Calibration-Free Error-Related Potential Decoding with Adaptive Subject-Independent Models: A Comparative Study

Article *in* IEEE Transactions on Medical Robotics and Bionics · July 2020

DOI: 10.1109/TMRB.2020.3012436

CITATIONS
0

READS
32

4 authors:

Florian Marian Schönleitner
Technische Universität München
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Lukas Otter
Technische Universität München
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Stefan K. Ehrlich
Technische Universität München
14 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Gordon Cheng
Technische Universität München
337 PUBLICATIONS **6,433** CITATIONS
SEE PROFILE

Calibration-Free Error-Related Potential Decoding with Adaptive Subject-Independent Models: A Comparative Study

Florian M. Schönleitner, Lukas Otter, Stefan K. Ehrlich and Gordon Cheng *Fellow, IEEE*

***Abstract*—Error-related potentials (ErrPs) provide an elegant method to improve human-machine interaction by detecting incorrect system behavior from the electroencephalogram of a human operator in real time. In this paper, we focus on adaptive subject-independent decoding models particularly suitable for ErrP classification. As individualized decoding models require a time-consuming calibration phase, such models provide a promising alternative. Based on an investigation of the characteristics of inter-subject variations in the signal and feature space, we evaluate the performance of a decoding model solely trained on prior data and the effectiveness of adapting this model to a new subject in a comparative study. Our results show that such a generalized model can decode ErrPs with an acceptable average accuracy of $72.7 \pm 9.66\%$ and that supervised adaptation can significantly improve the accuracy of the generalized model after adaptation with 85 trials by on average $+3.8 \pm 5.1\%$. We show that adaptation of subject-independent decoding models is superior to the traditional calibration procedure. Unsupervised adaptation only proved effective for some subjects and requires further attention to be practical for a broader range of subjects. Consequently, our work contributes to the development of calibration-free ErrP decoding in the broader scope of enhancing usability of ErrPs for human-machine interaction.**

***Index Terms*—Brain-Computer Interfaces, EEG, Human-machine interaction, Error-Related Potentials, Adaptive Classification**

I. Introduction

THE ongoing development of technological and information systems fundamentally changed the way we perceive and interact with our surroundings. We do not only interact with other humans, but more frequently also with intelligent artificial systems, which can be considered as human-computer (HCI) or human-machine interaction (HMI) [1], [2]. In such scenarios, the human operator has active control over the system and hence expects it to react with a distinct response to his command. If it does not, the system is considered flawed, which can have fatal consequences in the worst case. Consequently, one fundamental aspect of all intelligent technological systems is the correct interpretation and execution of the operator's intention. Hence, a much desired ability is the evaluation of system behavior in an instantaneous and seamless manner. As error perception and evaluation are essential abilities of all humans, characteristic changes in the neurophysiological activity are measurable in the prefrontal cortex directly after observing wrong responses of an interaction partner [3], [4]. Such error-related potentials (ErrPs) provide an elegant way to detect incorrect system behavior from the electroencephalogram (EEG) of a human operator. Recently, several studies have shown that such passive brain-computer interfaces (BCIs) can monitor a subject's neurophysiological behavior to decode ErrPs in real-time. Thus, this neuronal evaluation signal can be used to correct wrong system behavior [5]–[10], for label-free adaptation [11] as well as teaching robotic skills [12]–[14] and co-adapting robot behavior during human-robot interaction [15], [16]. Nonetheless, the usability of ErrPs is impeded by the non-stationarity of the EEG signals [17]–[19], which leads to variations in the signal of different subjects. As this also reflects in different feature distributions [20]–[22], simply transferring a pre-trained decoding model to classify data from a new person or data obtained from a different experimental context [8] is accompanied by grave performance losses. In addition, generating subject-specific decoding models involves a time consuming initialization phase as substantial training data is required for model calibration. However, for practical subject-independent decoders both approaches are impractical as the only information available would be labeled data from prior subjects and a few unlabeled trials from a new subject. Fortunately, information transfer and adaptive classification methods provide alternative techniques: by combining both methods it is possible to generate decoders which adapt to the subject-specific characteristics without the need for a time consuming initialization phase [11], [23]–[27].

Iturrate et al., 2011, replaced the calibration phase in ErrP classification by inter-subject information transfer [26]. They started classification in an early stage with a decoding model trained on pooled data of multiple subjects. Afterwards, they adapted the model towards a different subject by incorporating new labeled trials. This approach also proved useful for motor imagery classification. Lotte and Guan, 2009 [28] and Vidaurre et al., 2011 [25], applied a similar approach which also allows unsupervised adaptation as no labeled information from new subjects is needed. However, their approach is not easily transferable to other paradigms as it makes some assumption about the underlying nature of inter-subject variations, which

Florian M. Schönleitner is with the Chair for Cognitive Systems, Department of Electrical and Computer Engineering, Technische Universität München, Munich, Germany, e-mail: florian.schoenleitner@tum.de

Lukas Otter is with the Department of Electrical and Computer Engineering, Technische Universität München, Munich, Germany, e-mail: lukas.otter@tum.de

Stefan K. Ehrlich is with the Chair for Cognitive Systems, Department of Electrical and Computer Engineering, Technische Universität München, Germany, e-mail: stefan.ehrlich@tum.de

Gordon Cheng is the head of the Chair for Cognitive Systems, Department of Electrical and Computer Engineering, Technische Universität München, Germany, e-mail: gordon@tum.de

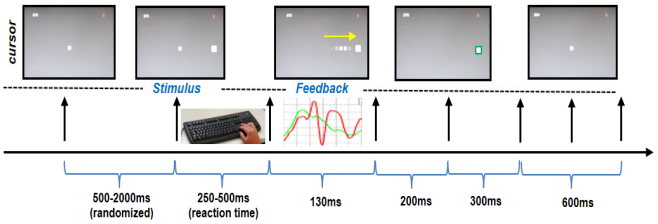

Fig. 1. Trial design of the experimental protocol. Adapted from [7] with permission

are not given by ErrPs.

Building upon our preceding study [29], we supplement the related work by analyzing the capabilities and limitations of subject-independent adaptive decoding models similar to the ones used in [25], [26], [28], [30] in the context of ErrP classification. Thereby, we start by analyzing the inter-subject inconsistencies of ErrP characteristics in the signal and feature space. Based on this analysis we assess the feasibility of reusing data from other subjects, and then evaluate the effectiveness of a suitable model adaptation algorithm. In result, we demonstrate that model initialization from prior data combined with supervised adaptation is superior to traditional BCI calibration. First, reusing prior data for model initialization is more efficient as it allows decoding of ErrPs from the first trial on which is not possible with a traditional calibration procedure. Second, supervised model adaptation proved more effective than a minimal sample calibration and led to a consistently higher decoding performance. Although unsupervised adaptation did not lead to a performance increase for most subjects, it is promising especially in the case of a high accuracy of the initial model.

II. Material and Methods

A. Experimental data

We based our study on an in-house recorded dataset[1] which contains EEG recordings of twelve subjects[2]. This allowed us to simulate a BCI used by different subjects with a sufficient amount of prior data. A detailed explanation of the experiment is available in [8]. In short, the experimental task consisted of a computer-based interaction scenario in which participants were asked to send directional commands depending on a visual stimulus (see Figure 1) and subsequently perceived the movement of a cursor as feedback. Participants indicated their direction of choice by pressing left or right arrow-keys on a keyboard with one finger of the right or left hand. The elicitation of error-related potentials in the subject's EEG was realized by wrong cursor responses which were introduced in a pseudo-randomized fashion. Previous analysis of this dataset showed that within-subject single-trial classification of ErrPs is possible with accuracies of around 90% average across subjects [8]. EEG signals were recorded with a Brain Products actiCHamp amplifier with 32 active electrodes, arranged according to the international 10-20 system. Three channels were used for post-hoc ocular artifact correction and all leads were referenced with respect to the average of the left and right mastoids (TP9 and TP10). With this setup, 500 trials per subject were recorded of which approximately 175 (35%) are wrong cursor responses. Data preprocessing consisted of band-pass filtering, artifact correction, re-referencing, and downsampling in accordance to the procedure described in [7]. Furthermore, the EEG data of each trial was segmented in a window of -200ms to 1000ms time locked to the onset of the cursor movement. After preprocessing, temporal features were extracted similar to the procedure described in [7] by calculating the arithmetic mean of the signal in ten partially overlapping time windows from 0ms to 800ms after feedback presentation. In total, 270 temporal features (27 channels x 10 windows) were extracted for every trial of every subject in the following windows: 0-100ms, 100-200ms, 150-250ms, 200-300ms, 250-350ms, 300-400ms, 400-500ms, 450-550ms, 500-600ms, 600-800ms. Note that we altered the original feature extraction in [7] by adding an additional window from 450ms to 550ms to better capture the late negativity around 400ms (see Figure 2).

[1]Dataset with detailed description available at https://github.com/stefan-ehrlich/dataset-ErrP-HRI

[2]the data of one subject was removed from the dataset due to technical problems during data acquisition

B. Analysis of error-related potentials characteristics

To find a suitable adaptation algorithm, a systematic analysis of the characteristics of variations between different subjects' ErrPs was performed. We evaluated and compared individual ErrPs in the signal space and the class-specific feature distributions after feature extraction and set the focus on gaining insight on inter-subject signal- and feature-shifts which must be considered by the adaptation algorithm. For the signal analysis, we computed the average of all error and non-error trials for each subject as well as the grand average of the signal by computing the mean of all trials and subjects. The results are visualized as time-series and topographic plots in Figure 2. Inter-subject inconsistencies between features were analyzed on a) the bivariate feature subspace spanned by the two most discriminative features and b) the distance between the means of the class-specific multivariate probability distributions of each subject.

a) Fisher Score Projections: For feature analysis we performed a dimensionality reduction based on the Fisher score to focus our analysis on the features which are most relevant for classification. This measure of linear separability of distributions is based on the within class scatter $\mathbf{S}_w$ and the between class scatter $\mathbf{S}_b$ and finds an optimal transforming axis $\mathbf{w}$ which maximizes Fisher score $J(\mathbf{w})$ of a feature.

$$\mathbf{S}_b = \sum_{i=1}^{K} p(c_i)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^T \tag{1}$$

$$\mathbf{S}_w = \sum_{i=1}^{K} p(c_i) E[(\mathbf{y} - \boldsymbol{\mu}_i)(\mathbf{y} - \boldsymbol{\mu}_i)^T | c_i] \tag{2}$$

$$J(\mathbf{w}) = \frac{|\mathbf{w}^T \mathbf{S}_b \mathbf{w}|}{|\mathbf{w}^T \mathbf{S}_w \mathbf{w}|} \tag{3}$$

Here, K is the number of classes, which reduces to two for binary classification, $p(c_i)$ denotes the prior of the class c_i,

$\boldsymbol{\mu}_i$ and $\bar{\boldsymbol{\mu}}$ are the class-specific and global mean respectively. We particularly chose the Fisher score as a metric as LDA classifiers, which were used for single-trial classification, are based on the same basic principle. To find the most relevant features, the Fisher score as defined in (3) was calculated and standardized for each feature and every subject individually and the two features with the highest average scores were chosen as the most discriminative features across all subjects.

b) Class-specific distance between the means: Besides the two-dimensional Fisher projection we also analyzed feature shifts between subjects in a quantitative manner taking into consideration all feature dimensions. To do so, we calculated the summed absolute distance between the means for each subject combination $S_k||S_m$ where $k \neq m$.

$$H_i(S_k||S_m) = \sum_{f=1}^{270} \mid \mu_{i,f}^{S_k} - \mu_{i,f}^{S_m} \mid \tag{4}$$

f denotes the feature (which in total sum up to 270), i denotes the class (error and non-error), S_k and S_m denote the subjects and $\mu_{i,f}$ is the class-specific mean of the subject. As H_i is symmetric, and thus $H_i(S_k||S_m) = H_i(S_m||S_k)$, repeated combinations were excluded from the analysis. We particularly focused on changes in the mean because we expect them to be the main source of inter-subject variations while the covariance appears to be rather stable across subjects (see also Figure 3). Furthermore, adaptation of only the mean was reported to sufficiently capture inter-subject variations in a MI task [25] which is in line with a preliminary study from our side.

C. Single trial classification

Due to its simplicity and effectiveness, linear discriminant analysis (LDA) is a widely used classification method for classifying EEG signals [31]. ErrP decoding is a binary classification problem. Hence, all trials can be split into the two classes $C1$ (error), $C2$ (non-error) and the class-wise Gaussian distributions with the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ as class-wise means and $\boldsymbol{\Sigma}$ as the shared covariance, which is assumed to be equal for both classes. For covariance estimation, a shrinkage approach with optimal shrinkage intensity was used [32]. Single trials were classified by a LDA classifier which computes a decision boundary as follows:

$$D(\mathbf{x}) = [b, \mathbf{w}^T] \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \tag{5}$$

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1}) \tag{6}$$

$$b = -\mathbf{w}^T \bar{\boldsymbol{\mu}} \tag{7}$$

$$\bar{\boldsymbol{\mu}} = \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) \tag{8}$$

Note that this description of an LDA model assumes equal prior probabilities of each class during training which was ensured by repeated random undersampling of the majority class (see also Section II-E). New trials are classified by computing the distance $D(\mathbf{x})$ of the feature vector to the separating hyperplane. Hence, a class decision is made based on the sign of $D(\mathbf{x})$:

$$\mathbf{x} \in \begin{cases} C_1 & \text{if } D(\mathbf{x}) < 0 \\ C_2 & \text{if } D(\mathbf{x}) > 0 \end{cases} \tag{9}$$

D. Model adaptation

For every new trial $\mathbf{x}_t$, classifier adaptation was implemented by adjusting the parameter $\boldsymbol{\mu}_i$ of the class-specific probability distributions by Equation (10). This formula was inspired by other works such as [25], [28], [30] where a similar adaptation approach was used.

$$\boldsymbol{\mu}_i(t+1) = (1-\lambda)\boldsymbol{\mu}_i(t) + \frac{\lambda}{N(t)} \sum_{t=1}^{N(t)} \boldsymbol{x}_t \tag{10}$$

λ is the update parameter which determines the rate of adaptation and $N(t)$ is the total number of available trials from the current subject.

This adaptation approach is particularly suitable for ErrP decoding mainly because of three reasons: first, class-wise mean adaptation was chosen instead of pooled mean adaptation as the previous feature analysis revealed the class dependency of variations in the individual subject's data (see Section III-A). Second, adapting only the mean values but not the covariance seems sufficient as additional covariance adaptation did not lead to a noteworthy performance increase in our experiments. This was also observed by other groups where adapting only the mean could sufficiently capture inter-subject variations in a MI experiment [25]. Further, when performing covariance adaptation $\mathbf{n}(\mathbf{n}+\mathbf{1})/\mathbf{2}$ covariance matrix entries need to be estimated from a n-sized input vector, which is - especially in the case of limited subject data - highly unreliable. By restricting to mean adaptation, this ratio reduces to a linear $\mathbf{n}$ to $\mathbf{n}$ relation. Third, when comparing single trial adaptation to trial average adaptation, the latter approach has the benefit of scaling the relevance of new trials with the usage time of the system. The fact that this approach neglects within-session adaptation is less relevant as ErrPs appear to be stable over time [33], [34].

E. Evaluation scenarios

For a systematic evaluation of the proposed adaptive LDA model, we will compare five scenarios. These scenarios use a different amount of information from new and prior subjects.

1) 10-fold cross validation (CV): To test the performance of the proposed LDA classifier, 10-fold cross validation was performed separately for every subject. Cross-validation is a commonly used evaluation scenario in which the data is repeatedly divided into a larger training set and a smaller evaluation set which comprise different portions of the data for each repetition. We performed ten repetitions (hence *10-fold* cross validation) in each of which all trials of one subject were divided randomly in a 10%-90% proportion for the testing and training set respectively. As our LDA description requires a balanced training set, random undersampling of the majority class was performed ten times for each fold and the results of

this ten individual undersampling-repetitions were averaged to compensate effects of differently balanced training sets. This ensures that each trial in the training set has a high probability of being used at least once for model training. Finally, the results of each fold were averaged to obtain the final evaluation metric for the CV scenario. Note that the LDA models in this scenario were tested on the same subject as they were trained on, though on different trials. Hence, this scenario investigates subject-specific decoding models which accuracies can be interpreted as the optimal decoding performance of the chosen classification model.

2) Minimal sample calibration (MSC): This scenario implements the traditional training of a subject-specific classifier with minimal samples. The calibration set consists of balanced calibration trials from the current subject only and increases over time as more trials get available. Data balancing was performed just like in the CV scenario. As the label information is needed for (re-)training and at least some labeled data is necessary for classifier training, this scenario is only possible in a supervised manner and cannot be used to classify the first trials of the subject. In our scenario, we used at least 50 trials from the current subject for training and then retrained the model by adding always one consecutive trial to the training set. This was repeated until all but the last 100 trials, which were used for testing, were incorporated in the training set. The main difference to the CV scenario is the reduced number of trials used for model training which would have a positive effect on the calibration time.

3) Generalized model (GM): A generalized model [35]–[38] can be built without any information of the current subject by training a LDA on a balanced training set which contains all prior data from other subjects. To compensate effects induced by random undersampling, 100 LDA models were trained on separately balanced training sets and their parameters were averaged for the final generalized model. We evaluated the model's performance as described in Section II-F. When applied to new subjects, instantaneous ErrP decoding[3] is feasible, though with presumably lower classification accuracy as no individual feature characteristics are captured in the generalized model.

4) Supervised adaptation (SA): To maximize the decoding performance of the generalized model, adaptation to a new subject was implemented according to Equation (10). If class labels for new trials are available, supervised adaptation can be applied by using the label information to determine which class should be adapted.

5) Unsupervised adaptation (UA): If labeled trials are not available, only unsupervised adaptation is feasible. A label for each new trial can be predicted by the adapted generalized model $LDA_{GM}(t)$ (adapted with all available subject trials at time t). Based on this label, the class parameters are adapted according to Equation (10). Provided that the adaptation rate and the generalized model can be initialized reasonably, this scenario would be suitable for instantaneous adaptive online ErrP decoding without the need of any label information from a new subject.

[3]*instantaneous ErrP decoding* in this context should be understood in the sense that a meaningful decoding model is available without calibration from the beginning on.

TABLE I
OPTIMAL AND INITIALIZED ADAPTATION RATES $\lambda_{optimal}$ AND λ_{init} FOR EACH SUBJECT AND ADAPTATION SCENARIO

S_i	$\lambda^{SA}_{optimal}$	λ^{SA}_{init}	$\lambda^{UA}_{optimal}$	λ^{UA}_{init}
1	0.018	0.014	0.001	0.007
2	0.019	0.014	0.009	0.002
3	0.002	0.018	0.002	0.007
4	0.041	0.014	0.002	0.007
5	0.011	0.018	0.012	0.002
6	0.014	0.018	0.007	0.002
7	0.042	0.014	0.044	0.002
8	0.009	0.018	0.013	0.002
9	0.018	0.014	0.012	0.002
10	0.001	0.018	0.001	0.007
11	0.008	0.018	0.001	0.007
12	0.042	0.014	0.001	0.007

F. Evaluation and reporting of results

The testing set for the MSC, the GM, the SA and the UA scenarios consisted of the last 100 trials of the current subject. The testing set for the CV scenario consisted of randomly chosen 10% of the subjects' trials which were not part of the training set. This random selection was repeated 10 times and the obtained classification accuracies of each individual evaluation step were averaged. As only the training but not the testing set is balanced, the performance of the different scenarios will be evaluated by comparing the balanced accuracies $bACC$ of the LDA classifiers, which is the average of the true positive and true negative rate. The optimal adaptation parameter λ is not known a priori, so we performed a grid-search by varying the adaptation rate between 0.001 to 0.05 in 0.001 steps. The rate which led to the highest accumulated $bACC$ increase for each subject and scenario was calculated post-hoc and is termed $\lambda_{optimal}$. The median of all individual optimal adaptation rates, excluding the subject used for evaluation, was used to initialize the adaptation process and is termed λ_{init}. Table I lists the optimal and initialized supervised and unsupervised adaptation rates $\lambda_{optimal}$ and λ_{init} for each subject. For model visualization, we implemented a two-dimensional projection of the class-specific distributions representing the distance of each trial to the decision hyperplanes of the GM and CV models. This projection method was inspired by [39] and slightly adapted to our purposes. Specifically, the x-axis and y-axis of Figure 3 are one-dimensional projections of the respective LDA decision hyperplanes and the x- and y-coordinates represent the distance $D(x)$ as in Equation (5) to those hyperplanes. Thus, this representation allows the interpretation of the models' performances and, more importantly, their differences in an informative graphical representation.

III. RESULTS

A. Signal and feature analysis

Figure 2 shows the grand average of all subjects and trials of class *error* (black) and the difference wave of classes *error*

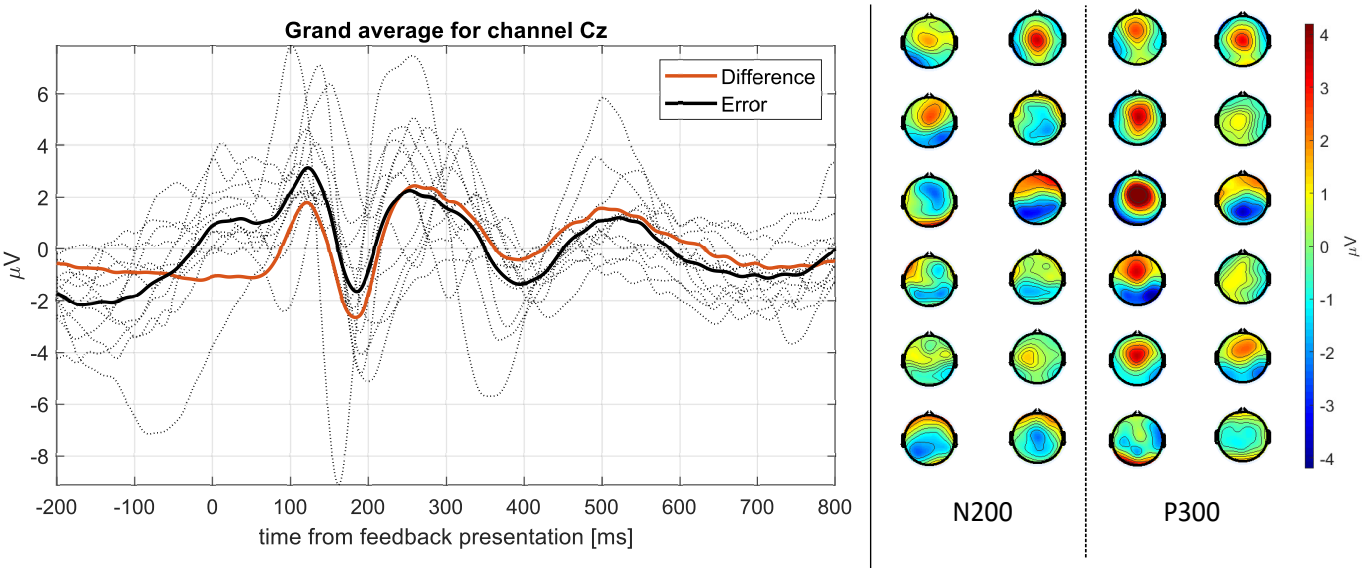

Fig. 2. Left: Grand average of EEG signal for channel Cz for *error* trials (black) and the difference between *error* and *non-error* trials (red). The dashed lines show the EEG signal of every individual subject averaged over all error-trials. Right: topographic plots of the averaged EEG signal between a time window of 187 ± 50 ms (N200) and 277 ± 50 ms (P300) for each subject (subject order from top to bottom).

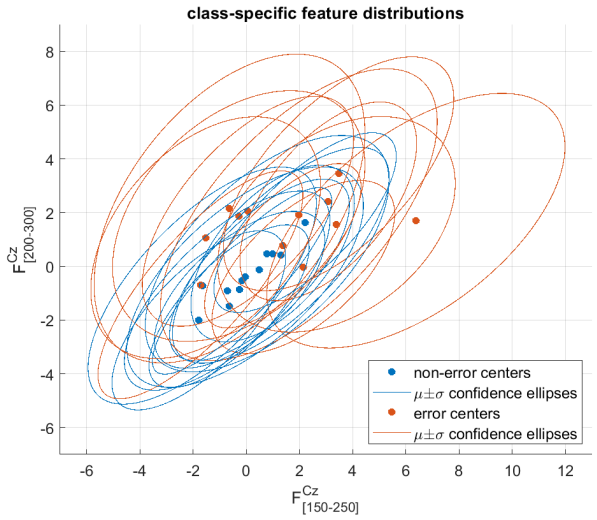

Fig. 3. Bivariate class-specific feature distributions for the two most discriminative features ($F^{Cz}_{150-250}$ and $F^{Cz}_{200-300}$) according to the normalized averaged Fisher scores for each subject. The confidence ellipses of the first standard deviation and their centers are plotted separately for error and non-error trials.

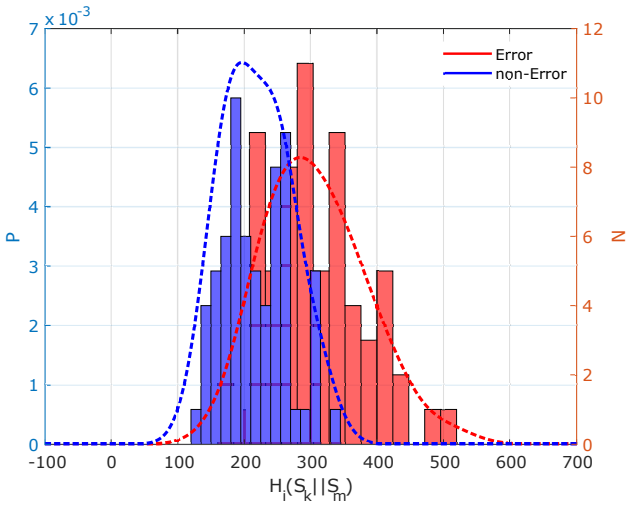

Fig. 4. Summed absolute distance $H_i(S_k||S_m)$ between the class-specific mean values for each subject combination $m \neq k$, separately for each class i. The number of bins for each class is 15 and the solid lines show a Gaussian kernel probability distribution fit. As H_i is symmetric (see Equation (4)), repeated combinations were removed.

minus *non-error* (red) for channel Cz (left) and topographic plots for each subject averaged between 185 ± 50 ms (N200) and 260 ± 50 ms (P300) after feedback presentation (right). The channel Cz was chosen since neuropyhsiological activity of error processing manifests most prominently in central regions [40] observable between 100 - 800 ms after feedback presentation. The two most significant deflections around 180 ms and 250 ms are commonly termed N200 and P300 components respectively [40], [41]. The two most significant deflections around 180 ms and 250 ms are commonly termed N200 and P300 components [40], [41]. Although both components are consistently observable across all subjects in our dataset, their amplitude and timing varies by on average $-3.06 \pm 3.07 \mu V$ (N200) and $+4.53 \pm 1.59 \mu V$ (P300) and by on average 187 ± 15 ms (N200) and 277 ± 37 ms (P300) between subjects[4]. This suggests that the N200 potential is temporally more stable than the P300 potential which is also in line with observations by other research groups [36], [42].

Comparing the bivariate distributions of error and non-error trials (Figure 3) reveals a clear class-dependency of inter-subject variations. On average, non-error trials seem to be affected primarily by linear shifts in the mean which show

[4]This comparison is based on the single value of the highest/lowest amplitude and the corresponding timepoint

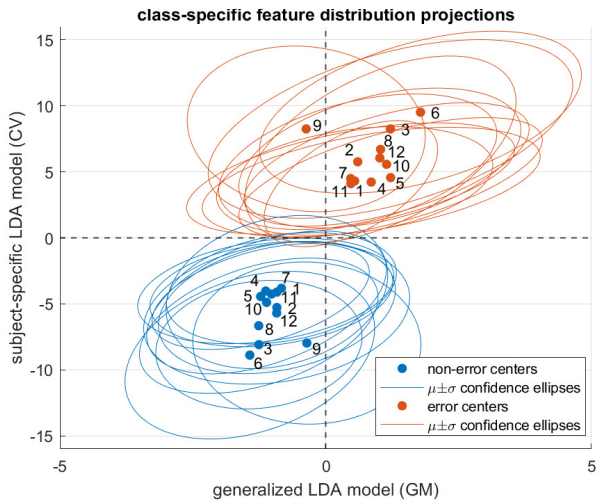

Fig. 5. Two-dimensional projection of the class-specific distributions representing the distance to the decision hyperplanes of the GM and CV models. The x- and y-axis are one dimensional projections of the LDA decision hyperplanes of the two models. The distance $D(x)$ of each data sample to those hyperplanes is shown on the x- and y-coordinates (see equation 5 for the definition of $D(x)$.

a strong positive correlation ($r = 0.92$) between both feature dimensions. Contrary, error trials exhibit an almost random behavior of variations and show no clear correlation. The covariance seems to be less affected by inter-subject variations and appears to be rather stable. To further investigate the assumption of class-dependent feature variations between subjects in a more quantitative manner, we computed the distance H_i between different subject as described in Equation (4). Figure 4 shows the histogram and Gaussian kernel probability distribution fits for the obtained distances for all non-repeating and non-equal subject combinations for both classes independently. In line with the two-dimensional projections in Figure 3, error trials exhibit a higher inter-subject distance and also a higher variance of the distance. Specifically, $H_{Error} = 306.7 \pm 73.8$ and $H_{non-Error} = 218.0 \pm 49.4$. Finally, the higher variability of the error class, which was also observed by other researchers [26], strengthens the assumption of a class-dependency of inter-subject variations. As a consequence for the adaptation algorithm, adaptation of class-specific parameters is indispensable as adaptation of class-unspecific parameters cannot capture class-dependent variations. This is also in line with in-house preliminary analysis where adaptation of only class-unspecific parameters could not improve the accuracies of the GM models.

B. Comparing subject dependent and subject independent models

Besides the decodability of ErrPs in general, a prerequisite for calibration-free systems is the feasibility of model initialization. Here, it is practical to rely on data which is available prior to system usage as we did in the GM scenario. Comparing the averaged balanced accuracies of this scenario to the CV scenario (see Table II), we can report three main results. First, in the case of optimal individualized decoding models, reliable ErrP classification is possible with an across subject average bACC of $88.29 \pm 3.93\%$ (CV scenario). Second, reusing data of prior subjects is practicable to generate a generalized model which resulted in acceptable accuracies of $72.7 \pm 9.66\%$ (GM scenario). Third, as expected, we observed a significant (two-sided Wilcoxon signed rank test, $p < 0.05$) performance decrease between the accuracies of the individualized (CV scenario) and generalized (GM scenario) models of on average $-(15.56 \pm 8.65)\%$. Hence, model initialization without any data from the current subject is feasible but not optimal.

Figure 5 provides further insights into the differences between the GM and CV models of each subject by visualizing the models' decision boundaries in a two-dimensional projection. It can be clearly seen that the subject-specific CV models obtain a much higher class-separability than the subject-independent GM models. This is in particular the case for subject S_9 where especially error trials are largely missclassified by the GM model, leading to a low $bACC$ of this model of only 57.90% (see also Figure 9 and Table II).

C. Effectiveness of model adaptation

To further increase the performance of the generalized model, we analyzed the effectiveness of model adaptation as described in Section II-D. Therefore, we defined five segments of 70 trials each in which we compared the difference between the static GM scenario and the adaptive SA and UA scenarios (Figure 6, right). The maximum number of available trials for adaptation is restricted to 350 because the last 100 trials comprised the testing set and hence could not be used for adaptation. Hence, Figure 6 indicates the performance increase due to adaptation with regard to the amount of data acquired from a new subject. Looking at the adaptation performance, the beneficial effect of supervised adaptation is clearly visible: the balanced accuracies of the SA scenario (green) are significantly (two-sided Wilcoxon signed rank test, $p < 0.05$, $n = 12$) greater than the ones of the GM (cyan) after 83 trials and also slightly greater than the MSC scenario (blue). During the first 50 trials, adaptation could not improve the GM scenario but accuracies were approaching to on average $83.04 \pm 6.92\%$ in the later stage for supervised adaptation (Figure 6, left). Thus, a high classification accuracy only slightly below the CV scenario is established after a certain amount of trials. Unsupervised adaptation as in the UA scenario (red curves in plot) behaves differently and leads to an unlearning effect: the accuracies drop noteworthy below the ones of the static GM scenario within the first 50 trials and converge to $67.07 \pm 13.47\%$, which is on average $-(5.65 \pm 7.79)\%$ below the GM scenario accuracy.

Besides the pooled balanced accuracy, Figure 9 shows the $bACC$ for the different scenarios for every individual subject. In line with results depicted in Figure 6, the $bACCs$ of the supervised adaptation (green) are above the ones of the GM model for all subjects, except S_{10}. For this specific subject, adaptation is not practical as the GM model already outperforms the CV model. On the contrary, the performance

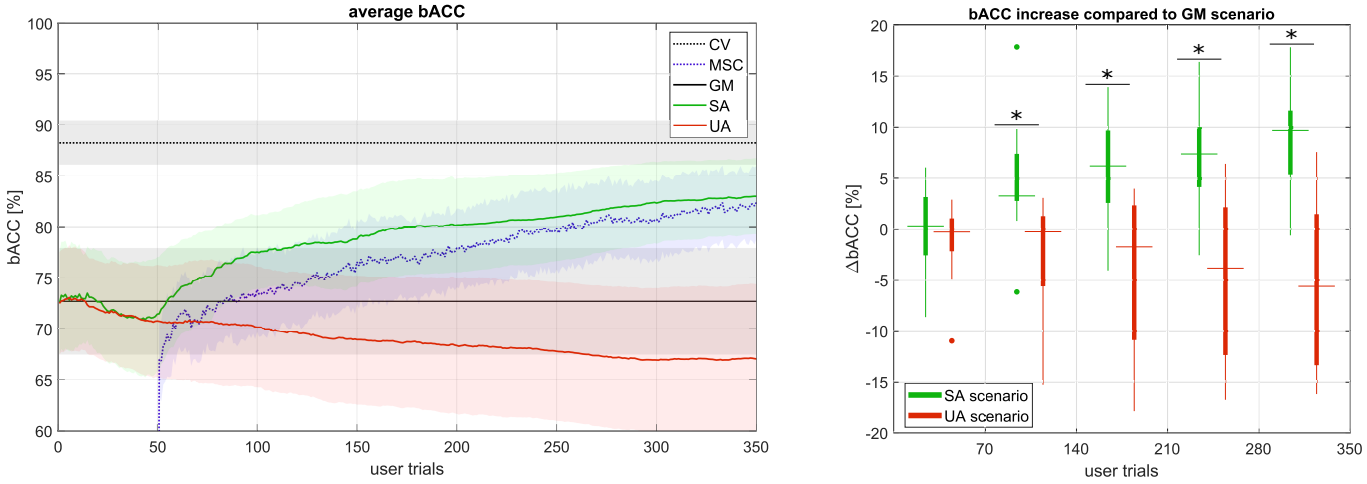

Fig. 6. Left: Averaged balanced classification accuracies for the CV (upper dashed black line), the MSC (dashed blue line), the GM (lower dashed black line), the SA (green solid line) and the UA (red solid line) scenario in a 95% confidence interval plot. Right: difference in model $bACC$ due to adaptation in blocks of 70 trials; $\Delta bACC_{SA} = bACC_{SA} - bACC_{GM}$ (green) and $\Delta bACC_{UA} = bACC_{UA} - bACC_{GM}$ (red). On each box, the median (central mark), the 25^{th} and 75^{th} percentiles (bottom and top edges) and the most extreme data points not considered as outliers are indicated. All outliers (n=3) are represented by dots. Boxes significantly different from zero are marked with a black star on top. Note: Figures are better readable when printed in color.

TABLE II
BALANCED ACCURACIES FOR GM AND CV MODELS

S_i	1	2	3	4	5	6	7	8	9	10	11	12
CV [%]	83.60	87.89	94.24	84.54	84.16	93.90	85.43	93.61	91.67	87.90	84.15	88.35
GM [%]	64.38	68.84	90.24	68.75	72.66	77.18	70.57	78.18	57.90	90.99	62.81	70.20
$\Delta bACC$ [%]	19.21	19.05	4.01	15.79	11.50	16.72	14.86	15.43	33.77	-3.09	21.35	18.15

of the UA scenario is not as homogeneous as the SA scenario: while unsupervised adaptation indeed improves the accuracy of the GM model for some subjects (S_3, S_4, S_5, S_8), it is detrimental for others ($S_1, S_7, S_{10}, S_{11}, S_{12}$). As the only difference between both scenarios is the reliability of the label of new subject trials - which is 100% in the SA scenario and the current model accuracy in the UA scenario - we analyzed the relation between the model accuracy and the performance increase when adapting the model. Figure 7 illustrates the performance difference of the supervised SA and unsupervised UA scenario. Depending on the current model accuracy $bACC_{model}(t)$ in the UA scenario, we computed the difference in the adaptation performance $\Delta bACC_{adapt}(t)$ to the SA scenario. One can observe that the difference between SA and UA scenario as well as the unlearning effect is especially prominent if the initial model accuracy is low (red region). On the contrary, unsupervised adaptation increased the GM performance for three out of four[5] subjects whose initial GM accuracy is above the mean GM accuracy of 72.73% (green region) and supervised adaptation is applicable.

To better understand the relationship between the GM model accuracy, the update parameter λ and the unsupervised adaptation performance, we performed a grid analysis in which we varied λ between 0.005-0.05 in steps of 0.005 and simulated a GM model misclassification rate (MCR) between 0-45% in steps of 5%. For the MCR simulation we randomly switched the labels of n adaptation trials and performed (semi-) unsupervised model adaptation as in the *UA* scenario. The overall balanced accuracy increase to the GM model was defined by the area under the curve (AUC) of the UA model accuracy minus the GM model accuracy. Adaptation was repeated 10 times and results were averaged to minimize a dependency on the random label switching. It turns out that the adaptation performance in the grid analysis is mainly dependent on the model accuracy (see Figure 8) which strengthens our assumption of a minimal GM model accuracy required for successful unsupervised adaptation. Concretely, the grid search analysis suggests a minimal model accuracy of approximately 75% for which model adaptation could increase the overall classification accuracy (see red intersection in Figure 8). Interestingly, this value almost matches the average GM model accuracy of 72.73% (see Figure 7). Furthermore, the update parameter λ is negatively correlated with the model accuracy: the lower the accuracy, the higher the influence of λ. Unsupervised adaptation with a more accurate model is less dependent on the adaptation parameter, at least in the tested region.

[5]for subject S_{10} adaptation with any of the presented methods is not helpful as the GM scenario already outperforms the CV scenario; for subject S_6 unsupervised adaptation with the first 70 trials is superior to supervised adaptation but also not helpful on the long term

IV. DISCUSSION

A. Error-related potential variability and the consequences for classification

Based on the signal and feature analysis, we cannot confirm the existence of stable spatio-temporal features which would dispense the need for model adaptation. Although the N200

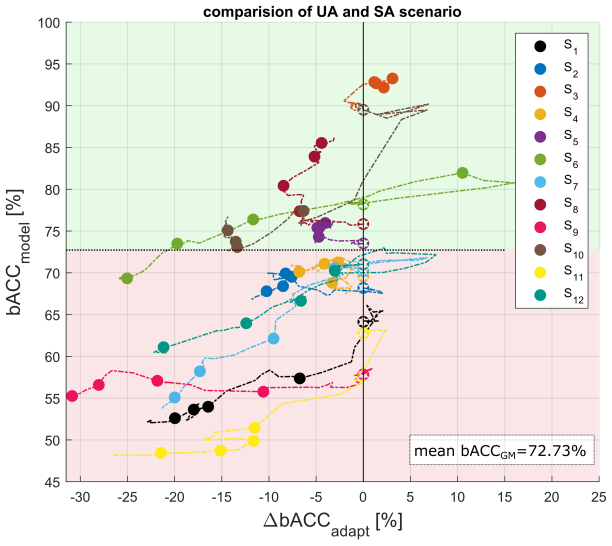

Fig. 7. Comparison of SA and UA scenario with regard to model accuracy and adaptation performance. $\Delta bACC_{adapt}(t) = bACC_{UA}(t) - bACC_{SA}(t)$ and $bACC_{model}(t)$ is the current balanced accuracy of the UA model. The dots mark the results obtained from each block of 70 trials as in Figure 6 (bottom plot). The individual initial model accuracy is equal to the GM accuracy (marked with a white star). The plot is separated in two regions (red and green) where the interception with the y-axis is at the mean $bACC_{GM}$. Note: Figures are better readable when printed in color

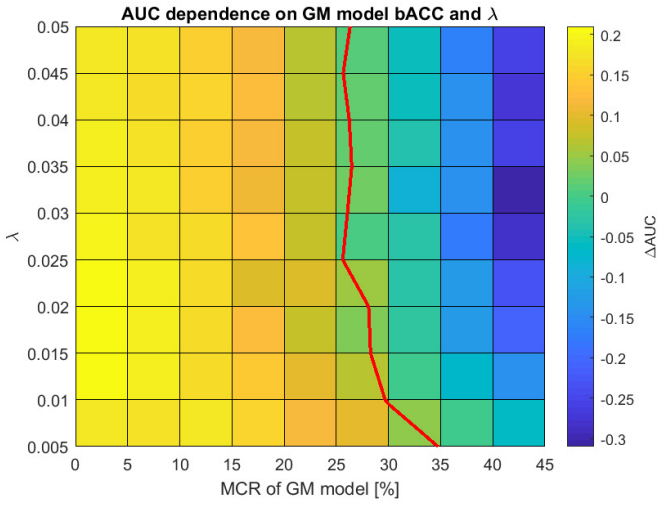

Fig. 8. Results of the grid analysis: the plot shows the color coded performance increase, averaged over all subjects, due to simulated unsupervised adaptation with respect to the GM model missclassification rate (MCR) and the update parameter λ. GM MCR was simulated by random label switching with a probability $p = MCR$. The performance increase is defined as the area under the curve (AUC) of the UA model minus the GM model. The red line separates positive and negative adaptation by the boundary of zero ΔAUC.

components seem less variant in their timing, they do not appear so in terms of amplitudes which are inconsistent across subjects. Hence, we decided against restricting the feature space to only one of the observed potentials as both exhibit noteworthy fluctuations between subjects. Furthermore, N200 *and* P300 components represent context relevant information as they are elicited *only* in erroneous trials. Hence, both components are distinct characteristics of error-trials and extracting corresponding features can help to increase the class-separability in the classification process. This ultimately also increases the classification performance if the subject-specific characteristics of the potentials are represented correctly. Wrongly estimated subject-specific characteristics however will lead to a decreased classification performance. Nonetheless, such subject-unspecific decoding models can be modified by appropriate model adaptation to better match the subject-specific signal characteristics as shown in our study and others [26], [39].

B. Model generalization

Within the context of inter-subject information transfer, an ideal generalized decoding model should capture subject-independent feature characteristics and neglect subject-specific information which could impede transferability [27]. However, each subject's ErrPs have an individual timing, amplitude and spatial distribution. Thus, the proportion of subject-specific and therefore noisy feature information is fairly high in the generalized model. In contrast, the CV scenario is unaffected by these subject-to-subject variations which explains the higher average performances of this scenario. For the same reason we suspect that the use of simplistic decoding models with a reduced number of parameters is preferable when constructing a generalized model as they are less prone to overfitting unreliable subject-independent feature characteristics from the shared training-set. This has also been suggested by other researchers [43]. Furthermore, simplistic models can be individualized efficiently as only few parameters need to be adapted. In line with the results reported in [26] we also suggest that a higher number of subjects leads to a better transferability and performance of the generalized model. This is based on the assumption that an increased number of subjects in the training set minimizes the risk of modeling subject-specific feature characteristics while at the same time enhances the modeling of subject-independent characteristics.

C. Adaptation

Regarding the adaptation, it is important to explain the lower performance of the UA scenario compared to its supervised counterpart. One obvious reason is the partially unreliable label information of new trials. While in the SA scenario, parameters of always the correct class are adapted, parameter adaptation in the UA scenario is correct only with the current accuracy of the decoder. This gives rise to the assumption of a minimal GM model accuracy needed for the UA approach to be practicable, which indeed is in line with

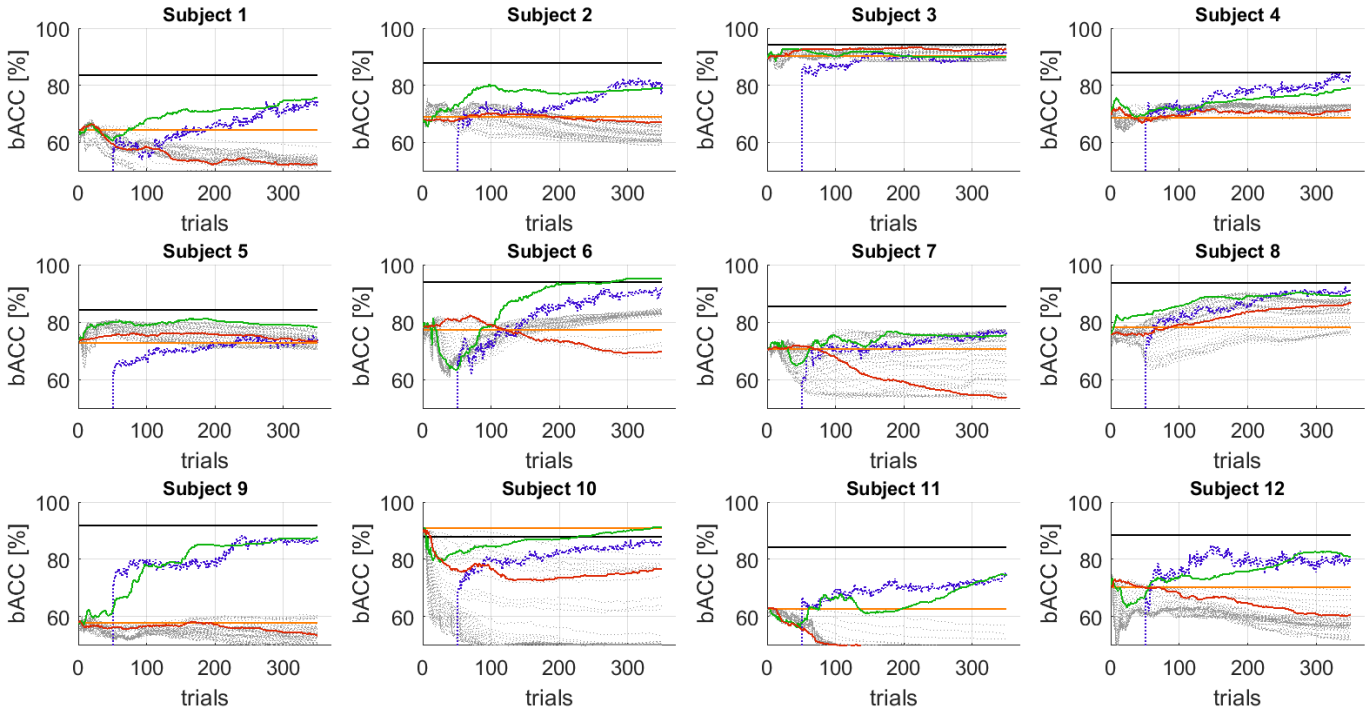

Fig. 9. Individual balanced classification accuracies for the CV (black), the MSC (blue), the GM (orange), the SA (green) and UA (red) scenarios. The grey dotted curves are the results obtained by varying the adaptation rate between the range covered by the grid search approach (0.001 - 0.050) for the UA scenario. Note: Figures are better readable when printed in color.

our findings. Consequently, increasing the generalized model accuracy is of high interest as this does not only improve the classification accuracy in an early stage but also the unsupervised adaptation performance. As we did not take into account that ErrPs of some subjects have a higher resemblance to a new subject compared to others, prior selection or prior weighting based on a suitable similarity criterion as done by other researchers [20]–[22], [28], [44], [45] form a key part of our future work. Furthermore, extending the proposed adaptation algorithm by incorporating the label certainty [24], [46], which could be used to scale the adaptation rate in the unsupervised scenario, will also be investigated in future work. Finally, in the case of unsupervised adaptation, also other adaptation algorithms which are less dependent on the generic model accuracy should be considered [47], [48].

D. Co-adaptation

In this paper we analyzed adaptive classification models. However, for any human-machine interaction (HMI) which involves adaptive systems, one also should consider bidirectional adaptation in which not only the system adapts to its user, but also the user adapts to the system. This process of co-adaptation has recently gained special attention from the BCI community [16], [30], [43], [49], [50] and ErrPs may constitute a promising complementary feedback signal for guiding adaptation towards human preferences in collaborative HMI [15].

E. Towards plug-and-play ErrP decoding

On a broader perspective, the ultimate goal are plug-and-play systems which enable reliable subject-independent calibration-free ErrP decoding in real world environments. Here, multiple other factors which go beyond the scope of this paper have to be considered. Besides system calibration, also hardware preparation is a time consuming process in classical EEG systems which can take up to almost one hour depending on the EEG system. Here, more user friendly and simplistic systems are highly needed [51]. Furthermore, not only subject-to-subject variations but also the task dependability of ErrPs [8], [17] limits the transferability of decoding models between different modalities. Although our work focused on inter-subject variations, the general principles can be easily transferred to other domain-transfer scenarios like session-to-session or task-to-task transfer. In our future work we will focus on developing a general method which is applicable to all scenarios alike and finally overcomes the domain specificity of ErrP decoding BCIs.

V. Conclusion

This work contributes to the development of practical ErrP decoding by investigating adaptive generic decoding models as an efficient and effective alternative to a laborious calibration procedure. We demonstrat the feasibility of calibration-free systems with acceptable classification accuracies of $72.7 \pm 9.66\%$, which can be initialized solely from prior subjects' data. To compensate performance losses induced by transferring non-individualized decoders, supervised adaptation to a

new subject's individual feature characteristics was investigated and results were at least comparable to the minimal sample calibration scenario. Accordingly, parameter adaptation of the class-specific means of a generalized LDA model led to an average performance increase of $+(10.31 \pm 7.69)\%$ after 350 trials. Contrary, unsupervised adaptation had an unlearning effect for most subjects. Model accuracies dropped on average $-(5.65 \pm 7.79)\%$ after 350 trials and unsupervised adaptation only proved effective for some subjects. Finally, our results demonstrate that first, instantaneous ErrP classification based on the dataset tested is feasible from the first trial on with satisfying accuracies. Second, supervised model adaptation could effectively improve the classification performance and proved superior to a traditional calibration procedure. Unsupervised adaptation is dependent on the generic model accuracy and increased the performance for most subjects with an initial model accuracy higher that the average GM $bACC$. A simulation analysis led to similar results, indicating a minimal initial model accuracy of $\sim 75\%$ for unsupervised adaptation to be practical.

ACKNOWLEDGEMENT

This work was supported by the Elite Master Program in Neuroengineering at the Technische Universität München, funded by the Elite Network Bavaria (ENB).

REFERENCES

[1] G. Cheng, "Humanoid robotics and neuroscience: Science, engineering and society," 2014.

[2] A. Wykowska, T. Chaminade, and G. Cheng, "Embodied artificial agents for understanding human social cognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1693, p. 20150375, 2016.

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

[4] H. T. van Schie, R. B. Mars, M. G. Coles, and H. Bekkering, "Modulation of activity in medial frontal and motor cortices during error observation," *Nature neuroscience*, vol. 7, no. 5, pp. 549–554, 2004.

[5] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Response error correction-a demonstration of improved human-machine performance using real-time eeg monitoring," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 173–177, 2003.

[6] M. Spüler, M. Bensch, S. Kleih, W. Rosenstiel, M. Bogdan, and A. Kübler, "Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a p300-bci," *Clinical Neurophysiology*, vol. 123, no. 7, pp. 1328–1337, 2012.

[7] S. Ehrlich and G. Cheng, "A neuro-based method for detecting context-dependent erroneous robot action," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 477–482.

[8] S. K. Ehrlich and G. Cheng, "A feasibility study for validating robot actions using eeg-based error-related potentials," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 271–283, 2019.

[9] A. Cruz, G. Pires, and U. J. Nunes, "Double errp detection for automatic error correction in an erp-based bci speller," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 26–36, 2017.

[10] F. Kalaganis, E. Chatzilari, K. Georgiadis, S. Nikolopoulos, N. Laskaris, and Y. Kompatsiaris, "An error aware ssvep-based bci," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 775–780.

[11] J. Blumberg, J. Rickert, S. Waldert, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Adaptive classification for brain computer interfaces," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 2536–2539.

[12] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. d. R. Millán, "Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control," *Scientific reports*, vol. 5, p. 13893, 2015.

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

[14] S. K. Kim, E. A. Kirchner, A. Stefes, and F. Kirchner, "Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction," *Scientific reports*, vol. 7, no. 1, pp. 1–16, 2017.

[15] S. K. Ehrlich and G. Cheng, "Human-agent co-adaptation using error-related potentials," *Journal of neural engineering*, vol. 15, no. 6, p. 066014, 2018.

[16] S. K. Ehrlich and G. Cheng, "A computational model of human decision making and learning for assessment of co-adaptation in neuro-adaptive human-robot interaction," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 264–271.

[17] I. Iturrate, L. Montesano, and J. Minguez, "Task-dependent signal variations in eeg error-related potentials for brain–computer interfaces," *Journal of neural engineering*, vol. 10, no. 2, p. 026024, 2013.

[18] I. Iturrate, J. Grizou, J. Omedes, P.-Y. Oudeyer, M. Lopes, and L. Montesano, "Exploiting task constraints for self-calibrated brain-machine interface control using error-related potentials," *PloS one*, vol. 10, no. 7, 2015.

[19] A. Myrden and T. Chau, "Effects of user mental state on eeg-bci performance," *Frontiers in human neuroscience*, vol. 9, p. 308, 2015.

[20] Y. Li, Y. Koike, and M. Sugiyama, "A framework of adaptive brain computer interfaces," in *2009 2nd International Conference on Biomedical Engineering and Informatics*. IEEE, 2009, pp. 1–5.

[21] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1318–1324, 2010.

[22] M. Sybeldon, L. Schmit, E. Sejdic, and M. Akcakaya, "Transfer learning for eeg based bci using learn++. nse and mutual information," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2632–2636.

[23] S. Lu, C. Guan, and H. Zhang, "Learning adaptive subject-independent p300 models for eeg-based brain-computer interfaces," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 2461–2465.

[24] S. Lu, C. Guan, and H. Zhang, "Unsupervised brain computer interface based on intersubject information and online adaptation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 2, pp. 135–145, 2009.

[25] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K.-R. Müller, "Toward unsupervised adaptation of lda for brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2010.

[26] I. Iturrate, L. Montesano, R. Chavarriaga, J. d. R. Millán, and J. Minguez, "Minimizing calibration time using inter-subject information of single-trial recognition of error potentials in brain-computer interfaces," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 6369–6372.

[27] P. Wang, J. Lu, B. Zhang, and Z. Tang, "A review on transfer learning for brain-computer interface classification," in *2015 5th International Conference on Information Science and Technology (ICIST)*. IEEE, 2015, pp. 315–322.

[28] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," in *2010 IEEE International conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 614–617.

[29] F. M. Schönleitner, L. Otter, S. K. Ehrlich, and G. Cheng, "A comparative study on adaptive subject-independent classification models for zero-calibration error-potential decoding," in *2019 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. IEEE, 2019, pp. 85–90.

[30] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Co-adaptive calibration to improve bci efficiency," *Journal of neural engineering*, vol. 8, no. 2, p. 025009, 2011.

[31] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[32] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of erp componentsa tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[33] R. Chavarriaga and J. d. R. Millán, "Learning from eeg error-related potentials in noninvasive brain-computer interfaces," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 18, no. 4, pp. 381–388, 2010.

[34] M. Abu-Alqumsan, C. Kapeller, C. Hintermüller, C. Guger, and A. Peer, "Invariance and variability in interaction error-related potentials and their consequences for classification," *Journal of neural engineering*, vol. 14, no. 6, p. 066015, 2017.

[35] S. Bhattacharyya, A. Konar, D. N. Tibarewala, and M. Hayashibe, "A generic transferable eeg decoder for online detection of error potential in target selection," *Frontiers in Neuroscience*, vol. 11, p. 226, 2017.

[36] S. K. Kim and E. A. Kirchner, "Handling few training data: classifier transfer between different types of error-related potentials," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 3, pp. 320–332, 2015.

[37] C. L. Dias, A. I. Sburlea, and G. Müller-Putz, "Asynchronous detection of error-related potentials using a generic classifier," in *Proceedings of the Graz Brain Computer Interface Conference 2019*, 2019.

[38] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multisubject learning for common spatial patterns in motor-imagery bci," *Computational intelligence and neuroscience*, vol. 2011, 2011.

[39] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K.-R. Müller, "Towards adaptive classification for bci," *Journal of neural engineering*, vol. 3, no. 1, p. R13, 2006.

[40] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán, "Errare machinale est: the use of error-related potentials in brain-machine interfaces," *Frontiers in neuroscience*, vol. 8, p. 208, 2014.

[41] S. H. Patel and P. N. Azzam, "Characterization of n200 and p300: selected studies of the event-related potential," *International journal of medical sciences*, vol. 2, no. 4, p. 147, 2005.

[42] S. K. Kim and E. A. Kirchner, "Choice of training data for classifier transfer in error related potentials based on signal characteristics," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2015, pp. 102–105.

[43] J. S. Müller, C. Vidaurre, M. Schreuder, F. C. Meinecke, P. Von Bünau, and K.-R. Müller, "A mathematical model for the two-learners problem," *Journal of neural engineering*, vol. 14, no. 3, p. 036005, 2017.

[44] F. Putze, D. Heger, and T. Schultz, "Reliable subject-adapted recognition of eeg error potentials using limited calibration data," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 419–422.

[45] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1352–1359, 2019.

[46] T. Zeyl, E. Yin, M. Keightley, and T. Chau, "Adding real-time bayesian ranks to error-related potential scores improves error detection and auto-correction in a p300 speller," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 24, no. 1, pp. 46–56, 2015.

[47] Y. Li and C. Guan, "An extended em algorithm for joint feature extraction and classification in brain-computer interfaces," *Neural Computation*, vol. 18, no. 11, pp. 2730–2761, 2006.

[48] D. Huebner, T. Verhoeven, K.-R. Mueller, P.-J. Kindermans, and M. Tangermann, "Unsupervised learning for brain-computer interfaces based on event-related potentials: Review and online comparison [research frontier]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 2, pp. 66–77, 2018.

[49] R. J. Kobler and R. Scherer, "Restricted boltzmann machines in sensory motor rhythm brain-computer interfacing: a study on inter-subject transfer and co-adaptation," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2016, pp. 000 469–000 474.

[50] J. Mattout, M. Perrin, O. Bertrand, and E. Maby, "Improving bci performance through co-adaptation: applications to the p300-speller," *Annals of physical and rehabilitation medicine*, vol. 58, no. 1, pp. 23–28, 2015.

[51] S. Ehrlich, A. Alves-Pinto, R. Lampe, and G. Cheng, "A simple and practical sensorimotor eeg device for recording in patients with special needs," in *Neurotechnix2017, CogNeuroEng 2017, Symposium on Cognitive Neural Engineering*, 2017.

Florian M. Schönleitner received his Bachelor in Engineering Sciences at the Technische Universität München (2018) and is currently enrolled in the elite master program Neuroengineering. He is member of the Elite Network of Bavaria (2019-) and employed at the Institute for Cognitive Systems (2018-), which is part of the Faculty of Electrical and Computer Engineering at the Technische Universität München. In his research he focuses on advanced machine learning algorithms for the decoding of neuronal signals in brain-computer interfaces.

Lukas Otter received his Bachelor's degree in Engineering Science at the Technische Universität München (2017). He is pursuing a Master's degree in electrical engineering and information technology with a focus on neuroengineering. He is interested in signal processing and machine learning algorithms for brain-computer interfaces and neuroprosthetics.

Stefan K. Ehrlich is doctoral researcher in electrical and computer engineering at the Chair for Cognitive Systems at Technische Universität München. He is conducting research on non-invasive EEG-based brain-computer interfaces (BCI). His work is particularly centered around neuro-adaptive brain-robot interfaces for augmentation of human-robot interaction, affective neurofeedback BCIs utilizing synthesized sound and music stimuli, and practical EEG-based ease-of-use neurotechnology. His long-term research goals include the development of BCI-based technology fostering novel approaches to neuro-rehabilitation and the understanding of neurological disorders.

Gordon Cheng holds the Chair for Cognitive Systems (2010-), Prof. Cheng is Founder and Director of Institute for Cognitive Systems, Faculty of Electrical and Computer Engineering at Technical University of Munich, Munich/Germany. He is the coordinator of the Center of Competence Neuro-Engineering (2013-). He is the director of the Elite Master of Science program in Neuroengineering (MSNE) of the Elite Network of Bavaria (2016-). He received a PhD (2001) in Systems Engineering from the Department of Systems Engineering, The Australian National University. Bachelor (1991) and Master (1993) degrees in Computer Science from the University of Wollongong, Australia. He has extensive industrial experiences in consultancy as well as contractual development of large software systems. He was also the Founder/CEO of the company, G.T.I. Computing (1995-2006), a company he founded specializing in networking and transport management systems in Australia. His research interests include, humanoid robotics, cognitive systems, artificial robot skin, brain machine interfaces, bio-mimetic of human vision, computational neuroscience of vision, action understanding, human-robot interaction, active vision and mobile robot navigation. He is the co-inventor of approximately 20 patents and author of approximately 300 technical publications, proceedings, editorials and book chapters. He has been named IEEE Fellow 2017 for "contributions in humanoid robotic systems and neurorobotics".