ERR@HRI 2024 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Interactions

Micol Spitale*
ms2871@cam.ac.uk
University of Cambridge
Cambridge, UK

Maria Teresa Parreira
Cornell University
Ithaca, NY, USA

Maia Stiber
Johns Hopkins University
Baltimore, MD, USA

Minja Axelsson
University of Cambridge
Cambridge, UK

Neval Kara†
Cankaya University
Ankara, Turkey

Garima Kankariya‡
Indian Institute of Technology
Delhi, India

Chien-Ming Huang
Johns Hopkins University
Baltimore, MD, USA

Malte Jung
Cornell University
Ithaca, NY, USA

Wendy Ju
Cornell Tech
New York, NY, USA

Hatice Gunes
University of Cambridge
Cambridge, UK

Abstract

Despite the recent advancements in robotics and machine learning (ML), the deployment of autonomous robots in our everyday lives is still an open challenge. This is due to multiple reasons among which are their frequent mistakes, such as interrupting people or having delayed responses, as well as their limited ability to understand human speech, i.e., failure in tasks like transcribing speech to text. These mistakes may disrupt interactions and negatively influence human perception of these robots. To address this problem, robots need to have the ability to detect human-robot interaction (HRI) failures. The ERR@HRI 2024 challenge tackles this by offering a benchmark multimodal dataset of robot failures during human-robot interactions, encouraging researchers to develop and benchmark multimodal machine learning models to detect these failures. We created a dataset featuring multimodal non-verbal interaction data, including facial, speech, and pose features from video clips of interactions with a robotic coach, annotated with labels indicating the presence or absence of robot mistakes, user awkwardness, and interaction ruptures, allowing for the training and evaluation of predictive models. Challenge participants have been invited to submit their multimodal ML models for detection of robot errors, to be evaluated against various performance metrics such as accuracy, precision, recall, F1 score, with and without a margin of error reflecting the time-sensitivity of these metrics. The results of this challenge will help the research field in better understanding the robot failures in human-robot interactions and designing autonomous robots that can mitigate their own errors after successfully detecting them.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; • **Computing methodologies** → *Machine learning algorithms*;

Keywords

Robot Failure, Error Detection, Human-Robot Interaction, Multimodal Interaction, Benchmarking.

ACM Reference Format:
Micol Spitale, Maria Teresa Parreira, Maia Stiber, Minja Axelsson, Neval Kara, Garima Kankariya, Chien-Ming Huang, Malte Jung, Wendy Ju, and Hatice Gunes. 2024. ERR@HRI 2024 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3678957.3689030

1 Introduction

Human-Robot Interaction (HRI) research is currently placing a greater emphasis on the development of autonomous robots that can be deployed in real-world scenarios to understand the implications of integrating such robots in our lives. However, past works [8, 12, 13] have shown that such autonomous robots are often characterised by making mistakes, for example when the robot interrupts people or when the robot takes a very long time to respond. These robot failures may disrupt the interaction and negatively impact the perception of people towards the robot [11]. To overcome this problem, robots should be able to detect HRI failures.

*The author is also affiliated with Politecnico di Milano, Milan, Italy
†Contributed to this work while undertaking a remote visiting studentship at Department of Computer Science and Technology, University of Cambridge.
‡Contributed to this work while undertaking a remote visiting studentship at Department of Computer Science and Technology, University of Cambridge.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3689030

The *ERR@HRI 2024* challenge aims at addressing this issue by providing the community with a benchmark multimodal dataset of robot failures during human-robot interaction. The challenge encourages researchers to benchmark and develop multimodal machine learning-based models designed to identify when failures occur during HRI.

We recruited challenge participants through email advertisements (e.g., ICMI announcements, robotics-worldwide) that included a link to our website[1] where they could fill out the registration form. An EULA agreement, approved by both the DPO and the Departmental Ethics Committee of the University of Cambridge, was shared with the teams who signed up. The signed EULA was then sent to the research office of the University of Cambridge for a final review and approval.

We provided participants with a dataset that includes 1) multimodal non-verbal features (i.e., facial, speech, and pose features) of interaction clips where individuals interact with a robotic coach delivering positive psychology exercises, and 2) binary labels in the form of 'interaction rupture present' (1) or 'interaction rupture absent' (0). These features and labels were to be used to train the predictive models. The dataset was annotated as a time-series with the following labels: robot mistake (e.g., interruption or non-responding, (0) absent, (1) present), user awkwardness (e.g., when the participant feels uncomfortable interacting with the robot without any robot mistakes, (0) absent, (1) present), and interaction ruptures (i.e., either when the user displays some cues of awkwardness towards the robot and/or when the robot makes some mistakes; (0) absent, (1) present). We invited the teams to submit their multimodal ML models for error detection to be evaluated and benchmarked against the pre-determined performance metrics, including accuracy, precision, recall, F1 score, with and without an error margin [4, 8].

1.1 Relevance to Multimodal Interaction

This challenge aims at addressing the problem of detecting robot failures in human-robot interaction, and as such it is extremely relevant to the multimodal interaction community. HRI is multimodal by nature because interactions often involve multiple types of social signaling, such as facial expressions, speech and body language of both humans and robots that, if better understood, can be used as cues to facilitate more natural interactions. ERR@HRI provides a novel multimodal dataset that can be used by participants to develop multimodal machine learning failure detection models. By highlighting the use of multimodal datasets and ML models for detecting failures, the ERR@HRI challenge contributes to advancing the understanding and enhancement of interactions between humans and autonomous robots in real-world settings. The increased interest of the ICMI community in HRI is also evident by the recent contributions published in ICMI proceedings that include 5 papers at ICMI'23 (e.g., [5]) and 2 at ICMI'22 (e.g., [14]) on HRI, and as well as a keynote talk by Prof Maja Mataric at ICMI'23 entitled "A Robot Just for You: Multimodal Personalized Human-Robot Interaction and the Future of Work and Care". The talk focused on multimodal aspects of HRI in healthcare, demonstrating the ICMI community's increasing attention to this field.

[1]https://sites.google.com/cam.ac.uk/err-hri/home

2 Related Work

Past works have shown that robot failures are known to commonly occur during human-robot interactions, and they can negatively impact the user's trust towards the robot. For example, Spitale et al. [10] demonstrated that participants experienced frustration when the robot interrupted them e.g., by erroneously detecting the end of the user's speech while they were still talking, or when the robot exhibited prolonged response times due to internet connectivity issues. Analogously, Kontogiorgos et al. examined human non-verbal behaviour reactions to conversational failures during a cooking instruction class delivered by a Furhat robot [6, 7]. They found that severe errors may decrease users' trust in the robot [7]. However, very few works attempted to address this problem by automatically detecting such failures. Spitale et al. [11] introduced a new multimodal LLM-based system that allows robotic coaches to autonomously adapt to individual's multimodal behaviours (facial valence and speech duration) and detect ruptures while delivering well-being coaching. Bremers et al. [2] used the bystander reaction dataset as input to a deep-learning model, BADNet, to predict failure occurrence without levering multimodal information. These studies represent the first stepping stone toward identifying robot failures during HRI, but they neither focused on benchmarking nor organising a challenge event to enable such comparisons under pre-defined settings and metrics.

The ERR@HRI initiative will provide a unique opportunity for benchmarking not only HRI data but also multimodal machine learning models to detect interaction ruptures, which is fundamental for the success of human-robot interactions. In this first edition, the challenge will focus on using a multimodal dataset collected in a real-world setting where a robotic coach delivered well-being coaching practices to each participant over four weeks. For future editions of the challenge, we plan to focus on additional datasets, such as REACT and Response to Errors in HRI [13], which have already been collected by the co-organizers of this challenge. This will enable a sustained engagement of the research community over the next couple of years and push the state of the art in multimodal robot failure analysis, detection and understanding.

3 The ERR@HRI 2024 Challenge

This section describes the dataset provided, including feature extraction, tasks, and evaluation process.

3.1 Materials

A challenge website was set up[2] with a commitment to be maintained at least for the next 3 years. A GitHub repository[3] has been established along with the official website to guide and support the challenge participants.

3.2 Feature Extraction

We used a dataset collected in a previous study [1, 12], in which we deployed a robotic positive psychology coach at a workplace over four weeks. We involved a total of 43 employees out which 23 gave consent to share their data in processed and aggregated form. The robotic coach conducted four positive psychology exercises over

[2]https://sites.google.com/cam.ac.uk/err-hri/home
[3]https://github.com/ERR-HRI-Challenge/baseline2024

four weeks. Please check the paper [10] for more detail on the study. During the interaction, we collected video recordings (coachee's face and a side view of the interaction) and audio recordings (both the coachee's and robot's speech) using two cameras (a frontal video camera and a lateral GoPro) and a Jabra microphone.

We used off-the-shelf state-of-the-art methods to extract multimodal behavioural features from the audio-visual data collected from the side-view camera as follows:

(1) Facial Features: We used the OpenFace 2.2.0 toolkit to extract the presence and intensity of 17 facial action units (AUs), in a total of 35 facial features per frame, at a rate of 30 fps.
(2) Audio Features: We used the openSMILE toolbox and extracted a total of 25 features, corresponding to the low level descriptors of feature set eGeMAPSv02, using a time window of 0.02 s and at a rate of 100 data points per second.
(3) Pose Features: We used the OpenPose toolbox [3] and extracted the 25-2D body key points per frame to estimate the movement of the torso, hands, arms, and head. The features provided (at 30 fps) do not correspond directly to the features extracted from Openpose, but rather the relational distance and velocity for pairs of spatial body points, in a total of 44 features, corresponding to relational features of 25 body points.

3.3 Labels

The video clips were labelled by 2 annotators using the ELAN video annotation tool. We marked instances of user awkwardness and robot mistakes with binary labels (i.e., 1: present, or 0: absent), marking the time when the displays of user awkwardness or robot mistakes start and end. These labels have been defined in [12] as follows:

- **User Awkwardness (UA)**: The coachee displays behaviours that signal the interaction is awkward — they may look confused, uncertain, distressed or uncomfortable.
- **Robot Mistake (RM)**: The robot makes a mistake such as interrupting or not responding to the coachee, or responding with an utterance that is not appropriate for what the coachee has just said.
- **Interaction Rupture (IR)**: We define an interaction rupture as either the presence of user awkwardness, a robot mistake, or both.

3.4 Sub-challenges

Accordingly, the ERR@HRI 2024 Challenge consists of the following three sub-challenges:

(1) Detection of robot mistakes (e.g., interrupting or not responding to the coachee)
(2) Detection of user awkwardness (e.g., when the coachee feels uncomfortable interacting with the robot without any robot mistakes)
(3) Detection of interaction ruptures (i.e. when the robot makes mistakes as described in 1) or when user displays awkwardness towards the robot described in 2))

3.5 Dataset

The dataset contains data from 23 users, in a total of 89 sessions and 700 minutes of interaction.

ERR@HRI 2024 participants are provided with 4 suggested dataset splits (i.e., subject-independent folds), with no overlapping participant data. Details of the data are provided in Table 1.

3.6 Metrics

This challenge contemplates three binary classification tasks. The metrics used to evaluate model performance are accuracy, precision, recall, f1-score, as well as metrics with a margin of error [4, 8] – for a sample margin of size k, and for a sample i, the model prediction is considered right if $y^i_{pred} \in [y^{i-k}_{pred}, y^{i+k}_{pred}]$

The motivation for considering metrics with a margin of error is due to considerations of real-life settings where effectiveness may still be achieved even if the model is slightly early or delayed in its error detection. Other options for real-use systems could be to use the median or mode of predictions within an interval, among others. Metrics with a margin of error, in this challenge, include accuracy, precision, recall and f1-score.

3.7 Evaluation

Challenge participants were given access to the training and validation sets to develop their ML models. Then, they were asked to submit the developed models and weights, and the organisers have evaluated the submitted models on the test set (the test set was released to the challenge participants without labels one week prior to the submission deadline). Each participating group was allowed to submit their models and results for the test set up to three times. The submitted models and predictions were automatically evaluated and ranked using various performance metrics, under two categories: overall performance and marginal performance. For both tracks, models are ranked based on the combined rankings of accuracy and F1-score (for the marginal track, we use the accuracy and F1-score considering an error margin of 1 sample). Metrics were calculated using the same script provided to participants in the study repository. Challenge participants were also asked to submit a paper describing their model via the EasyChair system, and their works were reviewed by the Technical Program Committee members of the challenge.

4 Challenge Baseline

We have provided a deep-learning multimodal baseline for each of the three tasks, as in [2] and [11] (where we reported results for interaction rupture prediction).

4.1 Training

For baseline models, and following previous work on a similar dataset, we decided to use Recurrent Neural Network models, which can conserve some feature history and are common approaches for time-series classification problems in HRI. Namely, we made use of Long Short-Term Memory networks (LSTMs), Bidirectional-LSTMs (BiLSTMs), Gated Recurrent Unit networks (GRUs), which tend to overfit less than LSTMs in smaller datasets. We used single-layer models, with dropout and a fully-connected layer. We wanted to

Table 1: Dataset and ground truth characteristics. Time per label includes the total amount of time within the dataset labeled as that type of interaction failure. Percentage refers to time per label over total time – which provides a sense of dataset label balance.

Subset	Subjects	Sessions	Total time (s)	Time RM (s)	% RM	Time UA (s)	% UA	Time IR (s)	% IR
Train + Val	18	71	33308	5320	0.16	5182	0.16	8679	0.26
Test	5	18	8048	1399	0.17	1875	0.23	2738	0.34

Table 2: Hyperparameters of best performing models. SL: sequence length. LR: learning rate.

Task	Model	Hyperparameters
RM	GRU	SL=5, Units=128, Dropout=0.2, LR: 0.0001 Activation: softmax, Optimizer: Adam Loss: Categorical Cross-Entropy, Batch size = 2048, Epochs = 500
UA	BiLSTM	SL=5, Units=256, Dropout=0.2, LR: 0.0001 Activation: sigmoid, Optimizer: Adam Loss: Categorical Cross-Entropy, Batch size = 512, Epochs = 200
IR	BiLSTM	SL=5, Units=256, Dropout=0.2, LR: 0.0001 Activation: softmax, Optimizer: Adam Loss: Categorical Cross-Entropy, Batch size = 4096, Epochs = 500

Table 3: Baseline (macro) performances. Margin of error metrics are noted with and e and represent a 1-sample tolerance.

Task	Accuracy	Precision	Recall	F1-Score	Accuracy$_e$	Precision$_e$	Recall$_e$	F1-Score$_e$
RM	0.71349	0.55593	0.54089	0.54184	0.71417	0.55756	0.54219	0.54335
UA	0.73074	0.56358	0.57356	0.56698	0.73207	0.56617	0.57676	0.56978
IR	0.68460	0.55541	0.50268	0.41964	0.68592	0.58794	0.50478	0.42395

provide a standard approach to model development, leaving room for participants to innovate their approaches for detection and classification. For training, we did hyperparameter tuning using test accuracy as the metric to pick the top performing hyperparameters. We used a 3-1 train-validation fold split, with the suggested folds provided in the study repository. For each task and each model architecture, we picked the top 3 performing model hyperparameters– a total of 9 models per task. These models were then trained using cross-validation on the 4 folds and the final model was picked based on the average metrics across all folds. In the end, each of these models was trained on the 4 folds and predictions on the test set were reported to all participants.

4.2 Results & Discussion

The hyperparameters and performance for each model, for each task, are described in Tables 2 and 3. The best performing models have short sequence lengths (5 samples) and the BiLSTM model performed best across two of the subchallenges. The obtained perfomances on the test set are slightly above chance level. While this baseline did not intend to be an exhaustive exploration of model architectures and training methods to generate the best possible performance, it is nonetheless notable that the performance results are not higher. This illustrates previously reported [9] challenges in obtaining generalizable models, due to the high range and diversity of human reactions to failure.

Interestingly, the task of detecting user awkwardness (UA) demonstrates the best overall performance with the highest scores in accuracy, precision, recall, and F1-score among the three tasks. This suggests that models are effective in detecting such expressions for predicting UA. This aligns with our previous analysis[12], which showed that expressions of user awkwardness are characterized by laughter, often marked by the intense activation of cheek raiser (AU6) and lip corner puller (AU12) action units, which correspond to the facial features that were fed into the model. The task of detecting robot mistakes (RM) shows moderate performance, with lower scores than UA but higher than IR, especially in accuracy and F1-score; while the task of interaction ruptures (IR) performs the worst across all metrics, with significantly lower Recall and F1-score compared to UA and RM. The task of detecting robot mistakes may be more difficult because the robot's mistakes caused coachees to limit their self-disclosure and, in turn, express less via their facial or auditory cues, as highlighted in [12]. Overall, these findings suggest varying levels of complexity in detecting user awkwardness and robot mistakes, as evidenced by the models performing the worst at detecting interaction ruptures (IR), which combines elements of both RM and UA. This highlights the importance of tailored approaches to improve model performance for each specific task.

5 Participation and Conclusion

This paper introduced the ERR@HRI 2024 Challenge organised in conjunction with the ACM International Conference on Multimodal Interaction 2024 (ACM ICMI'24), which focuses on detecting robot failures in human-robot interactions. A total of 10 teams from 5 countries registered for this challenge, and 3 teams from 3 European countries submitted their results for benchmarking and evaluation. The submitted models will be ranked under identical conditions using the specified evaluation protocol and metrics. We aim for the challenge data, code, systems, and results from competing teams

to be valuable resources for researchers and practitioners focusing on detecting failures in human-robot interactions. Our future efforts will be directed at continuing to organize ERR@HRI challenge events in conjunction with well-known conferences while introducing new datasets and modalities.

Acknowledgments

Funding: This challenge is possible due to the EPSRC/UKRI grant EP/R030782/1 (ARoEQ) and EP/R511675/1 that supported the HRI studies, and the work of M. Spitale and H. Gunes, that generated the data used in this challenge. M. Spitale's current work involving the organisation of this challenge and the writing of this paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

Data access: Raw data related to this publication cannot be openly released due to anonymity and privacy issues. However, challenge participants who signed the EULA agreement have been granted access to processed data in the form of aggregated feature statistics and models.

References

[1] Minja Axelsson, Micol Spitale, and Hatice Gunes. 2024. " Oh, Sorry, I Think I Interrupted You": Designing Repair Strategies for Robotic Longitudinal Well-being Coaching. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 13–22.

[2] Alexandra Bremers, Maria Teresa Parreira, Xuanyu Fang, Natalie Friedman, Adolfo Ramirez-Aristizabal, Alexandria Pabst, Mirjana Spasojevic, Michael Kuniavsky, and Wendy Ju. 2023. The Bystander Affect Detection (BAD) Dataset for Failure Detection in HRI. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11443–11450.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[4] IA de Kok and Dirk KJ Heylen. 2012. A survey on evaluation metrics for backchannel prediction models. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog, Stevenson, Washington, USA: Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. University of Texas, 15–18.

[5] Apostolos Kalatzis, Saidur Rahman, Vishnunarayan Girishan Prabhu, Laura Stanley, and Mike Wittie. 2023. A Multimodal Approach to Investigate the Role of Cognitive Workload and User Interfaces in Human-robot Collaboration. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 5–14.

[6] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[7] Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 112–120.

[8] Maria Teresa Parreira, Sarah Gillet, and Iolanda Leite. 2023. Robot Duck Debugging: Can Attentive Listening Improve Problem Solving?. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 527–536.

[9] Maria Teresa Parreira, Sukruth Gowdru Lingaraju, Adolfo Ramirez-Aristizabal, Manaswi Saha, Michael Kuniavsky, and Wendy Ju. 2024. A Study on Domain Generalization for Failure Detection through Human Reactions in HRI. arXiv:2403.06315 [cs.RO] https://arxiv.org/abs/2403.06315

[10] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. Robotic mental well-being coaches for the workplace: An in-the-wild study on form. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 301–310.

[11] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. VITA: A Multi-modal LLM-based System for Longitudinal, Autonomous, and Adaptive Robotic Mental Well-being Coaching. *arXiv preprint arXiv:2312.09740* (2023).

[12] Micol Spitale, Minja Axelsson, Neval Kara, and Hatice Gunes. 2023. Longitudinal evolution of coachees' behavioural responses to interaction ruptures in robotic positive psychology coaching. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 315–322.

[13] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[14] Xiang Zhi Tan, Elizabeth Jeanne Carter, Prithu Pareek, and Aaron Steinfeld. 2022. Group formation in multi-robot human interaction during service scenarios. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 159–169.