Why Robots Are Bad at Detecting Their Mistakes: Limitations of Miscommunication Detection in Human-Robot Dialogue

Ruben Janssens[1], Jens De Bock[1], Sofie Labat[2], Eva Verhelst[1], Veronique Hoste[2], Tony Belpaeme[1]

***Abstract*— Detecting miscommunication in human-robot interaction is a critical function for maintaining user engagement and trust. While humans effortlessly detect communication errors in conversations through both verbal and non-verbal cues, robots face significant challenges in interpreting non-verbal feedback, despite advances in computer vision for recognizing affective expressions. This research evaluates the effectiveness of machine learning models in detecting miscommunications in robot dialogue. Using a multi-modal dataset of 240 human-robot conversations, where four distinct types of conversational failures were systematically introduced, we assess the performance of state-of-the-art computer vision models. After each conversational turn, users provided feedback on whether they perceived an error, enabling an analysis of the models' ability to accurately detect robot mistakes. Despite using state-of-the-art models, the performance barely exceeds random chance in identifying miscommunication, while on a dataset with more expressive emotional content, they successfully identified confused states. To explore the underlying cause, we asked human raters to do the same. They could also only identify around half of the induced miscommunications, similarly to our model. These results uncover a fundamental limitation in identifying robot miscommunications in dialogue: even when users perceive the induced miscommunication as such, they often do not communicate this to their robotic conversation partner. This knowledge can shape expectations of the performance of computer vision models and can help researchers to design better human-robot conversations by deliberately eliciting feedback where needed.**

I. Introduction

In dialogue, individuals do more than merely interpret their interlocutors' words; they also seek feedback regarding the ongoing interaction. This process, as part of grounding the conversation, involves communicative partners signalling their understanding and agreement on the information exchanged. Through grounding, interlocutors adjust their contributions to ensure that the intended meaning has been successfully conveyed [1]. This capability is also crucial for socially interactive robots, as detecting miscommunications is key to handling the complexity and inherent imperfections in human-robot dialogue.

Detecting feedback on miscommunications is a multi-modal affair: people express their feedback through speech, but also through gaze, head gestures, body pose, and in particular facial expressions [2], [3]. Recent advances in computer vision, particularly through the application of Deep Learning, have significantly enhanced the performance of machine learning models for facial expression recognition [4]. However, it remains unclear how these advances transfer to real-world scenarios.

Fig. 1: An anonymised example frame from the REPAIR-Corpus dataset [5], showing the user from the robot's perspective during their conversation.

To build robotic systems capable of detecting mistakes and misunderstandings in dialogue, we set out to investigate the performance of machine learning models at detecting such communication breakdowns in real human-robot conversations. For this, we read the non-verbal signals of the human interlocutor during an interaction with a robot, and predict if and when they struggle to comprehend the robot.

Whereas previous research predominantly examines task-oriented collaborative interactions, in which humans and robots cooperate to complete a (physical) task with guidance from the robot [6]–[8], this study explores an educational context in which the robot assumes the role of an instructor, explaining a topic to the user while periodically posing questions to maintain engagement. Within this dialogue, four distinct types of intentional errors are introduced.

Previous studies have relied on external annotators to classify feedback as either positive or negative or solely relied on the intended effect of communication errors that were explicitly introduced [6]–[9]. In contrast, our approach seeks to determine whether users perceive the robot's speech acts as erroneous. To achieve this, we collected real-time feedback directly from users during conversations. This turn-by-turn feedback serves as the ground truth for training our machine learning models. Figure 1 shows an example frame from the dataset, showing the user from the perspective of the robot.

In this paper, we describe in detail the construction of a system that detects if a user felt the robot caused a

This research received funding from the Flemish Government (AI Research Program 2) and from the Research Foundation Flanders (FWO Vlaanderen, 1S96324N and 1S50425N).

[1]IDLab-AIRO, Ghent University–imec, Ghent, Belgium, contact: ruben.janssens@ugent.be

[2]LT3, Ghent University, Ghent, Belgium

miscommunication. The system operates on videos taken from the robot's perspective, showing the torso and head of the interlocutor, as both facial expressions and head movements were identified as strong predictors of negative feedback in earlier studies [9], [10]. The system contains two main components: a model to extract salient fragments from the interaction videos, and another model to classify those fragments as indicative of a misunderstanding or not.

Upon observing that the system's performance significantly underperforms relative to initial expectations, we conducted further validation by training and evaluating the model using a dataset of short videos featuring highly expressive emotional content. Additionally, we explored human performance on the same task by instructing participants to label the same videos used in the model's evaluation.

In conclusion, this study critically examines the successes reported in previous affect recognition and robot failure detection research. We apply state-of-the-art techniques for processing facial expressions for feedback detection to a social human-robot dialogue in an educational setting and provide insights into why this task is so challenging by leveraging (1) explicit feedback directly from the user in the interaction as a ground truth and (2) investigating human performance when acting as external annotators.

II. Related Work

As we aim to bridge the gap between research on affect recognition and adaptive human-robot dialogue, we begin by discussing related work in the theoretical foundations of affect, with a particular focus on emotion theories. We then explore recent advances in automatic affect recognition, and finally review related work in detecting mistakes in human-robot interaction.

A. Confusion as an affective-cognitive state

Affect is a broad term that encompasses various constructs such as sentiment, feeling, mood, and emotion. These constructs are central to understanding human experience and behaviour, and they have been studied across multiple disciplines such as psychology, cognitive science, and affective computing. Among these, emotions have received particular attention due to their observable impact on decision-making, communication, and social interaction. Despite this focus, researchers have yet to converge on a single, universally accepted definition of what constitutes an emotion. Emotion research is typically categorized into three major theoretical frameworks: discrete or basic emotion theories [11], [12], constructionist theories [13], [14], and appraisal theories [15], [16].

Discrete or basic emotion theories argue that humans have a set of biologically ingrained, universal emotions that are distinctly expressed through facial expressions. To apply this theory in data analysis, researchers use a fixed taxonomy of emotion labels to annotate texts, audio or videos [17]–[19]. Constructionist theories, on the other hand, suggest that emotions are learned constructs that are formed through a combination of core affective features (e.g., degree of pleasantness, level of arousal). For analytical purposes, emotions are measured in a constructionist way along the dimensions of arousal, valence, and dominance [20]–[22]. Finally, appraisal theories focus on how we cognitively evaluate or "appraise" salient stimuli along a range of checks (e.g., goal relevance, urgency, novelty). Although appraisals have recently found their introduction in the field of natural language processing (NLP), its application in the field of facial emotion recognition remains limited [23]–[25].

Building on these theoretical foundations, it becomes evident that not all affective phenomena fit neatly within the boundaries of emotion as traditionally defined. One such phenomenon is *confusion*, which we examine in this study as an affective-cognitive state that arises in natural, non-acted interactions. Confusion is often characterized by a blend of cognitive dissonance, uncertainty, and emotional arousal, making it a hybrid state that challenges the discrete categorization of emotions—though appraisal theories are capable of accounting for this nuanced state [26]. It typically emerges in response to complex or ambiguous stimuli, particularly in learning or problem-solving contexts, where individuals struggle to reconcile conflicting information or expectations [27]. Rozin and Cohen [28] observed that college students frequently found confusion, marked by eye-region facial movements, as a primary descriptor when interpreting facial expressions of emotion. Other researchers also remarked that the expression of confusion shares several characteristics commonly attributed to emotion, such as a consistent appraisal pattern and distinct facial cues [27], [29].

B. Affect recognition using machine learning

With advances in deep learning for computer vision, there has been a growing interest in automatically detecting emotional expressions from facial expressions, known as facial expression recognition (FER). Most FER research remains built on discrete emotion theories (i.e. classifying facial expressions into distinct classes as happy, sad, anger, surprise, fear, disgust, contempt, or neutral) [30]–[33]. FER systems are usually deep neural networks, built by taking models that are pre-trained for tasks such as detecting facial landmarks, and fine-tuning them on images of faces labelled for one of the basic emotions. Such models can achieve an accuracy of up to 96.8% [32].

However, such models are trained and evaluated in a rather black-and-white way: either the expression is indicative of one of the basic emotions, or it is neutral. This optimises them to learn extreme, prototypical expressions of these basic emotions [32]. However, as Kappas et al. note [34], such "full-blown patterns" are rarely present in the real world, and they are not a reliable indicator of emotions. This might limit the applicability of facial expression recognition models in real-world settings, such as human-robot interaction.

C. Mistake detection in human-robot interaction

Studying user responses to robot failures and automatically detecting them has been a topic of research in human-robot interaction for at least a decade [3]. Most of the research in

this direction focuses on some type of collaborative robots, where the user tries to accomplish a task together with the robot or with the robot's guidance.

For example, Trung et al. [8] investigate automatically detecting a Nao robot's failures while it is giving instructions to a user who is completing a LEGO construction task. By examining only the user's head and shoulder movements, their model can detect robot technical failures. However, it is less effective when the robot violates social norms or when it encounters unfamiliar users. Stiber et al. [7], [35] also look at a collaborative robot: their study employs a robot arm that picks up objects in response to user commands. They capture users' implicit feedback through facial expressions by employing a pre-trained model that detects Facial Action Units (FAUs), which are specific movements in the face. A neural network then classifies these facial movements to determine whether they are indicative of robot failure.

Kontogiorgos et al. [10] identify conversational failings of a Furhat robot during an assembly task by analyzing head movements, gaze, and speech signals. They find that external annotators can most reliably detect robot mistakes by identifying when the user is in a state of *confusion*. Comparing this task with another context, where the user and robot engage in a negotiation task [3], they suggest that non-verbal features indicating affect and emotion are more generalisable across interaction contexts, whereas verbal feedback is highly context-specific.

Perhaps most similarly to our study, Axelsson et al. [9] investigate multimodal user feedback to a robot that presents paintings, with external annotators labelling the user's feedback as positive, negative, or neutral. The robot, controlled through a Wizard-of-Oz setup, would occasionally misspeak (replacing part of the robot's speech by other words or silence). The authors find that facial expressions, backchannels, and speech tone are most indicative of negative feedback, and used Random Forests and LSTMs to automatically classify the feedback. In that study, facial expressions were manually annotated for specific movements such as frowning. In a follow-up study with an autonomous, real-time robot, the authors did not use automatic facial expression detection, instead using only head movements and speech features [36].

Notably, none of these studies collect explicit feedback from the user as ground truth for training or evaluation. Rather, they solely rely on external annotations or explicitly introduced failures and do not examine whether the user perceived them as failures—Trung et al. [8] have already found some cases where users do not perceive the robot's designed social norm violations as such.

Our study builds upon previous research by focussing on facial expressions, which have been shown to be a crucial feedback modality. We investigate this within a social, dialogue-only human-robot interaction, rather than a physical collaboration task, examining four distinct failure types and uniquely collecting ground truth from the users.

III. Datasets

Two datasets are used in this research: a dataset of natural user feedback in human-robot dialogues and a dataset containing videos with emotionally expressive content, sourced from an online repository.

A. Human-robot conversation dataset

The REPAIR-Corpus dataset is an in-house corpus that captures the natural reactions and social cues of participants during human-robot interactions. It is available through Zenodo [5]. The dataset consists of 240 videos featuring 40 participants, each of whom engaged in six conversations with the social robot Furhat [37]. The videos show the user from the robot's perspective and were recorded using an external webcam placed in front of the robot. Participants were told that they would be interacting with six different systems, while the conversations were in reality fully scripted and grounded in cooking recipes, with each recipe representing a different "system". The scripts incorporated four types of robot mistakes: interruptions, oversharing information, undersharing information, and irrelevant comments [38]–[40]. The order of these mistakes was randomized across the different recipes. Additionally, the order in which participants encountered the recipes was randomized. To address these communication errors, the experiment introduced repair strategies that differed in formality (formal vs. informal), including apologies, promises, and explanations [41]. However, participants often found these strategies to be somewhat unnatural, especially when they disrupted the flow of the conversation.

During the conversations, participants were asked to press a binary controller button after their reply to each robot utterance to indicate whether they perceived the utterance as a mistake or inappropriate for flow of the conversation. In our experiments, we use this feedback as the ground truth for whether the user experienced a miscommunication due to a robot mistake. The dataset includes a substantial amount of data capturing emotional expressions, thus distinguishing itself from many other existing datasets that rely on acted emotion expressions [19], [20], [42], which are often used in facial expression recognition research.

To use this dataset in our experiments, we first process the dataset into short video fragments which each contain one exchange, consisting of a robot utterance and user response. This collection of 2600 fragments, 27.2% of which were labelled by the user as containing a mistake, was further divided into a training, validation, and test set, which respectively contain 67.5%, 22.5%, and 10% of the samples. In each subset, labels follow the same distribution as in the complete dataset, and each participants' samples are contained within the same subset.

B. Expressive confusion dataset

To validate the performance of our mistake detection model and as existing datasets either do not label confusion or do not contain video data, we constructed a small dataset

that consists of short videos showing unambiguous expressions of confusion. As the videos contain clear expressions of confusion, we can investigate whether our model is capable of recognizing a confused state from facial expressions. Furthermore, the videos are short, so the problem of extracting salient moments showing emotional expressions does not impact performance on this task. 139 videos were collected from the public repository Tenor[1], selecting videos that contained either an expression of confusion or a neutral expression. All included videos contain only one person with their face clearly visible, who is exhibiting at least some movement, and appears to be in an interaction with another person. The videos also do not show any visual effects and the camera remains (mostly) stationary. All videos have a frame rate of 10 frames per second, reducing the complexity of the data, and are at most four seconds long. The dataset is imbalanced on purpose (30% neutral, 70% confusion) to resemble the imbalanced distribution of the REPAIR-Corpus dataset, although the imbalance is in the opposite direction, focusing on training the model to detect confusion. The dataset is split into a training and test set with an 80-20 split.

IV. Miscommunication detection system

In this section, we describe and evaluate our miscommunication detection system. The system contains two main components. First, the input – a video that shows one human-robot exchange – is cropped to contain only the user's face, and then given to an algorithm which extracts the most salient moments from this video. These moments are then passed on to the second component, which is a model that classifies the exchange as containing a miscommunication or not. Both components leverage high-level features which are extracted from the videos using existing off-the-shelf models. We first describe these feature extraction models and then discuss and evaluate the salient moment extraction algorithms. Then, we present the miscommunication classification model, first evaluating it on the human-robot dialogue "REPAIR-Corpus" dataset and then investigating how it performs on short videos containing more pronounced emotional expressions.

A. Visual feature extraction

Since raw videos are composed of numerous consecutive still images, their high dimensionality makes them unsuitable for direct input into our model. Therefore, we leverage pre-existing models to extract lower-dimensional representations of the images, capturing essential high-level visual features. We compare three approaches for this, of which the latter two are specialised for facial expressions.

First, we look at general-purpose computer vision neural networks that are originally trained for image classification. Specifically, we compare the convolutional neural network VGG16 [43] and the residual neural network ResNet50 [44]. For both models, the final classification layer is removed, so we use the output of the final hidden layer.

[1]https://tenor.com/en-GB/

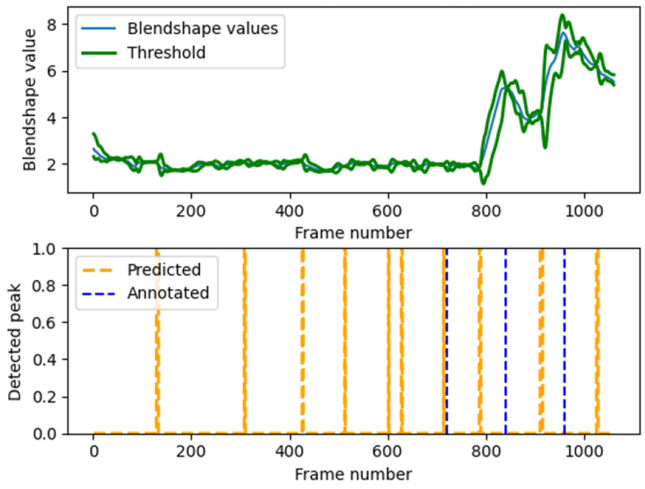

Fig. 2: Illustration of the real-time peak detection algorithm for extracting salient moments, applied on one video fragment. The top graph shows the sum-of-blendshapes signal and the threshold calculated by the algorithm. The bottom graph shows which salient moments the algorithm predicted in blue, with the manually annotated moments indicated in orange.

Second, we use a model that extracts facial keypoints, provided by Google's MediaPipe framework [45]. It detects the location of 468 keypoints on a human face, focusing on more expressive parts of the face.

Third, we use a different model from MediaPipe which extracts *blendshapes* instead. Popular in the domain of digital avatars, this approach extracts 52 scalar values that together control a 3D mesh model of a human face, and is similar to the detection of FAUs, which are often used in related work.

B. Salient moment extraction

As the video fragments in the dataset are too large to be efficiently used as input for a model, and as large parts of the videos often contain no movement, we attempt to extract the most salient moments for mistake detection. As a heuristic, we consider any moment where there is significant movement in the user's facial expression as a salient moment. We compare three approaches to detect these moments: one that is based on extracted blendshapes and the other two use facial keypoints.

The approaches are evaluated on a subset of 20 video fragments from the REPAIR-Corpus dataset. For each fragment, salient moments were manually annotated, for a total of 49 salient moments in the 20 fragments. Each salient moment is identified by a single frame in the fragment. Predicted salient frames are counted as correct when they are less than 60 frames (one second) removed from an annotated salient frame.

In the blendshape approach, each of the 52 blendshape values is first smoothed through a linear convolution over the length of the video fragment, with a window size of 45 frames. This reduces the influence of local extrema, increas-

ing the robustness of the approach. However, taking these 52 separate values into account would be overly complex, as the occurrence of salient moments is not easily mapped to specific behaviours of the separate blendshapes, and applying the same algorithm to all blendshapes at once would result in the entire video being seen as salient. Therefore, all 52 values are summed into one combined scalar signal. This combined signal is then passed onto a "real-time" peak detection algorithm adapted from [46]. For every frame, the algorithm calculates a threshold based on the values of the signal in the previous frames. If the signal exceeds that threshold, the frame is identified as salient. The working of this algorithm is illustrated in Figure 2. After tuning the parameters of the peak detection algorithm, this approach is able to extract 24 of the 49 annotated salient moments, or 49%.

Second, we evaluate an approach that uses extracted facial keypoints. As each keypoint is characterised by (x, y) coordinates in the image, we consider the sum, over all keypoints, of the square of the distance to the keypoint in the previous frame, again creating a single scalar signal. After smoothing this signal with a linear convolution with a window size of 45 frames, it is passed to the same real-time peak detection algorithm as in the previous approach. This approach is able to detect 17 out of the 49 keypoints, or 34%.

Finally, we improve the keypoint-based approach by using a different algorithm, that instead selects the three frames that have the highest sum-of-square-distances with the previous frame, selecting only frames that are at least 60 frames removed from each other. This approach extracts 30 of the 49 annotated salient moments, or 61%. As it performed best out of the three tested salient moment extraction approaches, we use this approach for our future evaluations.

C. Classification model

In the previous section, we developed a salient moment extraction algorithm. As this gives us the three most salient moments per video fragment, we extract the 60 frames around each moment, and concatenate these three mini-clips. The resulting video is a three-second compilation of the most salient fragments from the original video fragment, and can be used as input to our model.

Our model is constructed using the following basic architecture. First, one of the visual feature extraction models described in Section IV-A processes each frame of the video into a lower-dimensional representation. The resulting sequence is used as input for an LSTM neural network, which is designed to process temporal data. A final linear layer reduces the output of the LSTM to one scalar value, which is used to classify whether a mistake was present in the input video.

We evaluate three visual feature extraction models: VGG16, ResNet50, and MediaPipe's blendshapes. In addition, we explore three strategies to enhance model performance: applying a weighted loss function to address dataset imbalance, splitting the samples so that each relevant moment is treated as a separate sample, and shortening the video length by keeping only every n-th frame, comparing different values of n. For each approach, we tune the hidden size of the LSTM, the batch size, and learning rate of the model.

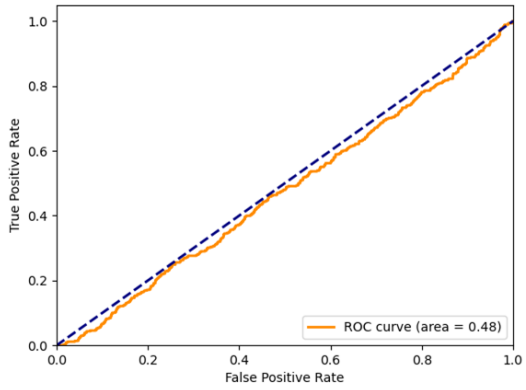

Fig. 3: ROC curve of the best-performing miscommunication detection model on the REPAIR-Corpus dataset.

We report the models' performance by showing their ROC curve, following the example set by prior work [8], [10]. To quantify that performance, we calculate the area under the ROC curve (AUC). An AUC of 1.0 represents a perfect classifier, 0.0 a classifier that is always wrong, and 0.5 a random classifier. Trung et al. reported AUC scores of up to 0.66 [8], while the ROC curves reported by Kontogiorgios et al. showed even better performance without providing the exact AUC [8], [10].

However, not a single one of the tested model configurations performs better than a random classifier. The best-performing model is the ResNet-based model, whose ROC curve is shown in Figure 3. It obtains an AUC of 0.48, which is slightly worse than random chance. In this model, the samples are split up into single salient moments and only every fifth frame is retained. The model was trained for 50 epochs, with the following optimal hyperparameters: batch size of 16, learning rate of $1e^{-4}$ and an LSTM input size of 52 and hidden size of 256. The training and validation accuracy curves show that the model quickly overfits on the training set, implying that longer training would not improve performance on the validation or test set.

D. Performance on expressive confusion dataset

Our mistake detection system performed worse than expected on the human-robot conversation dataset, underperforming prior research. This leaves a question to be answered: is this due to our methodology, meaning our model or training approach, or because of the data? In other words, is the model we designed capable of detecting facial expressions that indicate a user experiencing a miscommunication?

Therefore, as a toy problem to show the suitableness of our model, we apply the same methodology used for our classification model to our dataset of short videos with unambiguous expressions of confusion. This was identified by prior work to be the most reliable expression indicating miscommunications [10].

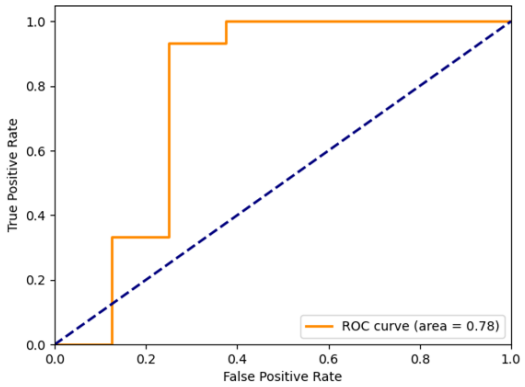

Fig. 4: ROC curve of the best-performing miscommunication detection model on the expressive confusion dataset.

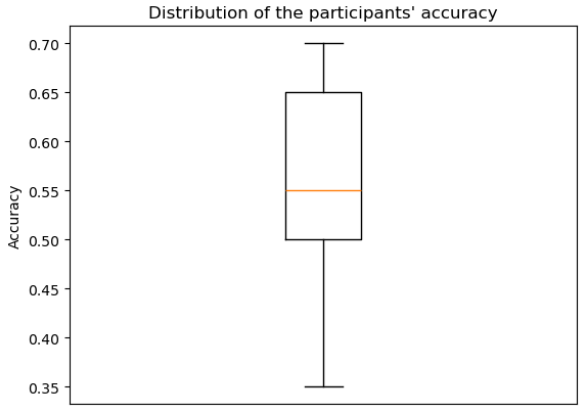

Fig. 5: Distribution of the participants' ($n = 17$) accuracy at identifying videos with robot mistakes.

Using the same model architecture as for the previous dataset, the results show that our approach is able to recognise confusion in facial expressions. As shown in Figure 4, the area under the ROC curve is 0.78, indicating that the model clearly exceeds random chance. The model reaches an F1 score of 74.1% and accuracy of 69.7%. Learning curves also show the model is not overfitting. These results were achieved when using blendshapes as visual features extractor. The model was trained for 75 epochs, with a learning rate of $1e^{-4}$, batch size 8, and an LSTM with input size 52 and hidden size 512. No frames were dropped, as the samples were only 40 frames long.

V. Human Evaluation

Up to this point, we have demonstrated that our mistake detection system, despite being effective at detecting confusion from facial expressions, performed no better than random chance on a dataset of real human-robot conversations. Now, we aim to explore the level of performance we can reasonably expect on this dataset. How accurately can humans identify user reactions to robot miscommunications?

A human evaluation was set up to answer this question, recruiting 17 annotators who were not familiar with the original study. Each annotator was shown the same 20 video fragments sampled from the REPAIR-Corpus dataset, each containing one exchange between the robot and the user. This sample reflects the distribution of the full dataset: 6 of the fragments were labelled as a miscommunication by the user's button feedback, the remaining 14 were not.

Whole fragments were shown to the annotators, meaning the fragments were not passed through the salient moment extraction algorithm and not cropped to the user's face. No audio was included, neither robot speech nor user speech, to completely match the information available to our model and to eliminate any bias from the content of the robot's speech – the aim is for raters to label whether the participant of the original experiment perceived the robot speech as a miscommunication, not whether they themselves perceive it as such.

Participants were asked to label each video one by one as "Mistake" or "No mistake", being instructed to label videos as "Mistake" whenever they thought the robot had made a mistake, the participant was confused, the interaction strayed off its normal course, or it seemed the robot said something strange or wrong.

Results show that human raters perform no better than our classification model. On average, participants correctly classified 56% of the fragments. Performance is also widely spread across participants: as shown in Figure 5, the lowest accuracy of any participant was 35%, with the highest 70%. Calculating the inter-annotator agreement (IAA) using Fleiss' Kappa [47], which ranges from -1 to 1, returns a value of -0.012, meaning absolutely no agreement between the raters.

Looking at performance for each video fragment separately, we see that on average, a video containing a miscommunication is correctly identified by 50% of the participants. More than half of the videos containing a miscommunication are misclassified by a majority of participants. Videos without a miscommunication are, on average, correctly identified by 59% of the participants. While some of the videos without miscommunication reach high human agreement, these videos tend to show a neutral facial expression.

These results clearly indicate that humans are not able to reliably identify miscommunications from only users' facial expressions and body language in these real-life human-robot conversations.

VI. Discussion and Conclusion

This work investigates the performance of a system that detects miscommunications in social human-robot interactions from users' facial expressions. Prior work has identified facial expressions as an important modality for users to express feedback to robots, but does not yet automatically detect facial expressions [9], [10] or focuses on physical human-robot collaborations rather than social dialogue-based interactions [7]. Furthermore, prior work relies on external annotators to label interactions as miscommunications [2] or uses purposefully designed robot failures without verifying that users perceive them as such [3]. This notion is particularly uncertain in the context of social failures [8].

Aiming to bridge the gap between stellar results in facial expression recognition and human-robot interaction, we built a system that processes robot-perspective videos of human-robot exchanges and detects whether they contain a robot-caused miscommunication, examining the user's facial expressions. This system first extracts the most salient moments in the video before classifying it. Given the lack of facial expression recognition models capable of processing videos and identifying complex affective-cognitive states such as confusion, crucial for detecting miscommunications, our system uses pre-trained models that track facial movements.

We trained and evaluated this system on REPAIR-Corpus: a multimodal dataset of educational human-robot conversations. While the robot was explaining recipes to the user, it introduced four types of scripted miscommunications. After each robot utterance, the user pressed a button to indicate whether they experienced this as a miscommunication.

However, our system performs below expectations. Despite evaluating multiple configurations, it never performs better than a random classifier. Investigating the cause of this low performance, we first tested our system on a toy dataset which is more similar to traditional facial expression recognition datasets, containing very clear expressions of confusion. The system performed well on this dataset, showing the suitability of our technical approach.

We then ran a human evaluation study to assess whether external annotators are able to identify miscommunications in our dataset. While prior work had shown reasonable success in detecting miscommunications from videos without audio [2], [10], our participants showed absolutely no agreement when classifying videos from our dataset.

These results highlight an underexposed question in human-robot interaction: when do users genuinely intend to convey (negative) feedback to a robot? Our results indicate a discrepancy between the occurrence of miscommunications, even as perceived by users, and the moments when users express them in a way that is noticeable even to other humans. Combined with the challenge that feedback is swift and much more subtle than can be classified into basic emotions, it can explain why automatic detection of miscommunications or mistakes performs worse than expected in social human-robot interactions.

Other questions are raised as well. Do people convey miscommunications through the same facial expressions and to the same extent to robots as to other humans? Would the human annotators do better at recognising miscommunications in human-human interactions? Additionally, we know that context is very important when interpreting facial expressions. What is the impact of different feedback modalities, such as the user and robot utterances, on recognition performance? Finally, we recognise that our study took place in a lab setting and cannot fully approximate a real-world interaction. Participants likely expected robot mistakes, and the miscommunications did not have any immediate consequences to them. Could the presence of a feedback button have impacted the users' tendency to convey non-verbal (or verbal) feedback? Or did the robot's lack of reaction to their feedback have an impact?

Future work should explore the factors that affect when users intend to communicate their feedback in a perceivable manner, such as the interaction context, the robot's embodiment or perceived autonomy, and other potential influences. Additionally, it is crucial to integrate these findings with the dialogue context, as well as verbal and other non-verbal feedback. These insights will aid in designing better human-robot interactions by highlighting the complexity of user feedback in social human-robot conversations. They also present clear challenges to facial expression recognition research regarding its applicability in autonomous human-robot interaction.

REFERENCES

[1] H. H. Clark, *Using language*. Cambridge university press, 1996.

[2] A. Axelsson, H. Buschmeier, and G. Skantze, "Modeling feedback in interaction with conversational agents—a review," *Frontiers in Computer Science*, vol. 4, p. 744574, 2022.

[3] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures," *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238992820

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[5] S. Labat, R. Janssens, L. Lismont, T. Belpaeme, T. Demeester, and V. Hoste, "Repair-corpus: Robot errors provoking affective-cognitive interactional reactions," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15711723

[6] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 53–62. [Online]. Available: https://doi.org/10.1145/3319502.3374782

[7] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 222–230.

[8] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi, "Head and shoulders: automatic error detection in human-robot interaction," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 181–188. [Online]. Available: https://doi.org/10.1145/3136755.3136785

[9] A. Axelsson and G. Skantze, "Multimodal user feedback during adaptive robot-human presentations," *Frontiers in Computer Science*, vol. 3, p. 741148, 2022.

[10] D. Kontogiorgos, A. Pereira, B. Sahindal, S. Van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 53–62.

[11] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980, pp. 3–33.

[12] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. The MIT Press, 1974.

[14] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, pp. 715 – 734, 2005. [Online]. Available: https://doi.org/10.1017/S0954579405050340

[15] P. C. Ellsworth and C. A. Smith, "From appraisal to emotion: Differences among unpleasant feelings," *Motivation and Emotion*, vol. 12, pp. 271–302, 1998. [Online]. Available: https://doi.org/10.1007/BF00993115

[16] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005. [Online]. Available: https://doi.org/10.1177/0539018405058216

[17] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[18] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[19] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[22] S. Buechel and U. Hahn, "EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 578–585. [Online]. Available: https://aclanthology.org/E17-2092

[23] K. R. Scherer, M. Mortillaro, I. Rotondi, I. Sergi, and S. Trznadel, "Appraisal-driven facial actions as building blocks for emotion inference," *Journal of Personality and Social Psychology*, vol. 114, p. 358–379, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3436924

[24] M. A. Stranisci, S. Frenda, E. Ceccaldi, V. Basile, R. Damiano, and V. Patti, "APPReddit: a corpus of Reddit posts annotated for appraisal," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, June 2022, pp. 3809–3818. [Online]. Available: https://aclanthology.org/2022.lrec-1.406

[25] E. Troiano, L. Oberländer, and R. Klinger, "Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction," *Computational Linguistics*, vol. 49, no. 1, pp. 1–72, Mar. 2023. [Online]. Available: https://aclanthology.org/2023.cl-1.1

[26] P. C. Ellsworth, "Confusion, concentration, and other emotions of interest: Commentary on rozin and cohen (2003)," *Emotion*, vol. 3, pp. 81–85.

[27] S. D'Mello and A. Graesser, "Confusion and its dynamics during device comprehension with breakdown scenarios," *Acta Psychologica*, vol. 151, pp. 106–116, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001691814001504

[28] P. Rozin and A. B. Cohen, "High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans." *Emotion*, vol. 3, pp. 68–75, 2003.

[29] S. D. Craig, S. D'Mello, A. Witherspoon, and A. G. and, "Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning," *Cognition and Emotion*, vol. 22, no. 5, pp. 777–788, 2008. [Online]. Available: https://doi.org/10.1080/02699930701516759

[30] S. M. Alarcão and M. J. Fonseca, "Identifying emotions in images from valence and arousal ratings," *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 17 413–17 435, July 2018. [Online]. Available: https://doi.org/10.1007/s11042-017-5311-8

[31] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2Exp: Combating Data Biases for Facial Expression Recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 20 259–20 268. [Online]. Available: https://ieeexplore.ieee.org/document/9879702/

[32] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition," Sept. 2016, arXiv:1609.06591 [cs]. [Online]. Available: http://arxiv.org/abs/1609.06591

[33] https://github.com/kdhht2334/awesome-SOTA-FER?tab=readme-ov-file, accessed on May 23, 2024.

[34] A. Kappas, R. Stower, and E. J. Vanman, "Communicating with robots: What we do wrong and what we do right in artificial social intelligence, and what we need to do better," *Social intelligence and nonverbal communication*, pp. 233–254, 2020.

[35] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling human response to robot errors for timely error detection," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.00565

[36] A. Axelsson and G. Skantze, "Do you follow? a fully automated system for adaptive robot presenters," in *Proceedings of the 2023 acm/ieee international conference on human-robot interaction*, 2023, pp. 102–111.

[37] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Springer, 2012, pp. 114–130.

[38] C. Torrey, A. Powers, M. Marge, S. R. Fussell, and S. Kiesler, "Effects of adaptive robot dialogue on information exchange and social relations," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, ser. HRI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 126–133. [Online]. Available: https://doi.org/10.1145/1121241.1121264

[39] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, "AI challenges in human-robot cognitive teaming," *CoRR*, vol. abs/1707.04775, 2017. [Online]. Available: http://arxiv.org/abs/1707.04775

[40] S. Serholt, L. Pareto, S. Ekström, and S. Ljungblad, "Trouble and repair in child–robot interaction: A study of complex interactions with a robot tutee in a primary school classroom," *Frontiers in Robotics and AI*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2020.00046

[41] C. Esterwood and L. P. R. Jr, "Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness," *Computers in Human Behavior*, vol. 142, p. 107658, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563223000092

[42] I. J. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015, special Issue on "Deep Learning of Representations". [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014002159

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[45] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[46] J. v. Brakel, "Robust peak detection algorithm using z-scores," https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/22640362#22640362, 2014. [Online]. Available: https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/22640362#22640362

[47] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.