesis that a PEP can be decoded by the electrical properties of this activity pattern.

The reason why we chose the window sizes used in this work is that 1 sample serves as a minimal example while 200 ms is enough to fully envelope the N1 component. 400 ms and 600 ms are used to analyze how classification performance is affected by longer time windows. Our investigation of different window sizes for the classification showed that there are no significant differences between the peak classification performances of the four tested window lengths. This is not surprising for the three longer window sizes (200 ms, 400 ms, and 600 ms) since the classifier mainly uses information provided by the N1 component for the detection of a PEP and all three window lengths are big enough to fully contain this component. It is interesting that a single sample is already enough to achieve a classification performance similar to those achieved with longer windows. This can be explained by the high amplitude of the N1 component, which is big enough to encode sufficient information in a single sample. However, the usage of only a single sample can prove to be problematic since such a restriction in terms of time makes the classifier vulnerable to artifacts. On the other hand, an increase in window size means that the static delay introduced by the window-based classification approach will also increase. The 200 ms window achieved peak accuracies around 100 ms after the amplitude peak of the N1 component in contrast to a delay of around 300 ms for the 400 ms window and a delay of around 400 ms for the 600 ms window. Our results suggest that a window length of 200 ms is a good trade off that allows for a classification that is stable and robust against artifacts and introduces a reasonable delay. We used the full layout to compare different window lengths in order to utilize all informations that were available to us when assessing the difference in classification performance of different windows.

Since our previous results show that there is no significant effect of window length on classification peak accuracy, we decided to use the 200 ms window for the comparison of different layouts because this window length achieves a good tradeoff between static delay and robustness. Four layouts were considered: the full layout of 29 electrodes, a medium layout with 15 electrodes, a small layout consisting of 5 electrodes, and a minimal layout with only 1 electrode. While the full, medium, and small layout performed without significant difference reaching peak classification accuracies around 94% on average, the minimal layout reached slightly lower accuracy peaks with a maximal classification performance of 87.6% on average around 240 ms after the perturbation onset. This finding agrees with the localization of the PEP N1. Since the component is distributed over frontal, central, and parietal areas with the peak amplitude located at Cz, the removal of channels that are far away from Cz does not change the classification accuracy and suggests that no additional discriminative information can be found in these channels. This hypothesis is supported by the high accuracies of well above 80% (peak) achieved with the minimal layout that uses only the Cz electrode.

To analyze our classifier in more detail, we looked at the classification results using a 200 ms window length and the minimal electrode layout. We chose these parameters due to the findings discussed above. We found that all participants were able to reach a peak classification accuracy above 70%. We calculated confusion matrices at different timepoints to judge the behaviour of our classifier. The confusion matrix calculated at perturbation onset (0 ms) indicated that almost all trials were classified as rest trials (see figure 6). This behavior is expected since all trials look like rest trials at perturbation onset (flat EEG). Therefore, classifying all trials as rest trials is the correct behavior at perturbation onset and our classifier showed this behavior. The confusion matrix showed that only half of the PEP trials were correctly classified at grand average peak accuracy. However, there was a high variety of classification peak latency, i.e. for many participants we did not take the best performing time point when calculating the confusion matrix. For this reason, we calculated confusion matrices for each participant at the specific peak time. These matrices showed much better results (TPR: 81.5 ± 12.2%; TNR: 93.5 ± 6.4%; Cohen-α: 0.804 ± 0.112). Furthermore, we calculated AUROC values at different time points to get a better performance measurement of our classifier at different points in our tROI. At perturbation onset, we calculated an AUROC value of 0.54. Since every trial looked like a rest trial at perturbation onset, as previously mentioned, we expected the classifier to fail at this time.
point and the AUROC value supports our expectation. An AUROC value of 0.73 indicates that our classifier only reached fair performance at grand average peak accuracy. Again, this behavior is not unexpected since there was a high inter-participant variability of peak accuracy. When only the best performances are considered, by combining participant-specific time points of peak accuracy, we found that our classifier reached a classification performance with an AUROC value of 0.93.

4.3. Limitations and future work
We prepared resting trials by using a virtual onset between perturbations. Participants were instructed to stay as relaxed as possible and did not perform a mental or physical task during these periods. In a real-world scenario, however, PEPs have to be discriminated from ongoing EEG during mentally or physically demanding tasks like flying an airplane or doing physical therapy for the purpose of rehabilitation.

Another limitation of selecting resting trials in this way is that subjects are constantly prepared for being tilted. Although we randomized perturbation onsets to some degree, this anticipation of the perturbation cannot be completely prevented.

Due to the manual tilting during trials, the acceleration showed a systematic divergence between the two tilting conditions. This is unfortunately a confounder, preventing further investigations between left and right perturbations. However, the ability to distinguish tilts to the left from tilts to the right is of great importance for several fields such as application of exoskeleton. Therefore, future experimental setups should incorporate controlled tilting procedures to validly investigate different tilt directions.

The use of a window-based classification approach will introduce a static delay in an online scenario. Although other online BCI systems do not suffer severely from that delay (Bin et al 2009, da Cruz et al 2015), a BCI that is supposed to detect loss of balance control works under strong time restrictions. One example would be a patient in rehabilitation therapy to restore lost walking functionality. The patient is walking with minimal support but is attached to a system that prevents falling if a loss of balance control is detected. Such a system has to react in an instant to prevent the patient from falling and hurting themselves. Future research should address the delay introduced due to the window-based classification approach and investigate whether the duration of that delay could be problematic in real-world scenarios.

The classifier used for discriminating PEP and rest condition, namely sLDA, has one important restriction: the performance depends on the estimated covariance matrix. However, the stability of this estimation decreases with an increase of the size of the feature space (curse of dimensionality; see Blankertz et al (2011)). The usage of a shrinkage algorithm already addresses this problem but the use of further dimension reduction techniques could improve the classification in terms of performance and stability. Successfully implemented methods are e.g. principal component analysis (PCA) (Lopes-Dias et al 2019), sequential forward selection (SFS) (Jochumsen et al 2016), or smoothing with a moving average filter (Pinegger et al 2015).

We have shown that classification can be performed with high accuracy using only a small number of channels (93.1% with five channels; 87.6% with one channel). This result is crucial for real-world scenarios since size is an important factor in integrating new parts into an existing system. Furthermore, companies that are specialized in EEG hardware are investigating into shrinking the size of their hardware during the last years to improve mobility. Although the developments in the hardware section are promising for PEP-detection systems, a thoroughly online study has to be conducted to really determine the viability of such a system. In such an online study, one has to investigate not only the reliability of PEP detection but also the false positive rate (TPR). The TPR should be well below one per minute since compensating loss of balance control usually restricts the user and such restrictions should only be put in place if they are necessary. Finally, tests have to be performed with actual use cases of a system that compensates loss of balance, e.g. with patients during gait rehabilitation, while using a VR headset, or during actual flights.

5. Conclusion
In this study we showed that perturbation-evoked potentials can be robustly discriminated from ongoing EEG using a linear classification approach. Furthermore, we show that this discrimination can be achieved using only a few electrodes for recording.

Our findings are a first step to make information about the user’s balance control state accessible for computers and machines. This would enable a machine to react to perceived balance perturbations which could improve interactions between humans and machines. An improvement of this interaction is important to enhance rehabilitation medicine or VR experience.

Acknowledgments

The authors would like to thank Dietmar Josef Schäfauer for his help in planning and constructing the tilting chair used to perturb participants. Supported by TU Graz Open Access Publishing Fund. The authors have confirmed that any identifiable participants in this study have given their consent for publication.
Author contributions
J D and G R M P had the idea; J D, A S, and G R M P designed the study; J D performed the experiment; J D performed the analysis supported by A S; J D, A S, and G R M P wrote the paper.

ORCID iDs
Jonas C Ditz https://orcid.org/0000-0002-4561-1900
Andreas Schwarz https://orcid.org/0000-0002-3883-4989
Gernot R Müller-Putz https://orcid.org/0000-0002-0087-3720

References
Ackermann H, Diener H C and Dichgans J 1986 Mechanically evoked cerebral potentials and long-latency muscle responses in the evaluation of afferent and efferent long-loop pathways in humans Neurosci. Lett. 66 233–8
Adkin A L et al 2006 Cortical responses associated with predictable and unpredictable compensatory balance reactions Exp. Brain Res. 172 85–93
Bhagat N A et al 2016 Design and optimization of an EEG-based brain machine interface (BMI) to an upper-limb exoskeleton for stroke survivors Front. Neurosci. 10 122
Billinger M et al 2012 Is it significant? Guidelines for reporting BCI performance Towards Practical Brain–Computer Interfaces (Berlin: Springer) pp 333–54
Bin G, Gao X, Yan Z, Hong B and Gao S 2009 An online multi-channel SSVEP-based brain–computer interface using a canonical correlation analysis method J. Neural Eng. 6 046002
Blankertz B et al 2011 Single-trial analysis and classification of ERP components—a tutorial NeuroImage 56 814–25
Cobb S V G et al 1999 Virtual reality-induced symptoms and effects (VRiSE) Presence 8 169–86
Crea S et al 2018 Feasibility and safety of shared EEG/EOG and vision-guided autonomous whole-arm exoskeleton control to perform activities of daily living Sci. Rep. 8 10823
da Cruz J N et al 2015 Adaptive time-window length based on online performance measurement in SSVEP-based BCIs Neurocomputing 149 93–9
Dietz V, Quintern J and Berger W 1984 Cerebral evoked potentials associated with the compensatory reactions following stance and gait perturbation Neurosci. Lett. 50 181–6
Dietz V, Quintern J and Berger W 1985a Afferent control of human stance and gait: evidence for blocking of group I afferents during gait Exp. Brain Res. 61 153–63
Dietz V, Quintern J, Berger W and Schenck E 1985b Cerebral potentials and leg muscle e.m.g. responses associated with stance perturbation Exp. Brain Res. 57 348–54
Dimitrov B, Gavrilenko T and Gavet P 1996 Mechanically evoked cerebral potentials to sudden ankle dorsiflexion in human subjects during standing Neurosci. Lett. 208 199–202
Duckrow R B et al 1999 Stance perturbation-evoked potentials in old people with poor gait and balance Clin. Neurophysiol. 110 2026–32
Faller J et al 2012 Autocalibration and recurrent adaptation: towards a plug and play online ERD–BCI IEEE Trans. Neural Syst. Rehabil. Eng. 20 313–19
Friedman J H 1989 Regularized discriminant analysis J. Am. Stat. Assoc. 84 165
Gerjets P, Walter C, Rosenstiel W, Bogdan M and Zander T O 2014 Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain–computer interface approach Front. Neurosci. 8 385
Jochumsen M, Niazi I K, Dremstrup K and Kamavuako E N 2016 Detecting and classifying three different hand movement types through electroencephalography recordings for neurorehabilitation Med. Biol. Eng. Comput. 54 1491–501
Kothe C 2020 Lab streaming layer (LSL) (https://github.com/sccn/labstreaminglayer)
Lopes Dias C, Shurlee A I and Müller-Putz G R 2018 Masked and unmasked error-related potentials during continuous control and feedback J. Neural Eng. 15 036031
Lopes-Dias C, Shurlee A I and Müller-Putz G R 2019 Online asynchronous decoding of error-related potentials during the continuous control of a robot Sci. Rep. 9 17596
Marlin A, Mochizuki G, Staines W R and McIlroy W E 2014 Localizing evoked cortical activity associated with balance reactions: does the anterior cingulate play a role? J. Neurophysiol. 111 2634–43
Millán J D R et al 2010 Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges Front. Neurosci. 4 161
Mochizuki G et al 2008 Cortical responses associated with the preparation and reaction to full-body perturbations to upright stability Clin. Neurophysiol. 119 1626–37
Mochizuki G, Sibley K M, Cheung H J, Camilleri J M and McIlroy W E 2009 Generalizability of perturbation-evoked cortical potentials: independence from sensory, motor and overall postural state Neurosci. Lett. 451 40–44
Müller-Putz G R et al 2008 Better than random: a closer look on BCI results Int. J. Bioelectromagn. 10 52–55
Öffner F F 1950 The EEG as potential mapping: the value of the average monopolar reference Electroencephalogr. Clin. Neurophysiol. 2 213–14
Osselton J W 1965 Acquisition of EEG data by bipolar, unipolar and average reference methods: a theoretical comparison Electroencephalogr. Clin. Neurophysiol. 19 527–8
Parra L C, Spence C D, Gerson A D and Sajda 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
Pérez M A O and Benjamin Knapp R 2007 BioTools: a biosignal toolbox for composers and performers International Symposium on Computer Music Modeling and Retrieval (Berlin: Springer) pp 441–52
Pinegger A, Faller J, Halder S, Wriessnegger S C and Müller-Putz G R 2015 Control or non-control state: that is the question! An asynchronous visual P300-based BCI approach J. Neural Eng. 12 014001
Quant S, Adkin A L, Staines W R and McIlroy W E 2004 Cortical activation following a balance disturbance Exp. Brain Res. 155 393–400
Quant S, Maki B E and McIlroy W E 2005 The association between later cortical potentials and later phases of postural reactions evoked by perturbations to upright stance Neurosci. Lett. 381 269–74
Scheffers M K and Coles M G H 2000 Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors J. Exp. Psychol.: Hum. Percept. Perform. 26 141–51
Schultz-Kraft M et al 2016 The point of no return in vetoing self-initiated movements Proc. Natl Acad. Sci. 113 1080–5
Schwarz A et al 2015 A co-adaptive sensory motor rhythms brain–computer interface based on common spatial patterns and random forest 37th Ann. Int. Conf. of the IEEE Eng. Med. Biol. Soc. pp 1049–52
Schwarz A et al 2019 Unimanual and bimanual reach-and-grasp actions can be decoded from human EEG IEEE Trans. Biomed. Eng. 67 1684–95
Schwarz A, Öffner P, Pereira J, Shurlee A I and Müller-Putz G R 2018 Decoding natural reach-and-grasp actions from human EEG J. Neural Eng. 15 016005
Shibasaki H and Hallett M 2006 What is the Bereitschaftspotential? *Clin. Neurophysiol.* 117 2341–56
Sibley K M *et al* 2010 The relationship between physiological arousal and cortical and autonomic responses to postural instability *Exp. Brain Res.* 203 533–40
Solis-Escalante T, van der Cruyssen J, de Kam D, van Kordelaar J, Weerdesteyn V and Schouten A C 2019 Cortical dynamics during preparation and execution of reactive balance responses with distinct postural demands *NeuroImage* 188 557–71
Staines R, McIlroy W and Brooke J 2001 Cortical representation of whole-body movement is modulated by proprioceptive discharge in humans *Exp. Brain Res.* 138 235–42
Tukey J W 1949 Comparing individual means in the analysis of variance *Biometrics* 5 99
Varghese J P, McIlroy R E and Barnett-Cowan M 2017 Perturbation-evoked potentials: significance and application in balance control research *Neurosci. Biobehav. Rev.* 83 267–80
Veneman J F, Kruidhof R, Hekman E E G, Ekkelenkamp R, Van Asseldonk E H F and van der Kooij H 2007 Design and evaluation of the LOPES exoskeleton robot for interactive gait rehabilitation *IEEE Trans. Neural Syst. Rehabil. Eng.* 15 379–86
Wolpaw J and Wolpaw E W 2012 *Brain–Computer Interfaces: Principles and Practice* (Oxford: Oxford University Press)
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
Zander T O *et al* 2009 Utilizing secondary input from passive brain–computer interfaces for enhancing human–machine interaction *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (Berlin: Springer) pp 759–71
Zander T O and Kothe C 2011 Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general *J. Neural Eng.* 8 025005