PRISCA at ERR@HRI 2024: Multimodal Representation Learning for Detecting Interaction Ruptures in HRI

Pradip Pramanick
pradip.pramanick@unina.it
Interdepartmental Center for Advances in Robotic Surgery,
University of Naples Federico II
Naples, Italy

Silvia Rossi
silvia.rossi@unina.it
Department of Electrical Engineering and Information
Technologies, University of Naples Federico II
Naples, Italy

Abstract

Interaction ruptures in human-robot interaction (HRI) refer to scenarios when seamless interactions are disrupted. Such ruptures can be directly observed by the robot at times, e.g., not responding to a human utterance. However, often the ruptures could be more passive and subtle and require an analysis of the human's behavior. In this work, we focus on detecting such ruptures by analyzing multimodal information in a face-to-face interaction setting. More specifically, this paper describes the PRISCA team's participation in the ERR@HRI Challenge 2024, which was recently proposed to benchmark multimodal learning approaches to interaction rupture detection in HRI. Central to our approach is a feature-fusion strategy for multimodal representation learning, where we train a neural network with separate recurrent layers that act as temporal encoders to learn modality-specific representations. Our approach was ranked 3rd in the ERR@HRI challenge. We present detailed experimentation on the released dataset from the challenge and a thorough analysis of the results. We further discuss the limitations of current approaches and implications for future works. Code will be made available at https://github.com/pradippramanick/prisca-errhri/.

CCS Concepts

• **Computing methodologies** → **Machine learning approaches**; • **Computer systems organization** → *Robotics*; • **Human-centered computing** → *Empirical studies in HCI*.

Keywords

Human-Robot Interaction, Robot Failure, Multimodal Learning, Feature Fusion, Affective Computing

ACM Reference Format:
Pradip Pramanick and Silvia Rossi. 2024. PRISCA at ERR@HRI 2024: Multimodal Representation Learning for Detecting Interaction Ruptures in HRI. 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/10.1145/3678957.3688387

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

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3688387

1 Introduction

Enabling seamless human-robot interaction (HRI) requires robust methods for processing information from various modalities in real time. In practice, this is quite challenging even if the interaction is limited to one modality, e.g., speech, where the quality of transcriptions can be affected by several factors and lead to incorrect interpretation of the human utterance [13, 18]. Further, without accurate turn-taking predictions, spoken dialog systems used in HRI are prone to timing errors [7]. For example, incorrect turn-taking can lead to a robot interrupting a person while they are still talking [15].

Detection of such interaction ruptures, when a seamless interaction between a human and a robot is disrupted is an important problem in HRI [16]. In most cases, the effects of interaction failures are negative, e.g., leading to reduced trust [10]. Whereas, being aware of the presence or even a high likelihood of interaction ruptures enables robots to employ strategies to repair the interaction. For example, language understanding errors and ambiguity can be repaired using dialog [14, 18]. Speech recognition errors can be minimized using additional information from other modalities, e.g., vision, when errors are expected [12]. Further, information from multiple modalities may contain robust indicators of interaction ruptures, compared to a single modality. Most prior works have focused on analyzing human reactions for automatic detection of such ruptures [4, 15, 18]. However, the lack of large-scale benchmark datasets with sufficient diversity and well-defined annotations remains a challenge. Recently, the ERR@HRI 2024 challenge was proposed to benchmark methods to detect interaction failures during human-robot interactions [16]. Subsequently, a multimodal dataset consisting of preprocessed features comprising facial expressions, speech, and relative pose from interactions with a robotic coach was released. The facial expressions are represented as Action Units (AU) activation and the corresponding intensities, extracted using OpenFace [1]. For speech, the dataset contains the eGeMAPSv02 feature set, extracted using openSMILE [8]. For the pose, the features represent the relative distance and velocity between keypoints extracted using Openpose [5]. In this challenge, an interaction rupture is defined as the presence of either a robot making a mistake or a human user showing awkwardness during the interaction. The objective is to detect binary interaction-rupture labels, given a sequence of the aforementioned features.

This work describes our participation in the challenge, where we develop an approach to learning modality-specific features on the released dataset. We apply this method to all three sub-problems in the challenge, namely detecting i) User Awkwardness (UA), ii) Robot Mistake (RM), and iii) Interaction Rupture (IR). Our experiments

show that our approach to learning modality-specific representations outperforms a baseline where an otherwise similar neural network learns from combined input vectors of all the modalities. We also highlight the issue of class imbalance in the dataset and find that our approach handles class imbalance better than the baseline. We also contribute to the discussion on the limitations of the current data collection approaches towards autonomous detection of interaction failures and suggest future directions.

2 Related Works

While many prior studies have examined interaction failures in human-robot interaction HRI [19], they focus on understanding the effects of such failures rather than developing methods for their automatic detection. Since this challenge aims to detect interaction ruptures by observing human reactions, we restrict our discussion to methods that process this type of information.

2.1 Datasets

Several datasets aim to collect implicit human feedback in the form of recorded reactions to robot's actions in various domains. The reactions are usually collected while a human directly observes the robot or watches recorded videos. In the EMPATHIC dataset [6], Cui et al. collect human reactions from observing videos in two domains - a simulated taxi game and an object sorting task with a physical robot. The *Response-to-Errors* dataset [18] contains facial responses in the form of facial action units (AUs) to robot mistakes during three physical human-robot interaction tasks - collaborative assembly, collaborative cooking, and programming by demonstration. Zhang et al. collect participants' facial expressions during interaction in virtual reality with a simulated robot for a navigation task [20]. Although these datasets are not directly aimed at detecting interaction failures, they highlight the prevalence of using implicit social signals to recognize robot mistakes. In REACT [4], the authors collect several implicit communicative signals such as head pose, and gaze, along with facial action units during collaborative gameplay and photography tasks. The dataset in ERR@HRI [16] also falls into a similar category. However, this dataset is explicitly annotated with the objective of interaction rupture detection.

Apart from collecting social signals as feedback, other datasets employ more explicit feedback metrics. Yu et al. collect a dataset of scalar ratings from human observers who watch a robotic arm perform several manipulation tasks. Overall, except for REACT [4], much of the existing datasets are collected from relatively shorter interactions. As such, there is not sufficient evidence of whether the models trained on these datasets can be applied to longer interaction sessions as well. We discuss this further in Section 5.

2.2 Methods

There are only a handful of prior works that specifically address detecting robot failures and disruptions by observing human reactions. Cui et al. [6] use fixed-length time windows to aggregate head pose and AU features extracted using OpenFace[1]. However, instead of explicit temporal modeling, the aggregated features are simply flattened and passed through a Multi-Layer Perception (MLP) for mapping implicit human reactions to an explicit reward model. Stiber et al. uses a two-layer neural network to detect robot mistakes from AU features [17]. Instead of temporal modeling at the input or feature level, their approach uses sliding window-based post-processing methods to filter spurious detections. Bremers et al. experiment with variants of 2D convolutional neural networks (CNNs) to detect machine failures from human reactions recorded through webcams [3]. CNNs are known to be effective for image understanding, although within the scope of our problem, we assume pre-processed features as inputs. Other experiments include [2], where Ben-Yousef et al. use logistic regression to classify engagement breakdowns from multimodal data. [20] compares CNN, graph-based, and transformer-based neural network architectures for jointly processing the robot's navigation-related and observer's facial features for error detection.

The baseline provided in ERR@HRI also follows a similar windowing approach [16], albeit with explicit temporal modeling with Gated Recurrent Units (GRUs). Given that numerous prior works have utilized Long Short-Term Memory Networks (LSTMs) for sequence modeling in several domains, we also base our temporal modeling approach using LSTM. However, in contrast to the baseline, we propose separate LSTM encoders for distinct temporal modeling of the different modalities.

3 Approach

To capture the temporal variations in the features from the three modalities, we use recurrent neural networks to encode the features. Specifically, we use LSTM layers that encode short-term temporal variations in human reactions. The decision to use separate modality encoders is based on our hypothesis that the information across different modalities may not exhibit strict temporal synchronicity. For instance, an awkward facial expression might be succeeded by a noticeable change in body posture, rather than occurring concurrently[1]. Thus, events that are slightly separated in time in different modalities can be a part of the same interaction rupture. Further, having separate temporal encoders allows us to compensate for approximately synchronized data. This is because the pre-processed features in the challenge dataset are collected at separate frequencies, i.e., AU and pose features are collected at 30 frames per second (fps); while the speech features are collected at 100 fps. Our empirical evaluation further supports this decision, as described in Section 4.

Figure 1 shows the proposed network architecture. More formally, given a sequence of AU features $x_0^A, \ldots, x_n^A$, pose features $x_0^P, \ldots, x_n^P$, and speech features $x_0^S, \ldots, x_n^S$, each having a sequence length of n, we compute the final state of the LSTMs in the forward and backward directions, obtaining three hidden representations $\mathbf{h}^A, \mathbf{h}^P$, and $\mathbf{h}^S$. These are concatenated to obtain a multimodal representation $\mathbf{z}$ which is passed through a fully connected layer with a sigmoid activation for classification. We train the network to classify a binary label $l \in \{0, 1\}$. We further introduce the following regularizers to stabilize the training. We apply three regularizers to the LSTMs, a kernel regularizer, a recurrent regularizer, and a bias regularizer. Further, we add a dropout layer before $\mathbf{z}$ is passed to the *FNN*. Our network architecture remains the same for all three

[1]We could not verify if such events exist in this dataset without the original videos.

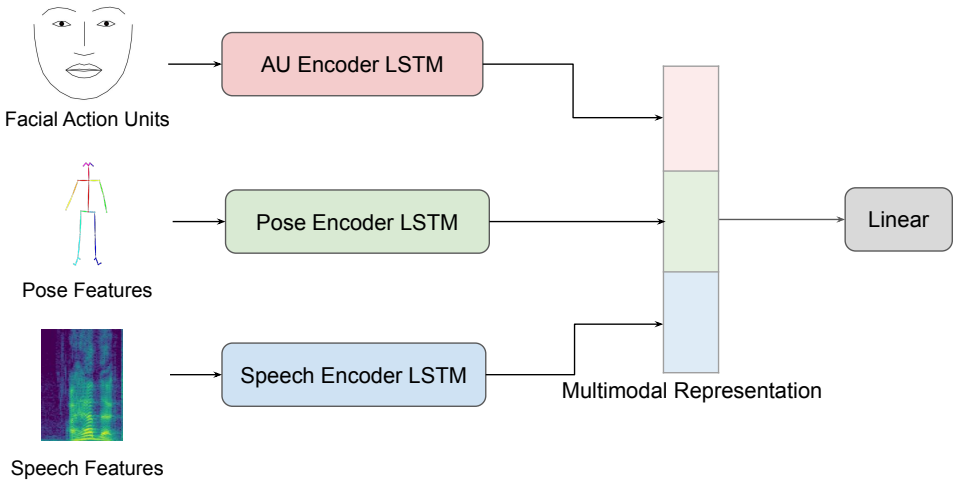

Figure 1: Our multimodal learning framework that was submitted to ERR@HRI 2024.

tasks - UA, RM, and IR. We describe the operations in the following.

$$\mathbf{h}^A = [\overrightarrow{LSTM}(x_0^A, \ldots, x_n^A); \overleftarrow{LSTM}(x_n^A, \ldots, x_0^A)]$$
$$\mathbf{h}^P = [\overrightarrow{LSTM}(x_0^P, \ldots, x_n^P); \overleftarrow{LSTM}(x_n^P, \ldots, x_0^P)]$$
$$\mathbf{h}^S = [\overrightarrow{LSTM}(x_0^S, \ldots, x_n^S); \overleftarrow{LSTM}(x_n^S, \ldots, x_0^S)]$$
$$\mathbf{z} = [\mathbf{h}^A; \mathbf{h}^P; \mathbf{h}^S]$$
$$l = \underset{l \in \{0,1\}}{\operatorname{argmax}} P(l|FNN(\mathbf{z})).$$

4 Experiments

4.1 Data

We experiment only with the ERR@HRI dataset [16], which contains features extracted from 89 interaction sessions with a robot and 23 participants. The dataset is annotated with binary labels for three tasks. Similar to the released baselines, we use the same participant distribution, the same set of features, and the same sequence length (5). We do not perform any feature normalization. We experiment with a total of 762624 training and 235610 validation examples. We observe a severe class imbalance in the dataset. The imbalance is most significant for UA and RM tasks, where only 16% of examples are marked as Awkward and Mistake, respectively. For IR, 23% of the examples are labeled as positive, i.e., a presence of interaction rupture.

4.2 Training

We train the models with a cross-entropy loss and Adam optimizer. We use early stopping when the validation loss does not reduce over three consecutive epochs. Further, we reduce the learning rate by a factor of 0.5 when the validation loss does not reduce over two epochs. Instead of using the usual approach of saving model checkpoints based on accuracy, we monitor the macro F1 on the validation set due to the class imbalance. We could not perform extensive hyper-parameter optimization experiments due to limited time. However, we experiment with slight variations of the batch size, learning rate, and the number of LSTM units heuristically. Table 1 shows the hyperparameters used for training the models submitted in the challenge. Apart from these, we used the same hyperparameters in all three tasks - Dropout rate 0.5, learning rate 0.0001, and regularization rate 0.01. We used the same seed (42) as the released baselines from the challenge. On a Quadro RTX 5000 GPU, training the 768-unit model takes about a minute per epoch.

Hyperparameter	UA	RM	IR
LSTM units	768	768	128
Batch size	512	512	256

Table 1: Training hyperparameters of the submitted models.

4.3 Results

As the test dataset was released without labels, we could not perform a detailed performance analysis on the test data. Instead, we report experiment results on the validation data. For completeness, we also report the official results on the test dataset in Table 2. On the test set, our models perform somewhat similarly for the three tasks, with slight improvements on the time-tolerant metrics. At the time of writing, our models rank 3rd in the challenge results.

In the following, we compare our approach to learning modality-specific representations with a baseline that uses a single encoder for all three modalities. This closely mimics the official baselines released as a part of the challenge [16]. We report the comparison results in Table 3. Overall, our first observation indicates that the severe class imbalance, as discussed in Section 4.1, leads to a highly imbalanced learning performance. Specifically, due to the significantly lower frequency of the positive classes (e.g., *Awkward*), the models can consistently reduce the total loss and thus increase accuracy, without properly learning to classify the positive examples. We posit that the accuracy metric can be quite misleading in this scenario, and thus focus on analyzing the F1 scores. Our

Task	Accuracy	Precision	Recall	Macro F1	Accuracy^{tolerant}	Precision^{tolerant}	Recall^{tolerant}	Macro F1^{tolerant}
UA	0.76	0.54	0.51	0.45	0.77	0.67	0.51	0.46
RM	0.82	0.53	0.5	0.46	0.82	0.53	0.5	0.46
IR	0.68	0.53	0.5	0.42	0.69	0.77	0.51	0.42

Table 2: Performance of the submitted models on the unreleased test set.

Task	Model	Label	Precision	Recall	F1
UA	Baseline	Non-Awkward	0.84	1.00	0.91
		Awkward	0.09	0.01	0.01
		Macro avg.	0.47	0.5	0.46
	Ours	Non-Awkward	**0.85**	0.99	0.91
		Awkward	**0.34**	**0.18**	**0.23**
		Macro avg.	**0.59**	**0.58**	**0.57**
RM	Baseline	No-Mistake	0.84	1.00	0.92
		Mistake	0.18	0.08	0.11
		Macro avg.	0.51	0.54	0.51
	Ours	No-Mistake	0.84	1.00	0.92
		Mistake	**0.31**	0.02	0.03
		Macro avg.	**0.58**	0.51	0.47
IR	Baseline	No-Rupture	0.77	1.00	0.87
		Rupture	0.27	0.01	0.01
		Macro avg.	0.52	0.50	0.44
	Ours	No-Rupture	0.77	1.00	0.87
		Rupture	**0.30**	**0.26**	**0.28**
		Macro avg.	**0.53**	**0.63**	**0.57**

Table 3: Performance on the validation set. We highlight cases where our model outperforms the baseline with boldface.

comparison with the baselines suggests that for both UA and IR tasks, our approach of separate modality encoders outperforms the approach of aggregated input features used in the baseline. For the UA task, we find a significant improvement in the macro F1 score (+0.11 points) in our approach. More importantly, our model learns a much better classifier for the low-frequency class, i.e., a significant improvement of 0.22 points in F1 for the *Awkward* class. We observe a similar trend in the IR class as well, where our model outperforms the baseline by 0.13 points in the macro F1 score, and by 0.27 points in the *Rupture* class.

However, for the RM task, our F1 scores are lower than the baseline, even though we improve on the macro-precision and the precision on the *Mistake* class. We suspect this is due to a limitation of our approach in data modeling, rather than the network design. More specifically, we did not perform speaker diarization on the speech features to separate the robot's and the human's audio features. Since the RM task is primarily defined by the robot's non-response and delayed responses, the speech features may not have been very informative without diarization. Somewhat surprisingly, our models performed better on the RM task in the test set, compared to UA and IR tasks (Table 2). Whereas, our experiments with the validation set suggest otherwise. In the future, we shall investigate further to understand this contradiction better. However, this may have been due to class imbalance and our checkpointing strategy to optimize performance on the validation set.

5 Discussion & Future Work

A crucial limitation of the contemporary data-driven approaches to interaction rupture detection in HRI is two over-generalized assumptions. One is to assume that given similar stimuli of interaction failures, people will behave similarly. Thus the data-driven methods may learn to associate common behavioral cues with failure instances. However, prior HRI studies suggest that personalization of such models may be needed [9]. Secondly, the current methods, datasets, and annotation schemes assume that the users of a robot will express themselves similarly over a long period. Again, several prior experiments seem to suggest otherwise. For example, Candon et. al find that users get less expressive over time during long-term interaction with a robot [4]. Therefore, an interaction failure detection model trained on more expressive data may not be accurate in the long term. We posit that this assumption should be relaxed and data collection and annotation strategies should consider this.

Another direction of future work involves processing raw data instead of pre-processed features. The reason is threefold. Firstly, since the features are extracted using statistical models (e.g., OpenFace and OpenPose), there is a significant chance of error propagation from noisy feature extraction by the models. Second, by learning on pre-processed features that are more abstract, the models may not be able to access subtle changes that would otherwise be present in the raw data. Finally, learning on raw data can take advantage of the robust vision and audio encoders by transfer learning, which can further improve generalization with small training sets. However, we also acknowledge the privacy concerns of using raw video recordings as training data.

For improving classification performance, several strategies can be explored to mitigate the effects of class imbalance, such as sampling and focal loss [11]. Since the UA, RM, and IR tasks are quite similar, and IR is essentially a logical OR operation of UA and RM labels, multi-task learning and post-processing based on logical reasoning could be worth exploring in the future.

6 Conclusion

In this work, we present experiments on the ERR@HRI dataset for autonomous interaction rupture detection in human-robot interaction. We propose an approach to learn multimodal representations by employing separate recurrent neural network (LSTM) layers to encode features from three different modalities. We compare our approach with a baseline that uses a single LSTM layer to encode concatenated features from the same three modalities. Our experiments suggest that our approach leads to better classification performance than the baseline in most cases. Further, we observe a better resistance to class imbalance in the dataset using our approach. We further point out limitations in the current data-driven approaches for this problem and discuss several future directions.

Acknowledgments

This work has been partially supported by the European Union's Horizon Europe research and innovation program under the TRAIL project, Marie Skłodowska-Curie grant agreement No 101072488, and by the Italian Ministry for Universities and Research (MUR) with the PNRR Project FAIR (Future Artificial Intelligence Research) PE0000013.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[2] Atef Ben-Youssef, Chloé Clavel, and Slim Essid. 2019. Early detection of user engagement breakdown in spontaneous human-humanoid interaction. *IEEE Transactions on Affective Computing* 12, 3 (2019), 776–787.

[3] 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.

[4] Kate Candon, Nicholas C Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. 2024. REACT: Two Datasets for Analyzing Both Human Reactions and Evaluative Feedback to Robots Over Time. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 885–889.

[5] 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).

[6] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*. PMLR, 604–626.

[7] Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*. 5190–5194. https://doi.org/10.21437/Interspeech.2022-10955

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[9] Norina Gasteiger, Mehdi Hellou, and Ho Seok Ahn. 2023. Factors for personalization and localization to optimize human–robot interaction: A literature review. *International Journal of Social Robotics* 15, 4 (2023), 689–701.

[10] 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.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[12] Pradip Pramanick and Chayan Sarkar. 2022. Can Visual Context Improve Automatic Speech Recognition for an Embodied Agent?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1946–1957. https://doi.org/10.18653/v1/2022.emnlp-main.127

[13] Pradip Pramanick and Chayan Sarkar. 2023. Utilizing Prior Knowledge to Improve Automatic Speech Recognition in Human-Robot Interactive Scenarios. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 471–475. https://doi.org/10.1145/3568294.3580129

[14] Pradip Pramanick, Chayan Sarkar, Sayan Paul, Ruddra dev Roychoudhury, and Brojeshwar Bhowmick. 2022. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters* 7, 4 (2022), 10826–10833.

[15] 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.

[16] 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. *arXiv preprint arXiv:2407.06094* (2024).

[17] Maia Stiber, Russell Taylor, and Chien-Ming Huang. 2022. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 676–683.

[18] 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.

[19] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.

[20] Qiping Zhang, Nathan Tsoi, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, and Marynel Vázquez. 2023. Towards Inferring Users' Impressions of Robot Performance in Navigation Scenarios. *arXiv preprint arXiv:2310.11590* (2023).