Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

IntEr-HRI Competition: Intrinsic Error Evaluation during Human-Robot Interaction

Kartik Chari[1*] , **Niklas Kueper**[1*] , **Su Kyoung Kim**[1] , **Frank Kirchner**[1,3] and **Elsa Andrea Kirchner**[1,2*]

[1]Robotics Innovation Centre, German Research Centre for Artificial Intelligence (DFKI), Bremen, Germany.

[2]Institute of Medical Technology Systems, University of Duisburg-Essen, Duisburg, Germany.

[3]Robotics Lab, University of Bremen, Bremen, Germany.

[*]{kartik.chari, niklas.kueper}@dfki.de, elsa.kirchner@uni-due.de

Abstract

Reliable detection of human intentions from electroencephalogram (EEG) to improve human-robot interaction (HRI) has recently gained significant importance. To ensure safe and satisfactory interactions, implicit detection of erroneous behavior of robotic systems, particularly assistive devices, is essential. This can be achieved by detecting error-related potentials (ErrPs) in EEG, evoked by visual, tactile, or visuo-tactile stimuli. Of these, the ErrPs evoked tactilely with the help of a robot remains unexplored and has been the main focus of this competition. The task for participating teams was to develop robust AI models for continuous real-time classification of erroneous behavior of assistive robotic devices from the human EEG. Even though the competition results prove its feasibility, a performance gap (balanced accuracy and computation time) of more than 10% was observed between the offline and online classification of errors in real-world scenarios. In addition to the competitive AI models developed by the participating teams, this competition also contributed towards a one-of-its-kind open-access EEG and EMG dataset, a lossless live streaming solution for EEG data, and a novel quantitative metric for benchmarking online asynchronous EEG detection solutions.

1 Introduction

EEG-based detection of human intentions and their dynamic changes have gained significant importance in human-robot interaction (HRI). It has been thoroughly studied in brain-computer interface (BCI) applications, especially in robot learning and human-robot co-adaptation. In particular, error-related potentials (ErrPs) evoked in the brain when observing erroneous actions of other humans or systems, such as robots, have been extensively studied across various research areas (see [Chavarriaga *et al.*, 2014] for a comprehensive review). Such studies enable the decoding of human intentions using robots, thereby enhancing communication with humans and improving behavioral strategies based on real-time updates (e.g., [Kim *et al.*, 2017], [Kim *et al.*, 2020]).

In most ErrP-based applications, erroneous behavior of robots was detected with the help of visual cues (e.g., [Iturrate *et al.*, 2010],[Kim and Kirchner, 2013], [Kim and Kirchner, 2015]). However, certain studies have used visuo-tactile stimuli to evoke ErrPs. These studies either integrated both visual and tactile channels to perceive erroneous behaviors of systems ([Tessadori *et al.*, 2017], [Schiatti *et al.*, 2019]) or used the visual channel to recognize errors and the tactile channel (e.g., vibration) solely to indicate upcoming errors([Perrin *et al.*, 2008], [Chavarriaga *et al.*, 2012], [Ahkami and Ghassemi, 2021]). Moreover, to the best of our knowledge, only tactile-based ErrP detections have not been investigated yet. Thus, this competition served as a means to encourage systematic investigation of tactile-based ErrP detection in HRI using machine learning approaches.

Furthermore, there were a few competitions that recorded openly accessible EEG datasets where specific brain patterns in the time or frequency domain were evoked synchronously in response to, for example, motor imagery or workload without the use of robots (e.g., [Roy *et al.*, 2022], [Blankertz, 2004], [Blankertz, 2008]). There are also some open EEG datasets where ErrPs were evoked synchronously by visual stimuli from observing robots' incorrect behavior (e.g., [Ehrlich and Cheng, 2019]). However, there are no open-access datasets where robots are used to evoke ErrPs tactilely.

Thus, to address this gap, for this competition, we developed an HRI scenario in which subjects tactilely perceived the incorrect behavior of an active orthosis device (see [Kueper *et al.*, 2024]) during the execution of arm movements (flexions and extensions). The recorded open source EEG and EMG dataset (see Section 2.1) from the offline stage of the competition is accessible on Zenodo[1]. This dataset can be used to further research the asynchronous detection of tactilely evoked ErrPs, which was also the goal of our competition.

This paper is structured as follows: Section 2 describes the task and structure of the competition. In Section 3, we provide details on the evaluation metrics used to determine com-

[1]https://zenodo.org/records/8345429

petition winners, while Section 4 presents an overview of the results. Finally, Section 5 discusses the key takeaways and insights provided by the competition.

2 Task and Structure of the Competition

As described in Section 1, the competition's goal was to encourage the scientific community to contribute towards improving the online classification of tactile-based error onsets in EEG data. Thus, the participating teams were tasked with developing robust and reliable signal processing and machine learning approaches to detect erroneous behaviors through single-trial EEG analysis asynchronously.

The competition comprised two stages - *offline* stage and *online* stage, and only the top-performing teams from the offline stage progressed to the online stage. The winners of the competition were announced at the IJCAI 2023 conference. During the conference, we also organized five keynote lectures (available on our competition homepage[2]), a live panel discussion and paper presentations by participating teams.

2.1 Offline Stage of Competition

In the offline stage, we recorded the first open-access EEG and EMG dataset[1], wherein tactile errors were deliberately introduced via an active orthosis device (see [Kueper *et al.*, 2024] for detailed information). Eight healthy subjects participated in the study, and 10 data sets were recorded, each consisting of 30 flexion or extension movement trials. Six trials in each set were randomly chosen for introducing tactile errors for a duration of 250 ms. Whenever the subjects felt an error, they were instructed to press an air-filled ball in their left hand (direct response), and these events were marked in the labeled EEG data.

The challenge consisted of a *training* phase and a *testing* phase. During the training phase, participants accessed only eight of the 10 recorded datasets, each containing labeled EEG data, to train their classifiers in detecting error onsets (six errors in each set). Furthermore, teams validated their models via 10-fold cross-validation and submitted results in a short paper (accessible on our competition homepage[2]). Subsequently, in the testing phase, the remaining two sets were provided as unlabeled EEG data, and the teams tested their pre-trained models offline and submitted error onset prediction results.

2.2 Online Stage of Competition

During the online stage, unlabeled EEG data was streamed via the Lab Streaming Layer (LSL) ([Kothe *et al.*, 2018]) with the help of a VPN tunnel. The challenge was to employ a pre-trained classifier for real-time error onset detection in the streamed EEG data. As we wanted to evaluate the robustness of the model in real-world session transfer scenarios, no calibration phase was included. Instead, a dry run (without real data) was conducted a day before the competition to test live data streaming through LSL and to ensure the general functionality of the teams' code.

On the competition day, a subject from the offline stage was prepared for the experiment, and the teams were notified about the subject in advance. The online experiment consisted of two cases: *direct response* case and *no response* case. The direct response case involved the subject pressing an air-filled ball upon recognizing an error, whereas in the latter case, the subject refrained from any response. These cases were introduced to evaluate the impact of motor-related activity in the EEG in real-world scenarios. Furthermore, each case included two sets of 30 movement trials wherein five trials were randomly selected to deliberately introduce tactile errors lasting 250 ms.

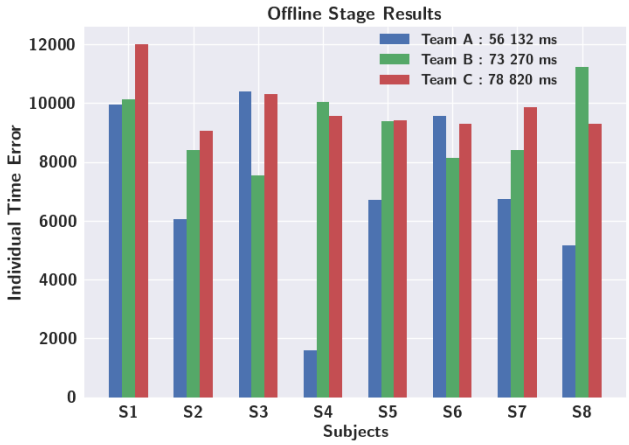

Figure 1: Bar plot of offline stage results for the top three teams. Each bar in a group represents the accumulated time error for all samples in the two test sets for each subject. The total accumulated score for each team is provided in the legend.

3 Performance Evaluation

The winners of the competition were decided through a quantitative performance evaluation scheme as detailed on our competition home page[2]. Each of the two stages had different evaluation metrics, as described below.

3.1 Offline Stage Evaluation

In the offline stage, as mentioned in Section 2.1, each team tested their classifier models on the 16 test sets (8 subjects; 2 test sets per subject) and submitted the sample indices corresponding to the predicted error onsets. Each submitted sample index was compared against reference ground truth to calculate the time error. The accumulated error score for the team was the sum of all individual time errors. If the error was negative (early prediction) or greater than 1000 ms (late detection), it was classified as a false positive. Each false classification incurred a penalty of 1000 ms. This method was designed to reward timely detection of the robot's erroneous behavior, which is critical when using assistive robotic devices.

3.2 Online Stage Evaluation

To determine the winners of the online stage, we devised the Final Performance Score (FPS). This novel quantitative metric combines balanced accuracy ($BAcc$) with a new metric

[2]https://ijcai-23.dfki-bremen.de/competitions/inter-hri/

termed time score (t_{score}). This evaluation was based on two critical aspects: accurate error onset prediction and real-time classification of errors. Both these aspects were quantified into a single parameter known as the individual sample score using logistic regression. Thus, each set's time score (t_{score}) was computed by summing up all the individual sample scores and normalizing them between 0 and 1. Finally, the FPS was calculated by assigning a weightage of 70% to $BAcc$ and 30% to t_{score}. The closer the FPS is to 1, the better the trained classifier's ability to accurately classify tactile-based error onsets in real time with minimal computation delay.

4 Competition Results

As discussed in Section 2, this competition was designed to allow only the best-performing teams from the offline stage to participate in the online stage. This section presents the results of both stages of the competition. Note that the names and affiliations of the top-performing teams are mentioned in the demonstration video (https://youtu.be/JFC-Kc3FHEc).

4.1 Results for the Offline Stage

A total of six teams had registered for the offline stage, out of which only three teams could solve the task and qualify for the online stage. The offline stage results of these three teams are provided in Figure 1.

As the accumulated error score was the summation of time errors across all the sets, a lower score was more desirable. Thus, team A was the clear winner of the offline stage, followed by team B and team C.

4.2 Results for the Online Stage

Of the three teams that qualified for the online stage, only one team (team B) could detect and classify errors in real time. Its performance was evaluated per the metrics described in Section 3.2 and provided in Figure 2. The arithmetic mean of all four individual FPS gave the final overall performance metric for the online classifier model, which was 0.805 on a scale of [0,1].

5 Key Insights

This competition proved to be a success for varied reasons. In addition to providing valuable technical insights to the participating teams and the organizers, it successfully demonstrated the challenges of asynchronous online classification of error-related activities from the human EEG.

Despite the top three teams having an average validation $BAcc$ score of around 90% in the offline stage, their performance dropped greatly in the online case. Two of the three finalists failed to detect a single error, and the winning team scored an average $BAcc$ of less than 80%. The main reason for this discrepancy lies in the low signal-to-noise ratio (SNR) and high variability of EEG signals across different sessions, even for the same subject. These inherent properties of EEG make the development of a robust machine-learning model quite challenging (see [Wu *et al.*, 2022] for a comprehensive review).

We also encountered several challenges while organizing this competition. The most significant among them was designing the experiment to evoke an ErrP through robot-induced tactile error introduction. Additionally, it was essential to design a framework that could stream live EEG data with minimal communication delay and near-zero sample loss for the online stage. This framework was created by combining a VPN tunnel and the LSL protocol as mentioned in Section 2.2. Furthermore, we developed a novel performance evaluation metric that could be used for benchmarking asynchronous online EEG classifiers. This metric considers key evaluation aspects, such as accurate prediction, compensation of class imbalance, and computation time, that determine how well the asynchronous detection works. Moreover, we integrated this metric with our live-streaming framework to ensure that the communication delays and sample losses did not impact the comparison of teams' online performances.

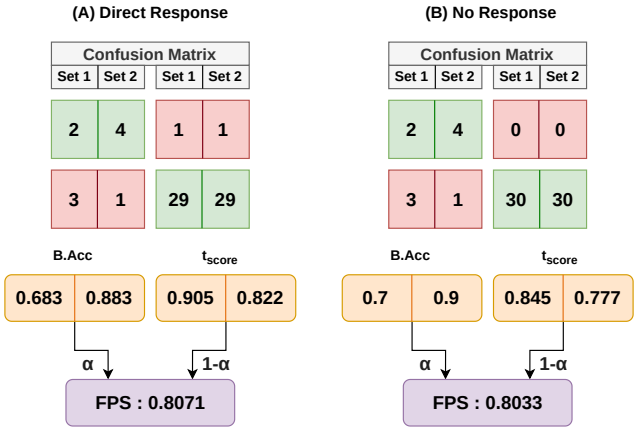

Figure 2: Online stage results. (A) Performance metrics for the direct response case. (B) Performance metrics for the no response case. Each of the 2 cases was evaluated separately to generate the average FPS for each case. The used notation for the confusion matrix is $\left(\begin{smallmatrix} TP & FP \\ FN & TN \end{smallmatrix}\right)$.

6 Conclusion

The IntEr-HRI competition at IJCAI'23 addressed the challenges associated with the asynchronous detection of intrinsic feedback using EEG signals. A specific HRI scenario was designed to evoke ErrPs in the human EEG through robot-induced tactile stimuli. The scientific community was encouraged to develop state-of-the-art machine learning models to detect erroneous behavior of the robotic device continuously. The competition was structured considering real-world application scenarios, and special emphasis was placed on robustness and reliability in the evaluation metrics. To stimulate further research, an open-access EEG and EMG dataset was made available, alongside a detailed account of experimental design and methodology in [Kueper *et al.*, 2024]. We hope that through this competition, we could demonstrate research gaps in the online detection of tactile-based ErrP and direct the scientific community's attention towards further improving assistive technology using psychophysiological data such as EEG data.

Ethical Statement

The studies involving human subjects were approved by the Ethics Committee of the Department of Computer Science and Applied Cognitive Science of the Faculty of Engineering at the University of Duisburg-Essen. The studies were conducted following the local legislation and institutional requirements. The participants provided their written informed consent to participate in the study. Written informed consent was also obtained from the individual(s) for the publication of any potentially identifiable images or data included in this paper or the video.

Acknowledgements

We sincerely thank our colleagues, Judith Bütefür and Julia Habenicht, for their invaluable assistance in conducting the studies. Additionally, we acknowledge the contributions of Tobias Rossol, Marc Tabie, and Jan-Philipp Brettschneider in the mechanical design and setup of the active orthosis device. We would also like to thank Mathias Trampler for his support in developing the live data streaming solution.

References

[Ahkami and Ghassemi, 2021] Bahareh Ahkami and Farnaz Ghassemi. Adding tactile feedback and changing isi to improve bci systems' robustness: An error-related potential study. *Brain topography*, 34(4):467–477, 2021.

[Blankertz, 2004] Benjamin Blankertz. https://www.bbci.de/competition/iii/, 2004. Accessed: 2024-02-19.

[Blankertz, 2008] Benjamin Blankertz. https://www.bbci.de/competition/iv/, 2008. Accessed: 2024-02-19.

[Chavarriaga *et al.*, 2012] Ricardo Chavarriaga, Xavier Perrin, Roland Siegwart, and José del R Millán. Anticipation- and error-related eeg signals during realistic human-machine interaction: A study on visual and tactile feedback. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6723–6726. Ieee, 2012.

[Chavarriaga *et al.*, 2014] Ricardo Chavarriaga, Aleksander Sobolewski, and José del R Millán. Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Frontiers in neuroscience*, page 208, 2014.

[Ehrlich and Cheng, 2019] Stefan K Ehrlich and Gordon Cheng. A feasibility study for validating robot actions using eeg-based error-related potentials. *International Journal of Social Robotics*, 11:271–283, 2019.

[Iturrate *et al.*, 2010] Inaki Iturrate, Luis Montesano, and Javier Minguez. Single trial recognition of error-related potentials during observation of robot operation. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 4181–4184. IEEE, 2010.

[Kim and Kirchner, 2013] Su Kyoung Kim and Elsa Andrea Kirchner. Classifier transferability in the detection of error related potentials from observation to interaction. In *2013 IEEE international conference on systems, man, and cybernetics*, pages 3360–3365. IEEE, 2013.

[Kim and Kirchner, 2015] Su Kyoung Kim and Elsa Andrea Kirchner. Handling few training data: classifier transfer between different types of error-related potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(3):320–332, 2015.

[Kim *et al.*, 2017] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes, and Frank Kirchner. Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction. *Scientific reports*, 7(1):1–16, 2017.

[Kim *et al.*, 2020] Su Kyoung Kim, Elsa Andrea Kirchner, and Frank Kirchner. Flexible online adaptation of learning strategy using eeg-based reinforcement signals in real-world robotic applications. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4885–4891. IEEE, 2020.

[Kothe *et al.*, 2018] Christian Kothe, David Medine, Chadwick Boulay, Matthew Grivich, and Tristan Stenner. Lab streaming layer (lsl). https://github.com/sccn/labstreaminglayer, 2018. Accessed: 2023-08-09.

[Kueper *et al.*, 2024] Niklas Kueper, Kartik Chari, Judith Bütefür, Julia Habenicht, Tobias Rossol, Su Kyoung Kim, Marc Tabie, Frank Kirchner, and Elsa Andrea Kirchner. Eeg and emg dataset for the detection of errors introduced by an active orthosis device. *Frontiers in Human Neuroscience*, 18, 2024.

[Perrin *et al.*, 2008] Xavier Perrin, Ricardo Chavarriaga, Céline Ray, Roland Siegwart, and José del R Millán. A comparative psychophysical and eeg study of different feedback modalities for hri. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 41–48, 2008.

[Roy *et al.*, 2022] Raphaëlle Roy, Marcel Hinss, Ludovic Darmet, Simon Ladouce, Emilie Jahanpour, Bertille Somon, Xiaoqi Xu, Nicolas Drougard, Frederic Dehais, and Fabien Lotte. Retrospective on the first passive brain-computer interface competition on cross-session workload estimation. *Frontiers in Neuroergonomics*, 3, 2022.

[Schiatti *et al.*, 2019] Lucia Schiatti, Giacinto Barresi, Jacopo Tessadori, Louis Charles King, and Leonardo S Mattos. The effect of vibrotactile feedback on errp-based adaptive classification of motor imagery. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6750–6753. IEEE, 2019.

[Tessadori *et al.*, 2017] Jacopo Tessadori, Lucia Schiatti, Giacinto Barresi, and Leonardo S Mattos. Does tactile feedback enhance single-trial detection of error-related eeg potentials? In *2017 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 1417–1422. IEEE, 2017.

[Wu *et al.*, 2022] Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022.