Choice of training data for classifier transfer in error related potentials based on signal characteristics*

Su Kyoung Kim[1,2] and Elsa Andrea Kirchner[1,2]

Abstract— **In brain computer interfaces, different amounts of training data can be generated during the same recording time depending on the type of e.g., error related potential (ErrP) that is evoked. In our previous study (Kim & Kirchner, 2013), we obtained more training data containing observation ErrPs compared to interaction ErrPs within the same recording time under similar scenario conditions. Thus we trained a classifier on observation ErrPs to detect interaction ErrPs. This led to the reduction of calibration time. In this previous study we assumed that features extracted from a window, in which both types of ErrPs show a similar shape of averaged activity (0.16–0.6 s after error events), are optimal for classifier transfer. In this study we test this assumption on a larger group of subjects. Further, we evaluate an extended training window that covers a late negativity at 0.6–0.8 s, which has a stronger amplitude in case of observation ErrPs. Such an extension of the training window allows to improve the classification performance in case that observation ErrPs are used to train and test a classifier (*no* transfer case). However, in this study we will show that for the *transfer* case this long window [0.16–0.8 s] is outperformed by the short window [0.16–0.6 s], which contains only the part of both types of ErrPs with similar shape. The results indicate that the signal characteristics can guide the choice of training data for classifier transfer between different types of ErrPs.**

I. Introduction

Brain computer interfaces (BCIs) that link a human user and external systems are applied in different research areas (reviews for EEG-based BCIs in general see [1], [2], review for error-related potentials (ErrPs)-based BCIs [3], review for P300-based BCIs [4], review for movement-based BCIs [5], review for EEG-EMG based BCIs [6], and visual-evoked potential (VEP)-based BCIs [7]).

A successful and robust detection of specific pattern in the electroencephalogram (EEG), which correlates with the user's intent, is often a challenge in real-world BCI applications with complex application environments or situations (e.g., [8]–[11]). A further challenge is to develop scenarios which allow to generate enough training data for a BCI based support of such real-world or realistic application scenarios. Especially real-world applications using error-related potentials (ErrPs), which are elicited by recognizing erroneous behaviors, are challenging, since in general erroneous behaviors do not often occur in real-world applications. This leads to a long recording time to collect enough training data. Hence, from the perspective of an application it is of interest to develop scenarios which allow to collect a sufficient amount of training instances within a reasonable recording time.

The amount of training data can be affected by the nature of ErrPs. Our scenario is designed to generate two different types of ErrPs by performing different tasks (interaction/observation). Depending on who performs the task (e.g, a subject or an artificial agent) the different types of ErrPs are generated using the same scenario concept. Interaction ErrPs are elicited in the EEG of users who recognize interaction errors, which occur when an interface misinterprets the user's intent and sends a wrong control signal to an external system [12]. On the other hand, observation ErrPs are elicited in the EEG of a human observer who monitors the erroneous behavior of an external system [13]. Thus, observation ErrPs are elicited without any interaction with interfaces or external systems, whereas during interaction additional time that the user needs for decision making processes is necessary for the interaction task. Such different natures of involved brain processes eliciting both ErrP types lead to different amounts of training data during the same recording time (i.e., twice as much observation ErrPs are collected compared to interaction ErrPs during the same calibration time in our scenario), although the scenario concept used to generate both types of ErrPs was the same for both tasks.

In the previous study, where we showed that a classifier transfer between different scenarios can reduce calibration time [14], we used features extracted from the window, in which both types of ErrPs show a similar shape of averaged activity. The reason for such window selection was based on the assumption that the use of a window containing the part of both types of ErrPs which have a similar shape is optimal for classifier transfer. In this study, we test our previous assumption by investigating another window that contains a late negativity with an amplitude that is higher for observation ErrPs compared to interaction ErrPs on a larger group of subjects (i.e., 8 instead of 4 subjects). In the previous study, we achieved a higher classification performance for single trial detection of observation ErrPs in the *no* transfer case, i.e., in case that observation ErrPs including this late window are used to train *and* test a classifier [14]. Thus, in this paper we investigate a possible contribution of this later negativity for classifier transfer.

To this end, we investigate two time windows for classifier transfer: 1) A short time window [0.16 s–0.6 s] containing a similar averaged shape of each type of ErrPs and 2) A longer time window [0.16 s–0.8 s] containing an additional

*This work is supported by the German Ministry of Economics and Technology (grant no. 50 RA 1011 and grant no. 50 RA 1012).

[1]Su Kyoung Kim and Elsa Andrea Kirchner are with the Robotics Innovation Center, German Research Center for Artificial Intelligence (DFKI) GmbH, Robert-Hooke-Strasse 1, 28359, Bremen, Germany. {su-kyoung.kim, elsa.kirchner}@dfki.de

[2]Su Kyoung Kim and Elsa Andrea Kirchner are with the Robotics Lab, Faculty of Mathematics and Computer Science, University of Bremen, Germany. {sukim, ekir}@informatik.uni-bremen.de

late negativity that is only observed in observation ErrPs. The selection of different windows is based on the averaged shape of each type of ErrP (see section II-D). We assume that the averaged shape of each type of ErrPs can guide window selection for feature extraction in case of classifier transfer. Hence, the comparison of classification performance between two time windows can support our assumption.

II. Methods

A. Experimental Design

We used the scenario as depicted in Figure 1–(a) which was developed in our previous study [14]. Using this scenario, two different tasks (interaction/observation) were performed to generate two different types of ErrPs (interaction ErrPs/observation ErrPs) separately (more details, see [14]).

In both tasks, 20 targets had to be reached in numeric order by moving the cursor (see, Figure 1–(a)). Here, both obstacles placed among the targets and the spikes of targets had to be avoided. If any task rule was violated, a penalty was given (e.g., the cursor went back to the start position, when touching a target spike). In case that a target was reached in the wrong order, the target color remained red. For the correct case, the target color changed to green.

In the interaction task, subjects were instructed to check the target points by using a computer keyboard. All subjects needed about 2 minutes to finish one set. Here, interaction errors were programmed with a probability of 9%. Thus, the current movement direction of the cursor did not always correspond to the movement direction that was chosen by the subjects by pressing a certain key. The possible directions of wrong movements were uniformly distributed. In this way interface errors (i.e., errors of a classifier) were simulated as for the scenario described in [12]. When the subjects recognized wrong movements of the cursor (i.e., interaction errors) interaction ErrPs were elicited in the subject's EEG. In comparison with the scenario used in [12], our scenario was closer to a more realistic application. Thus, we could not exclude response errors made by the subjects themselves [15] (e.g., violating the target order or touching the spikes of a target). Therefore, two kinds of errors were expected: a) interaction errors and b) response errors

In the observation task, an artificial agent performed the task and the subjects observed the behavior of the agent. Task rules were the same as for the interaction task. Here, wrong behaviors of the agent were again programmed with a probability of 9%. When monitoring the wrong behaviors of the agent, observation ErrPs were elicited in the subject's EEG.

For both tasks, the order of targets was randomized for each set to avoid the same task pattern. The empirical ratio of erroneous and correct trials was 1:10 in both tasks. However, the average time between consecutive cursor movements was slower for the subjects (227 ms) compared to the agent (110 ms), since the subjects paused often to find the correct path to reach the targets. In contrast, the speed of key pressing was hard coded in the observation task. Further, the path to reach the targets and its deviation were also hard coded to obtain as much trials as possible within a fixed time. Thus, the way to reach the targets were not optimally programmed compared to the strategy to reach the targets which was chosen by subjects. Accordingly, twice as much erroneous trials were collected from the observation task compared to the interaction task during the same recording time.

B. Data Acquisition

Eight subjects (two females/six males, age: 26.5 ± 3.25, right-handed, normal or corrected to normal vision) participated in the study. All subjects provided written consent to participate in the study approved by the ethics committee of the University of Bremen. The study was conducted in accordance with the Declaration of Helsinki. EEGs were recorded using an actiCap system (Brain Products GmbH, Germany), in which 64 active electrodes were arranged in accordance to the extended 10-20 system with reference at FCz. Impedance was kept below $5\,\mathrm{k}\Omega$. EEG signals were sampled at 5 kHz, amplified by two 32 channel BrainAmp DC Amplifiers (Brain Products GmbH, Germany), and filtered with a low cut-off of 0.1 Hz and high cut-off of 1 kHz.

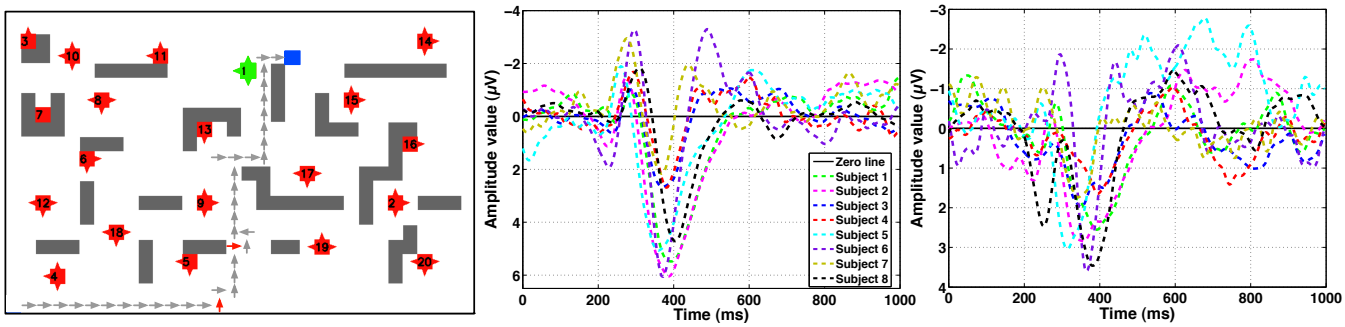
C. Data Set

We collected seven data sets from each task. For each task, one set took 2 minutes. One data set from the observation task contains 99 erroneous trials and 990 correct trials, whereas one data set from the interaction task contains 48 erroneous trials and 480 correct trials. That means, a different amount of trials was recorded for each task during the same recording time per set. From the perspective of an application a classifier transfer from the observation task to the interaction task can be useful to reduce calibration time needed to detect interaction ErrPs. To this end, the data set collected from the observation task was used to train the classifier for detecting interaction ErrPs, which were generated in the interaction task. One data set from each task was used for classifier transfer.

D. Preprocessing and Classification

The continuous EEG signal was segmented into epochs from 0 s to 1 s after each event type (correct/erroneous trial). All epochs were normalized to zero mean for each channel, decimated to 50 Hz, and band pass filtered (0.5 to 10 Hz). The xDAWN [16] was used as a spatial filter to enhance the signal-to-noise ratio. By applying the xDAWN the number of 64 physical channels was reduced to 8 pseudo channels.

Two time windows were used for feature generation: [0.16 s–0.6 s] and [0.16 s–0.8 s]. The selection of the two time windows was based on the shape of averaged interaction and observation ErrPs. As shown for Figure 1–(b) and 1–(c), both types of ErrPs showed a first negative peak around 0.27 s after the erroneous events, followed by a positive peak around 0.38 s. As for the averaged interaction ErrPs, we observed a narrow negativity peak in the late time window [0.4 s–0.6 s], whereas a broad negative peak including two negative peaks in the late time window [0.4 s–0.8 s] was



(a) Experimental paradigm (b) Interaction ErrPs (c) Observation ErrPs

Fig. 1. (a): A subject or an artificial agent moves the cursor (blue) towards one of 20 targets (red) in numeric order. When a target is reached in the correct order, the color of the target changes from red to green. The track of cursor movements is depicted by gray arrows towards the chosen direction and the track of wrong cursor movements is depicted by red arrows (direction of errors). (b) and (c): Averaged event related potential (ERP) for the difference error-minus-correct trials at channel FCz for each subject. Only artifact-free EEG trials were used.

TABLE I

CLASSIFICATION PERFORMANCE (MEAN±STANDARD DEVIATION) IN CASE OF CLASSIFIER TRANSFER.

Observation ErrP (training: calibration time of 2 min) → Interaction ErrP (test) / feature extraction from the *shorter* window of 0.16 s–0.6 s									
	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Average
bACC	0.83±0.01	0.70±0.04	0.76±0.01	0.81±0.01	0.82±0.01	0.76±0.02	0.77±0.01	0.88±0.01	0.79±0.06
TPR	0.74±0.01	0.55±0.13	0.58±0.03	0.80±0.02	0.75±0.05	0.77±0.04	0.66±0.05	0.81±0.01	0.71±0.05
TNR	0.91±0.01	0.84±0.06	0.94±0.01	0.82±0.02	0.88±0.03	0.74±0.02	0.89±0.02	0.94±0.01	0.87±0.07
Observation ErrP (training: calibration time of 2 min) → Interaction ErrP (test) / feature extraction from the *longer* window of 0.16 s–0.8 s									
	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Average
bACC	0.81±0.02	0.65±0.01	0.75±0.01	0.74±0.06	0.81±0.01	0.72±0.01	0.78±0.01	0.85±0.01	0.76±0.06
TPR	0.74±0.05	0.65±0.01	0.58±0.02	0.56±0.13	0.72±0.09	0.68±0.01	0.66±0.01	0.73±0.01	0.67±0.07
TNR	0.87±0.01	0.65±0.01	0.92±0.02	0.91±0.02	0.68±0.04	0.76±0.02	0.73±0.01	0.96±0.01	0.86±0.1

observed for the averaged observation ErrPs. Compared to both types of averaged activity, we observed a similar shape of averaged event related potential (ERP) activity for the time window of [0.4 s–0.6 s], whereas a difference in the averaged shape between two types of ErrPs was shown for the later time window [0.6 s–0.8 s].

For each time window, features were extracted from the obtained 8 pseudo channels after spatial filtering, between 0.16 s and N s where N $\in \{0.6, 0.8\}$. We obtained 176 features (8 channels $\times$ 22 data points = 176) from the shorter time window [0.16 s–0.6 s] and 256 features (8 channels $\times$ 32 data points = 256) from the longer time window [0.16 s–0.8 s].

The extracted features were normalized and used to train a classifier. We used a linear support vector machine (SVM) [17] to classify *correct* and *erroneous* trials. For each training, the complexity parameter of the SVM was optimized with an internal 5-fold cross validation using a grid search among the predetermined complexity values [10^0, 10^{-1}, ... , 10^{-6}]. The parameter optimization was repeated ten times and the construction of splits was different for each repetition. Due to the unbalanced ratio of *erroneous* and *correct* trials (1:10), different penalty constants were used for both classes [18]. We determined a class weight of 5 for the under-represented class as penalty so that making errors on under-represented instances was costlier than making errors on over-represented instances.

E. Evaluation

As mentioned earlier, the classifier trained on observation ErrPs was transferred to interaction ErrPs. To this end, we used one data set from each task. The classifier was trained on one data set containing observation ErrPs (99 erroneous trials, calibration time of 2 min). After that, the trained classifier was used to evaluate one data set containing interaction ErrPs (48 erroneous trials).

As performance metric, we used the arithmetic mean of true positive rate (TPR) and true negative rate (TNR), the so-called *balanced accuracy (bACC)*. Here, the erroneous trials were the positive instances. This metric is less sensitive to imbalanced data (i.e., unbalanced ratio of the two classes) compared to other metrics, e.g., accuracy (details, see [19]).

To compare both time windows that were used to extract features for classifier transfer, repeated measures ANOVA were performed with *time window* and *subject* as two within-subjects factors.

III. RESULTS

Table I shows the classification performance on interaction ErrPs using the classifier trained on observation ErrPs, in which two time windows (short [0.16 s–0.6 s] vs. long

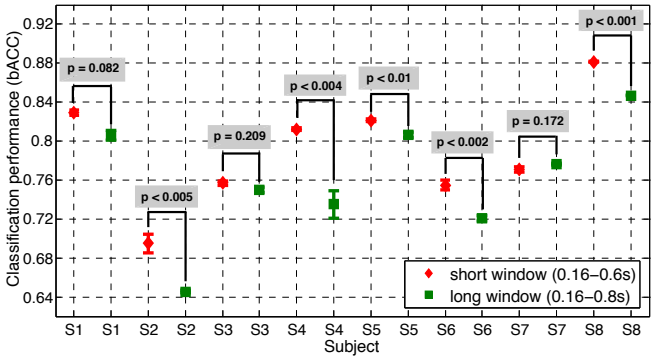


Fig. 2. Classification performance in case of classifier transfer [observation ErrPs (Training) → interaction ErrPs (test)] when using different time windows of observation ErrPs: short time window vs. long time window

[0.16 s–0.8 s]) were used to extract features from observation ErrPs.

We achieved a bACC of 0.79 (across all subject) with the *short* time window and a bACC of 0.76 (across all subjects) with the long time window. In case of using the *short* time window five subjects showed a *higher* classification performance compared to the case of using the long window. For three other subjects, there was no statistically significant difference between both time windows. Statistical values for each subject are depicted in Figure 2.

In summary, a higher classification performance was achieved for five subjects when using the *shorter* time window containing the part of the averaged ERP activity that has a similar shape for both types of ErrPs compared to the window including also a later negativity in case that observation ErrPs are evoked.

IV. Conclusion

In this study, we have shown that the short window was sufficient to transfer a classifier trained on interaction ErrPs to observation ErrPs. The long time window did not lead to a higher classification performance as in the *no* transfer case (i.e., observation ErrPs were trained and tested) [14]. In case of classifier transfer, we achieved no significant difference between the long and short time window for three subjects. However, a significant higher classification performance was obtained with the short time window for the remaining five subjects. Further, the use of the short window is beneficial with regard to computation time needed to detect single trial detection.

The paper shows that the window selection for feature generation may be relevant for classifier transfer between different types of ErrPs. The assumption that similarity in averaged activity between different types of ErrPs can guide the selection of training windows in case of classifier transfer is supported by this study. Furthermore, a later negativity with a higher amplitude in case that observation ErrPs are evoked plays a relevant role for single trial detection of observation ErrPs (i.e., during *no* transfer) [14]. However, this later negativity is not relevant for single trial detection of interaction ErrPs or does even reduce classification performance (i.e., during *classifier transfer*).

Acknowledgment

This work is supported by the German Ministry of Economics and Technology (grant no. 50 RA 1011 and grant no. 50 RA 1012).

References

[1] J. R. Wolpaw, N. Birbauer, D. J. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, pp. 767–791, 2002.

[2] J. d. R. Millán, R. Rupp, G. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb, C. Neuper, K.-R. Müller, and D. Mattia, "Combining brain-computer interfaces and assistive technologies: State-of-the-art and challenges," *Front. Neurosci.*, vol. 4, no. 161, 2010.

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

[4] J. Mak, Y. Arbel, J. Minett, L. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, and D. Erdogmus, "Optimizing the P300-based brain–computer interface: current status, limitations and future directions," *J. Neural Eng.*, vol. 8, no. 2, p. 025003, 2011.

[5] P. Ahmadian, S. Cagnoni, and L. Ascari, "How capable is non-invasive EEG data of predicting the next movement? A mini review," *Front. Hum. Neurosci.*, vol. 7, p. 124, 2013.

[6] T. D. Lalitharatne, K. Teramoto, Y. Hayashi, and K. Kiguchi, "Towards hybrid EEG-EMG-based control approaches to be used in bio-robotics applications: Current status, challenges and future directions," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 147–154, 2013.

[7] G. Bin, X. Gao, Y. Wang, B. Hong, and S. Gao, "VEP-based brain–computer interfaces: time, frequency, and code modulations," *IEEE Computational Intelligence Magazine*, vol. 4, no. 4, pp. 22–26, 2009.

[8] E. A. Kirchner, S. K. Kim, S. Straube, A. Seeland, H. Wöhrle, M. M. Krell, M. Tabie, and M. Fahle, "On the applicability of brain reading for predictive human-machine interfaces in robotics," *PLoS ONE*, vol. 8, no. 12, p. e81732, Dec 2013.

[9] A. Seeland, H. Wöhrle, S. Straube, and E. A. Kirchner, "Online movement prediction in a robotic application scenario," in *Proc. 6th Int. IEEE EMBS Conf. Neural Eng. (NER)*, Nov 2013, pp. 41–44.

[10] D. C. Jangraw, J. Wang, B. J. Lance, S.-F. Chang, and P. Sajda, "Neurally and ocularly informed graph-based models for searching 3D environments," *J of Neural Eng.*, vol. 11, no. 4, p. 046003, 2014.

[11] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. Millán, "Latency correction of event-related potentials between different experimental protocols," *J of Neural Eng.*, vol. 11, no. 3, p. 036005, 2014.

[12] P. W. Ferrez and J. d. R. Millán, "Error-related EEG potentials generated during simulated brain-computer interaction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 923–929, March 2008.

[13] I. Iturrate, L. Montesano, and J. Minguez, "Single trial recognition of error-related potentials during observation of robot operation," in *Proc. 32th Annu. Int. Conf. IEEE Eng. Med. and Bio. Soc.*, 2010, pp. 4181–4184.

[14] S. K. Kim and E. A. Kirchner, "Classifier transferability in the detection of error related potentials from observation to interaction," in *Proc. IEEE Int. Conf. Sys., Man, and Cybern. (SMC)*, Oct 2013, pp. 3360–3365.

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

[16] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, 2009.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans Intell. Sys. Technol. (TIST)*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[18] K. Veropoulos, C. Campbell, N. Cristianini, *et al.*, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. Artific. Intell.*, 1999, pp. 55–60.

[19] S. Straube and M. M. Krell, "How to evaluate an agent's behavior to infrequent events? – reliable performance estimation insensitive to class distribution," *Front. Comput. Neurosci.*, vol. 8, no. 43, 2014.