Designing a Multi-Modal Social Robot for Enhanced Emotion Recognition and Empathy

Diana Cernetchi, 2737113, d.cernetchi@student.vu.nl[1]
Anesa Ibrahimi, 2693458, a2.ibrahimi@student.vu.nl[1,2]
Lisann Becker, 12558141 , lisann.becker@student.uva.nl[2]
Michael Moor, 2672846, m.s.moor@student.vu.nl[1]
My Vu, 2773098, m.t.t.vu@student.vu.nl[1]
Hao Yu Wang, 2666479, h.y.wang@student.vu.nl[1]
Raj Kasimahanti Raj Gopal, 2766494, kasimahanti.raj.gopal@student.vu.nl[1]

Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
Universiteit van Amsterdam,1012 WP Amsterdam, Netherlands

Abstract. This research evaluates a NAO robot built to understand and respond to human emotions via text-based and multimodal interactions. The robot detects and predicts human emotional states by integrating voice, text, and visual emotion recognition technologies, including speech-to-text, text emotion classification, audio emotion classification, and visual emotion analysis. Participants in the study engaged with the robot in two groups: text-based and multi-modal. The study evaluated the robot's ability to reliably identify emotions, its response to emotional stimuli, and overall participant satisfaction. The results suggest that, while the robot performed better in text-based exchanges, the multi-modal systems were as capable or even better in many cases when classifying broader emotional states. The findings highlight the challenges and potential of multi-modal emotion recognition systems in human-robot interaction, offering insights into improving emotional intelligence in robotic systems.

Keywords: Multimodal Emotion Recognition · Human-Robot Interaction · Emotional Speech Synthesis · NAO Robot · Social Robotics.

1 Introduction

1.1 Use Case

The integration of empathetic social robots holds significant potential to enhance human experiences and well-being across diverse settings. Research [6], [10], [4] highlights the capability of social robots to provide emotional support and improve the quality of interactions. Our project aims to develop an empathetic robot designed to recognise and respond to a broad spectrum of human emotions, without targeting any specific demographic group. This work aims to combine auditory, visual, and textual information to enhance the robot's ability to detect and respond to human emotions. Our contributions include the development of a real-time emotion recognition and response generation framework.

1.2 Related Work

Empathy in human-robot interaction (HRI) has been a topic of growing interest over the past few decades, drawing from fields such as affective computing, psychology, and artificial intelligence. The work of Breazeal [4] highlighted the need for social robots to exhibit behaviours and cues that resonate with human social expectations. She argued that robots designed with social intelligence should incorporate nonverbal signals and expressive modalities to facilitate better interactions. This perspective laid the groundwork for understanding that empathy, as a core aspect of human interaction, is also critical in HRI.

[10] delved deeper into how empathic behaviour impacts long-term interactions with robots, showing that consistent and adaptive empathy fosters user engagement and builds rapport. Their

research supports the notion that empathy should be a dynamic aspect of robotic behaviour, capable of adjusting to a user's emotional state. In line with this, Cavallo et al. [5] studied emotional expressions in robots, concluding that empathetic responses significantly enhance user perception of the interaction and engagement. This work aligns with findings from Picard [13], who pioneered the field of affective computing and emphasised the importance of systems that recognise and simulate human emotions to improve interaction quality.

The psychological underpinnings of empathy and nonverbal communication are essential in robotic design. Kleinsmith et al. [8] explored how nonverbal cues contribute to emotional understanding and stressed that prosody, such as variations in pitch and tempo, plays a key role in conveying emotions effectively. Similarly, [1] focused on how prosodic elements in robotic speech influence users' perception of empathy, finding that appropriate pitch modulation can greatly enhance a robot's emotional expressiveness. These insights are relevant for informing design choices that aim to make robots appear more empathetic.

Riek [15] expanded on this by discussing how empathetic behaviour in robots must be carefully structured to maintain user comfort and authenticity. She noted that over-exaggerated or poorly timed expressions of empathy could appear insincere or even unsettling. This is complemented by research from Paiva et al. [12], who studied artificial empathy in autonomous agents and found that response timing plays a significant role in the perceived naturalness of interactions. Timing that mimics human-like reaction delays, they argue, can contribute to the believability and effectiveness of empathetic responses.

Empathy's role extends beyond emotional recognition to the robot's ability to respond thoughtfully. [6] highlighted the importance of social robots exhibiting behaviours that contribute to social interaction, proposing that empathy should not only be reactive but also proactive. This involves generating responses that show active listening and appropriate emotional support. Building on this, [19] explored multimodal emotion recognition, emphasising that the integration of vocal, visual, and contextual data allows for more nuanced behaviour in robots.

The integration of empathy also benefits from insights within the field of social psychology. Empathy involves understanding not just explicit emotional expressions but also subtle cues that indicate underlying emotional states. For example, Rogers and Tannen [16] investigated the impact of conversational pauses and timing on perceived empathy in interpersonal communication. Their findings indicate that measured, thoughtful responses create an impression of attentiveness and care. Applying this knowledge to robotics, implementing strategic pauses can make the robot's responses appear more thoughtful.

1.3 Design Question

How to design an emotionally aware multi-modal robot capable of recognising and responding to human emotions by integrating auditory , visual and textual information ?

1.4 Research Question

How can the integration of vision-based and language-based emotion recognition models enhance a social robot's ability to detect and empathetically respond to a person's emotional state?

1.5 Hypotheses

Hypothesis 1: *Integrating emotion recognition from a vision model and language models (RoBERTa for text emotion and SpeechBrain for voice emotion) will improve the accuracy of the robot's emotion detection compared to using a single modality.* **Motivation:** Multi-modal emotion recognition systems have been shown to outperform unimodal systems in accurately detecting human emotions.

Hypothesis 2: *A greater consistency between the emotion predictions of the multi-modal model and language models will correlate with increased user perception of the robot's empathetic response.* **Motivation:** Multimodal systems often outperform unimodal systems because they integrate information from multiple sources, providing a richer and more comprehensive understanding of

context and emotions.

Hypothesis 3: *Using specific interaction scenarios intended to trigger different emotions will effectively induce target emotions in the users, and therefore allow for sound measurement of ground truth emotions.* **Motivation:** Controlled prompting of emotions is essential for evaluating emotion recognition models and requires reliable methods to measure users' true emotional states.

1.6 Evaluation Metrics

An Empathy Perception Scale will be employed as a questionnaire to measure how users rate the empathetic behaviour of the robot during interactions, capturing perceived empathy as the dependent variable. Additionally, the User Satisfaction Score will be collected through post-interaction surveys to quantify how well users feel the interaction met their emotional needs, aligning with user expectations of empathetic behaviour.
Modal Alignment: Metric assessing the degree to which the emotion predictions of the vision model (VLM) and the language models (LM) correspond.
Emotion Elicitation: How successfully the interaction scenarios triggered the intended emotions from users, measured by the match between the target emotions and the reported ground truth emotions.
Empathy Perception Scale: A questionnaire where users rate the robot's empathy during interactions, capturing perceived empathy as a dependent variable.

2 Interaction Design

2.1 Approach overview

In order to enable the multi-modality capabilities of a social robot which facilitates seamless interaction with humans, verbal and non-verbal cues have been integrated within the robot's framework to accurately perceive human emotions. As highlighted in [2], social robots are considered to be 'effective' when they're able to imitate emotions portrayed by humans in such a way that the interaction feels natural rather than artificial. The naturalness of emotion expression of social robots is therefore imperative for humans to treat them equally and ensure social compatibility [2]. As mentioned in [9], a significant accuracy in emotion-recognition is achieved when speech and image modalities are combined and analysed together rather than processed separately. Therefore we aimed at utilising multi-modal emotion recognition through speech-based verbal cues, which take into consideration both text and voice emotion cues. In parallel, vision-based non-verbal cues, involve analysing facial expressions, gestures, and other body language to interpret emotional states.

2.2 Design of robot interaction modalities

2.2.1 Input Modalities

2.2.1.1 Speech-Based Verbal Cues: The verbal cues extracted during interaction involved primarily two channels jointly working together. The first channel, **audio**, enables the robot to detect emotions from user's tone, pitch, and intensity. Whereas the second channel, **text**, allows the robot to discern emotions via the generated text output which is transcribed from the audio. The purpose of these two channels is to allow the robot to effectively detect emotions based on the linguistic content *and* acoustic properties of the user's input. In more detail, the linguistic content is extracted through the reliance of a transformer-based language model (RoBERTa [11]) which processes the text and then predicts an emotion label corresponding to the tone of textual content. On the other hand, the acoustic properties, are handled by SpeechBrain [17] model which directly uses the audio to output an emotion label.The importance of making use of these channels *jointly* rather than separately has been carefully studied and showcased in [3]. More closely, it studied

three different settings (Audio-only, Text-only,Audio & Text). It then concluded that in the Audio & Text Setting where both features were combined , a significant improvement in performance for emotion-detection was present while efficiently resolving cases of ambiguities.

2.2.1.2 Vision-Based Non-Verbal Cues: Humans have shown to simultaneously apply several different emotion modalities. This phenomenon has been extensively covered by [18] where emotional cues from face, body, and voice all interacted with one another. The study [18] therefore stresses the importance of multi-modal recognition approaches and observes that body motion alongside posture were impacted with the emotion expressed in a person's voice and face. It therefore claims that multi-modality plays a significant role in higher emotion recognition performance which is why non-verbal cues were added to the robot's detection framework.More specifically, a webcam is used to capture frames of the user's face and body where each frame is then processed via OpenCV [3] in specific intervals to extract non-verbal cues such as facial expressions and general body language(posture,leaning,movement). After extracting these cues, a detected emotion label is generated alongside its confidence levels. These detected non-verbal cues are then able to provide additional context in cases of ambiguity from verbal cues as well as more comprehensive analysis on human emotions.

2.2.2 Output Modalities

2.2.2.1 Emotionally Adaptive Responses: The abundant cues fed into the robot's framework have to all be efficiently handled and ultimately mirror the right , if not the most adequate, emotion back to the user. The first major block that can be attributed to handling the task above, is the OpenVoice[14] speech synthesis model. This model, adjusts the *tone* and *emotion* of the robot's speech based on the final detected emotion across the multi-modal analysis phase. Emotional mappings from text and speech are carefully adjusted by us to match with the already-available tone and emotion parameters of a pre-trained model (V1) from OpenVoice. Therefore, providing natural and emotionally compatible verbal interaction. Alongside the speech synthesis model, responsible for generating the robot's voice, is the contextual awareness framework. This framework can be further decomposed to two essential parts. First, the user-robot interaction log, which provides context to the robot by serving as shared memory of the robot.In return, allowing new responses to be referenced and get influenced from previous messages to ensure smooth conversation flow. Second, the emotional analysis prompt, which provides a reasoning framework to infer the final emotional state and assess the emotional signals before passing to the LLM which is used to generate robot's response. In exchange, the speech synthesis model makes the necessary adjustments to robot's acoustic properties to align with the inferred emotional state.

2.2.2.2 Visual Feedback Mechanisms: Real-time responsiveness is a desired attribute throughout this project in order to interact in real-time with users. In efforts of enabling this feature (more details provided in 2.3) we made use of some visual feedback to guide and indicate operational status to the current user. This in effect, modifies the LED states of robot's eyes to visually reflect when the robot is actively listening/engaging or when it is processing user's input. The former displays *green* while the latter displays *red* LED signals. Essentially allowing the user to understand the current state of the robot and interaction flow.

2.3 Implementation details

The implementation of our system integrates several components to achieve real-time, multi-modal emotion recognition and response generation. Below, we discuss those components, their roles, and their strengths and weaknesses. The interaction diagram in figure 1 shows a high-level visualisation of the interaction flow and it's individual components.

Speech-to-Text uses Google Speech-to-Text[1] to convert spoken input into text. It is accurate with clear speech but can struggle in noisy environments. We use the laptop microphone instead

[1] https://cloud.google.com/speech-to-text

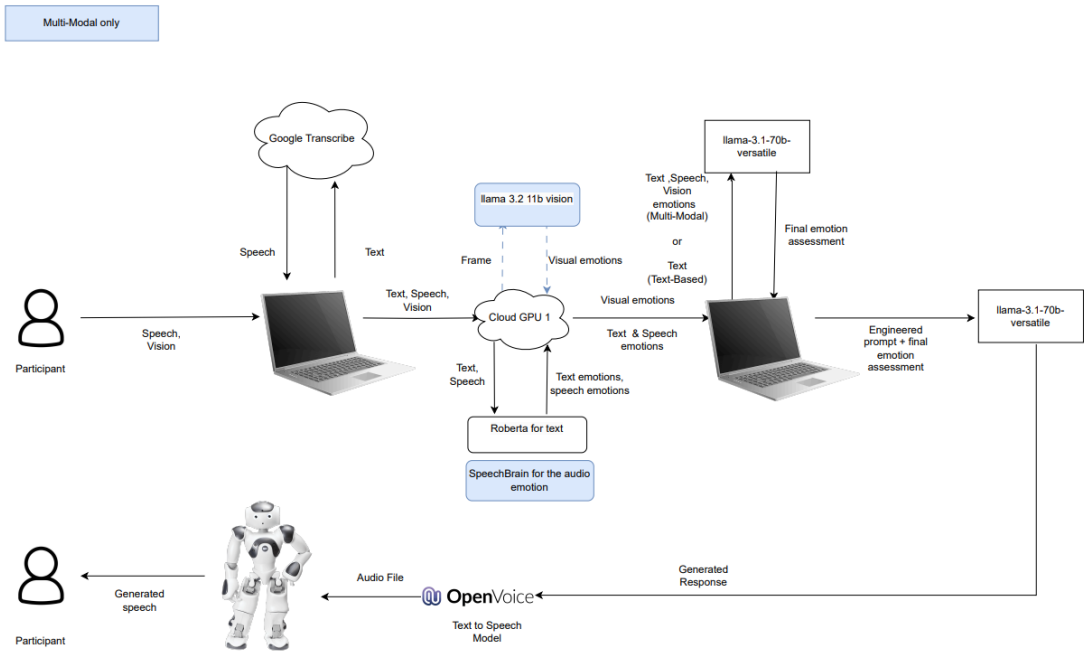

Fig. 1. NAO bot interaction diagram.

of the NAO microphone to get a reliable input.

Text Emotion Classification is done by RoBERTa[7], which analyses the transcribed text to detect emotional cues. It predicts 6 emotions: anger, disgust, fear, joy, neutral, sadness and surprise.

Audio Emotion Classification is done by SpeechBrain[17] to classify the emotion from audio files. It predicts 4 emotions: anger, sad, happy and neutral. We considered using other models with a broader range of emotions that better align with the speech synthesis system, but SpeechBrain[17] offered better real-time performance and accuracy.

Visual Emotion Analysis is done by LLaMA 3.2 11B Vision. It processes facial expressions and body language from visual input for emotion assessment. It predicts an emotion and a description of the body language. It delivers high accuracy in ideal lighting, but its performance lowers with poor-quality video input. Which is why we used the laptop webcam instead of the NAO's camera.

The Final Emotion is predicted by Groq[2] with LLaMA 3.1-70B[3] Versatile. We combine the emotional cues from text, audio, and visual inputs into a prompt and let the LLM decide on the final emotion. The final emotion is left for the LLM to decide, as determining the weights for each modality is complex and context-dependent.

Response Generation also uses Groq with LLaMA 3.1-70B Versatile. We combine the user input from the speech-to-text component with the final emotion into a prompt to generate a reply with the final emotion taken into consideration.

Text-to-Speech implements OpenVoice[14] for emotional speech generation. It maps detected emotions to appropriate speech tones and supports various emotional states (terrified, sad, cheerful, neutral, and angry).

Real-Time Performance is achieved through a distributed system architecture, using Docker containers to encapsulate each component. Emotion analysis is implemented in one Docker container and OpenVoice runs in another. Groq is accessed via its API. Google Speech-to-Text is also accessed through its API.

[2] https://groq.com
[3] https://huggingface.co/meta-llama/Llama-3.1-70B

3 Study Design

3.1 Methodology

3.1.1 Participants Twelve students were assigned to our study by the course. Our subjects' main task would be to assess the capacity of our NAO robot for emotional perception and adaptive reaction. The participants were split equally into two groups of six. One group had conversations with the text-based NAO robot while the other group was given the multi-modal NAO robot that combined text, visual, and auditory input to, in theory, better identify and adjust to a human's emotional state. Under carefully monitored circumstances intended to elicit and assess emotional reactions, each participant interacted with the given robots in particular setup described below.

3.1.2 Experimental design The experiment used a within-subjects design to evaluate the robot's capacity to recognise and adjust to human emotions in two different interaction modes: multi-modal and text-based only. The text-only group focuses on emotion detection and responses based on textual content, while the multimodal group assesses the integration of visual, audio, and textual cues for enhanced emotion recognition and empathy. Carefully chosen visual stimuli were intended to produce both good and negative emotional states in each subject to illicit a diverse set of responses. The purpose of the study was to evaluate how well the robot recognised emotions and how responsive it was in both models. Participant input regarding the perceived quality of the interaction and the robot's accuracy in identifying participants' emotional states were the main dependent variables. The whole experimental environment was conducted in a Python script application and experimental results were directly saved for later use.

3.1.3 Measures & Instruments Several metrics were used in the experiment to assess the robot's performance:

1. **Mini-Surveys:** Participants answered a brief survey on their feelings after seeing each image. This acted as a starting point for the emotional state that the robot was supposed to identify in addition to the baseline visual stimuli presented to them.
2. **Robot Logs:** Based on the input of the participants in each model, the robot documented what emotion it was able to identify from the human subject.
3. **Final Surveys:** At the end of the trial, a 5-point Likert scale survey with 13 questions was distributed. The participants' opinions on the overall satisfaction, flexibility, and accuracy of the robot's interaction in emotional recognition were assessed by this survey.

3.1.4 Materials & Set-up The experimental setup was designed to ensure a controlled environment for evaluating the NAO robot's emotional recognition and response capabilities. The following materials and setup were used:

Materials

- **NAO Robot:** The primary interaction device used in the study. The robot was equipped with speech and visual recognition capabilities, controlled from a laptop.
- **Laptop:** Used for running the emotion recognition models and managing the interaction between the participant and the NAO robot. The laptop's microphone and webcam were utilised instead of the NAO's built-in sensors to ensure higher quality input.
- **Survey Instruments:** Mini-surveys and a final questionnaire were administered to the participants to capture their emotional states and feedback on interaction.
- **Intercoder Questionnaire:** Used by two independent coders to assess the emotions expressed by each participant during the experiment. This questionnaire helped in comparing the robot's performance with participant-reported emotions.

Set-up

- **Experimental Room:** The experiment was conducted in a quiet, well-lit room to minimize environmental interference and ensure optimal performance of the visual and audio models.

- **Participant Seating:** Participants were seated comfortably in front of the NAO robot, with the laptop positioned to capture clear audio and video input.
- **Robot Positioning:** The NAO robot was placed on a table at eye level with the participant to facilitate natural interaction.
- **Software Setup:** The interaction framework was implemented using a Python script that managed the presentation of emotional stimuli, collection of survey responses, and logging of the robot's detection capabilities.
- **Data Logging:** All interactions, survey responses, and emotion detection logs were systematically documented and stored.

3.1.5 Procedure The experiment procedure was created to achieve uniformity and reliability between sessions. Each participant went through the following steps to facilitate a structured assessment of the emotional intelligence capabilities of the NAO robot in two distinct interaction models:

- **Emotion Induction:** The participants were shown two photos that trigger different emotional states using the International Affective Picture System (IAPS) dataset. The first image was intended to provoke a good sentiment (excitement), whilst the second image was chosen to elicit a negative reaction (disgust). After viewing each image, participants took a mini-survey to report their perceived emotions.
- **Robot Interaction:** After each emotion induction, the participants engaged in conversation with the robot. Depending on the group, the robot either employed a textual or multimodal model to communicate with the subjects. After around a minute of conversation, the robot processed the participant's input and attempted to determine their emotional state, which was recorded for study.
- **Final survey:** After the two interactions were concluded, participants were asked to complete a final survey that evaluated the robot's performance. This survey offered valuable insights into the robot's ability to recognise emotions, adjust its responses accordingly, and create a sense of understanding among participants.
- **Data Logging:** The progression of the experiment, which included the presentation of emotional stimuli, collection of survey responses, and logs of the robot's detection capabilities, was systematically managed and documented through a Python-based experimental script.

3.2 Analysis Plan

3.2.1 Data Collected The collected data was stored as `.csv` files. Below is an overview of what has been collected.

- **Inter-Coder Assessments:** Two independent coders will evaluate the emotions expressed by each participant during the experiment. These assessments will be used for comparing the robot's performance and participant-reported emotions.
- **NAO Robot's Emotion Predictions:** The NAO robot's emotion detection logs, based on its framework.
- **Self-Reported Emotions:** Participants will self-report their emotional states through mini-surveys administered immediately after an emotional stimuli.

3.2.2 Inclusion Criteria for Data All participant interactions with the NAO robot have been included in the analysis, as there were no obvious anomalies or incomplete survey responses.

3.2.3 Analysis Given the small sample size of this study (10 participants), no statistical tests will be performed as it would not provide relevant or meaningful insights. Instead, the analysis will employ qualitative and descriptive methodologies to extract insights from the collected data. Visualisations, such as bar charts, will allow us to assess trends and discrepancies visually, providing an intuitive understanding of the overall patterns.

While formal statistical testing will not be conducted, conclusions will be drawn by assessing trends and patterns from the visual data. For example, alignment or divergence between the NAO's emotion predictions, inter-coder assessments, and self-reported emotions will be analysed to evaluate the robot's accuracy and responsiveness.

4 Pilot Study

4.1 Methodology

The pilot study methodology did not differ from the study design too much. However, there were two aspects where changes had to be made to adjust to experimental conditions. First, instead of the original 12 participant subjects, due to time restrictions, had to be cut to only 10 subjects. These participants were split into equal groups of five to maintain equality and consistency and did not have substantial impact on the experiment results. Second, due to the environmental conditions, intercoder analysis was implemented to observe subjects and their responses. Metrics include the precision of robot emotion detection (compared to self-reported emotions), perceived empathy, and participant satisfaction across interaction modes. Due to the small sample size, the analysis focused on descriptive statistics and qualitative insights into system effectiveness and empathy.

4.2 Pilot Results

4.2.1 Emotion Detection Accuracy The main aim of the experiment was to evaluate the NAO robot's capability to identify and correctly forecast the emotional states of the subjects depending on the NAO robot mode: text-based or multi-modal. Every participant experienced two emotional stimuli: excitement and disgust, aiming for these feelings to elicit a response from the individual that would be evident in their self-reported emotions as well as the robot's forecasts. To comprehend the relationship between the robot's predictions and the subject's perceived emotions, we examined the correlation among three important variables: the emotion elicited by the image (trigger), the emotion chosen by the subject (subject emotion), and the emotion forecasted by the robot (predicted emotion). In Table 1 and Table 2 in the Appendix, a comprehensive table of the experiment results can be found. Furthermore, the emotions noted by the participants were considered by two intercoder evaluators to offer a thorough insight into the precision of emotion recognition.

In Figure 2, results of the emotion detection are visualised. Full hit marks a match between all three categories of emotion trigger, selected emotion, and predicted emotion. Hit marks a match between only selected emotion and predicted emotion. Among the 10 participants and 20 robot interactions, there was a full hit in just two cases. These two examples appeared in the text-based model for the emotional response of disgust. In a general assessment of the robot's predictive accuracy, it was discovered that, when disregarding the emotional triggers, the robot accurately identified the subject's chosen emotion in 12 out of 20 cases (60%). Out of these 12 occurrences, 8 took place in the text-only model, whereas 4 took place in the multimodal model. This outcome indicates that, although the robot could identify emotions more often in text interactions, the multi-modal model displayed comparatively fewer accurate predictions.

The analysis additionally showed that "disgust" and "fear" were the most prevalent emotional predictions among all subjects and robot models, whereas intense emotions like "happiness" and "anger" were predicted less often. This could suggest that the robot's emotional recognition system is more attuned to negative emotions, while positive or intense emotional responses might be harder for the robot to identify or are less frequently activated by the displayed images.

In Figure 3, when categorising emotions into three main superclasses - neutral, negative, and positive - 18 of 20 cases (90%) demonstrated alignment between the emotion chosen by the subject and the emotion predicted by the robot within the same emotional superclass. This suggests that, although the specific emotional labels weren't consistently aligned, the robot's predictions typically fell within the same emotional category as the participants' self-reported emotions, with the multi-modal model having two more hits and no misses while the text-based model had one more full hit but missed more often. Interestingly, the matching rates were higher for the multi-modal model, suggesting that this type of robot interaction did indeed have a slight edge in detecting the broader emotional category but lacked when identifying the specific emotions compared to the text-based model.

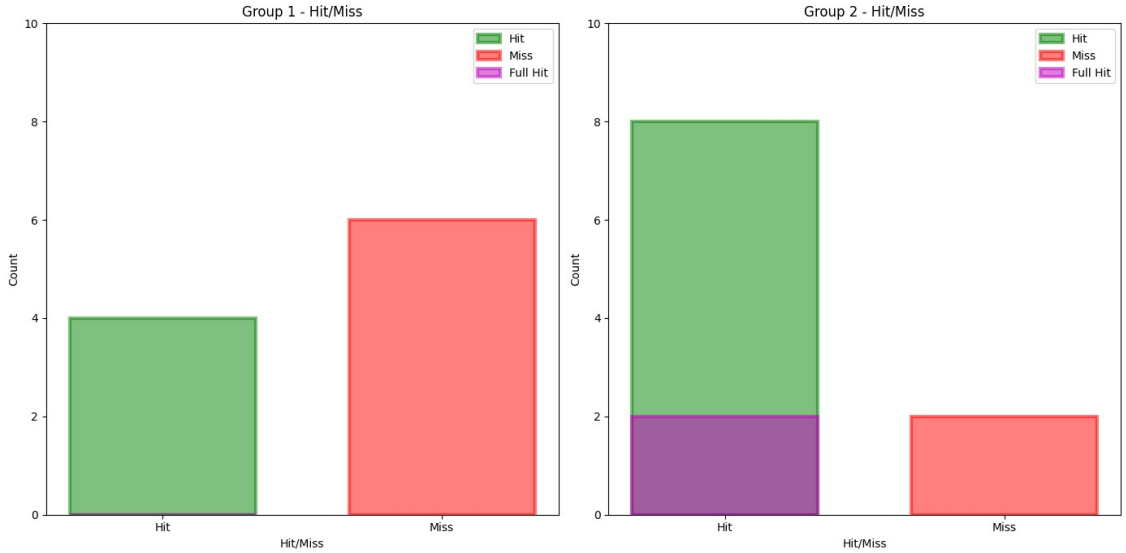

Fig. 2. Results of emotion detection

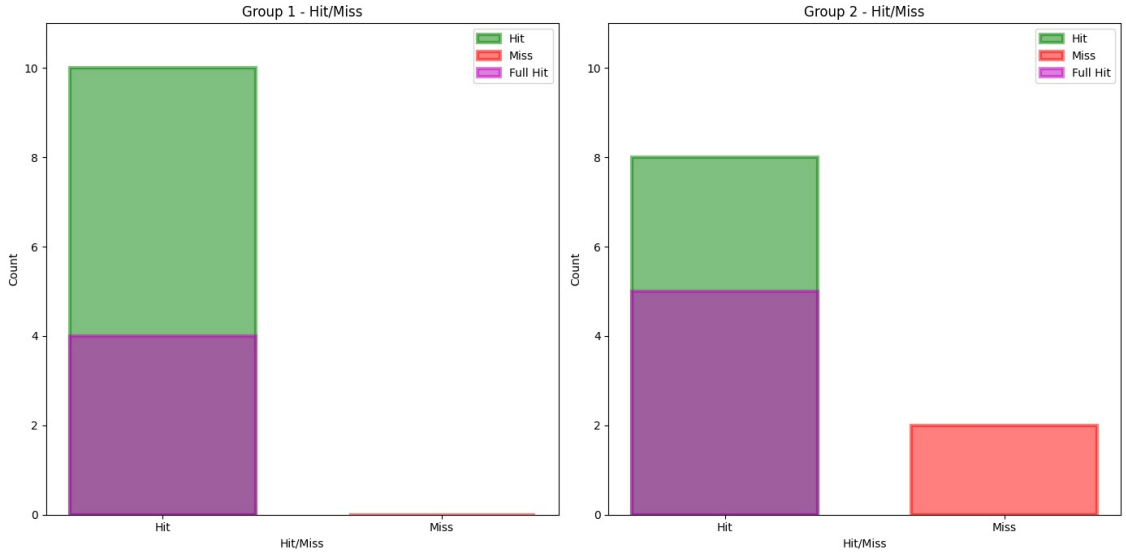

Fig. 3. Results of emotion detection when dividing emotions into superclasses of positive, neutral, and negative

The analysis indicated a pattern in which both the emotions chosen by the subject and those predicted by the robot tended to be more frequently negative, with "disgust" and "fear" being the most dominant emotion. Positive emotional triggers seemed to be less successful in eliciting the intended emotions, as both the robot and participants demonstrated a reduced tendency to choose neutral or positive emotions. This indicates that negative emotions could be simpler to trigger or maintain in a lab environment, while positive emotions might need more subtle engagement or tend to last for less time.

The intercoder analysis entailed contrasting the emotions noted by two independent judges with the emotions felt by the subject and the subject's reactions to the robot interaction, adding another level of verification for the experiment. It showed a fair agreement in emotion classification among

the intercoders, but notable discrepancies in the intensity ratings of those emotions existed, with one coder typically giving higher intensity ratings than the other. The results of the intercoders' annotations can be found in the Appendix.

Cohen's Kappa scores show varying levels of agreement:

- **Happiness**: Substantial agreement (**0.70**).
- **Fear**: Very low agreement (**0.09**).
- **Surprise, Neutral**: Fair agreement (**0.36**–**0.37**).
- **Disgust**: Moderate agreement (**0.59**).

This implies that happiness was simpler to recognise, whereas fear and neutral responses posed more difficulties and showed greater variation for the intercoders. Overall, the intercoders primarily over-selected three emotions from the seven provided: happiness, surprise, and neutral.

When comparing selected, predicted, and observed emotions, patterns start to emerge. Fear and disgust were the most frequent emotions, with happiness and anger being less common but more common than those emotions that are not even represented. For subject 1 (multimodal, excitement trigger), the subject chose disgust while the robot predicted fear after a short conversation, all the while the intercoders observed neutral and happiness. This instance highlights the difficulty in identifying emotions even for human intercoders. In this instance, the robot is observed to have identified an emotion closer to the one that the subject has chosen, compared to the human intercoders. The difficulty both humans and robots face in emotion identification suggests that misinterpretations by the robot reflects human-like imperfections and identifying the emotions of human beings is a tedious and complex task that is often impossible for humans to perform as well.

4.2.2 Emotion Satisfiability: Robot's Adjustment to Emotional Needs Following the complete experimental interaction, participants filled out a comprehensive questionnaire to assess the robot's empathetic reactions and its overall effectiveness in responding to their emotions. The survey comprised 13 questions, evaluating elements like the robot's capacity to soothe the participant, motivate them, and offer suitable emotional reactions. The findings from this questionnaire showed notable disparities between the multi-modal and text-based models regarding perceived empathy and interaction quality.

Overall, individuals in the text-based condition typically assessed the robot's empathetic replies more positively than those in the multi-modal condition, reflected in the higher average agreement scores for Group 2, table of results can be found in the Appendix. This discovery indicates that although the multi-modal interactions incorporated extra sensory inputs (visual and audio), the text-based model might have been more concentrated and accurate in its emotional recognition and reaction. Furthermore, participants in the text-based condition reported feeling more at ease and comprehended by the robot, as indicated by elevated empathy scores on the Likert scale.

Although no statistically significant conclusions can be drawn, observational comparisons between the multi-modal and text-based conditions indicated that the text-based robot interactions tended to be more effective in adjusting to the subjects' emotional requirements. This was especially clear in the increased empathy ratings and more positive views of the robot's emotional comprehension in the text-based scenario. Although the multi-modal model provided a more enhanced interaction by including visual and auditory elements, it did not notably enhance the robot's capability to recognise or react to emotions in comparison to the text-based model. This brings up significant inquiries regarding the significance of multi-modal interactions in recognising emotions and fostering emotional involvement, indicating that, at least within this experimental framework, the increased complexity of multi-modal inputs did not significantly improve the robot's emotional responsiveness or the user's experience.

In conclusion, the data suggest that while the NAO robot exhibited some ability to detect and respond to human emotions, its effectiveness was limited, particularly in the multi-modal condition. The text-based model showed greater success in emotion detection and empathetic interaction, showcasing the need for further refinement in multi-modal emotion recognition.

5 Discussion

This study researched multi-modal emotion recognition systems and their ability to detect and respond to human emotions. Our findings suggest that while the multi-modal approach added sensory inputs, its performance did not consistently surpass that of the text-only model. We discuss these findings in relation to the three hypotheses we evaluated:

Hypothesis 1: Integrating emotion recognition from a vision model and language models (RoBERTa for text emotion and SpeechBrain for voice emotion) will improve the accuracy (Y) of the robot's emotion detection compared to using a single modality (X).

The pilot study's results do not strongly support this hypothesis. Across 20 robot interactions, and 12 correctly predicted emotions, the multi-modal model achieved fewer accurate predictions (4 out of 12) than the text-only model (8 out of 12). While the multi-modal model showed higher alignment with broader emotional categories (e.g., neutral, positive, negative), it struggled to accurately identify specific emotions. Conflicting outputs between modalities (e.g., cheerful audio cues combined with anxious visual signals) often led to incorrect predictions, highlighting a need to improve fusion mechanisms.

Hypothesis 2: A higher alignment (X) between the emotion predictions of the vision model and language models will correlate with increased user perception of the robot's empathetic response (Y).

Participant feedback indicated limited support for this hypothesis. The final survey showed that participants rated the text-based model higher in perceived empathy and interaction quality. Participants in the text-only condition reported feeling more understood and at ease with the robot's responses. While the multi-model model added input modalities, and hence could make use of more information, this did not translate into more empathetic interactions, potentially due to the inability to reconcile conflicting predictions.

Hypothesis 3: Using specific interaction scenarios intended to trigger different emotions (X) will effectively induce target emotions in the users, and therefore allow for sound measurement of ground truth emotions (Y).

This hypothesis was partially supported. The pictures of the IAPS dataset successfully induced the intended emotional states in most participants, as validated by self-reported emotions and intercoder assessments. However, the robot's ability to correctly predict these emotions was inconsistent. For instance, while *disgust* and *fear* were frequently and accurately predicted, positive emotions such as *happiness* were less often recognised, which suggested that the emotion recognition system is biased toward negative emotions.

Evaluation of Interaction Design. The pilot study revealed that the text-based model outperformed the multi-modal model on perceived empathy and emotional alignment, suggesting that a more parsimonious design, like a text-only model, might have led to better outcomes. Verbal and non-verbal cues need to be integrated more effectively to justify the added complexity. It is also conceivable that the visual model used in the multi-modal system (LLaMA Vision) might be underperforming on the NAO robot due to hardware limitations or suboptimal configuration. Additionally, although the specific emotional labels weren't consistently aligned, the robot's predictions typically fell within the same emotional category as the participants' self-reported emotions, with the multi-modal model showing a slight edge in matching the broader emotional category. This suggests that while the multi-modal model can broadly categorise emotions well, it struggles with pinpointing specific emotions compared to the text-based model. Lastly, we potentially find ceiling effects in the robot's emotional recognition, especially in the multi-modal context, implying that the robot's emotional recognition might already be as good as its underlying models and architecture allow.

Evaluation of Study Design. The experimental procedure and survey instruments were effective in capturing participant feedback. As aforementioned, negative emotions were more easily detected than positive emotions. However, this is likely not due to bias in our stimuli, as the

IAPS dataset is well-validated. Instead, this trend may reflect inherent biases in the detection algorithms, warranting further research into model-based biases. The intercoder agreement score of approximately 0.5 also highlights the challenge in achieving consistent emotional assessments, even among human coders. This indicates the subtle and subjective nature of emotional detection and underscores the need for further refinement in both human and machine emotional assessments. Future iterations should aim to refine detection mechanisms and investigate how these biases affect the model's performance across different emotional categories.

Limitations and Future Research. Our study was limited by the following factors: Firstly, the visual model might be underperforming on the NAO robot. Second, environmental constraints, such as lighting and background noise, impacted the reliability of the visual and audio inputs. Third, the small sample size of 10 participants limits the generalisability of these findings, though some trends provide useful insights. Lastly, as discussed, integration challenges were evident, as the multi-modal system relied on simple aggregation techniques that struggled to resolve conflicts between modalities. These findings highlight the need for future research to optimise integration techniques, address biases in detection algorithms, and explore performance with increased sample sizes.

6 Conclusion

While we successfully integrated the multi-modal emotion recognition and response system, we observed specific strengths and areas for improvement. Multi-modality offered a richer framework for emotion recognition but did not consistently outperform the text-only model in accuracy or user satisfaction. The study highlights the importance of ensuring high performance for each individual model and developing improved techniques for integrating modalities effectively. Despite these challenges, the system highlighted opportunities for leveraging rich multi-modal input to enhance empathetic interactions and ultimately more lead to engaging conversations. Addressing biases in detection algorithms, expanding the range of detectable emotions, and improving environmental robustness remain critical areas for future research. This study illustrates both the potential and the complexity of designing multi-modal empathetic robots, emphasising the need for refined integration and robust individual components to achieve reliable emotion recognition and meaningful interactions.

7 Appendix

7.1 Individual Contribution Summary

Michael: During the project, my main areas of contribution were technical development, research, experimentation, and result analysis. In the first two weeks of the project, I could not complete all tutorials due to a connection issue with the NAO bot, however, I assisted my teammates instead. In the first stages of development Raj and I were tasked to handle the non-verbal cues component of the robot. As Raj already had a working implementation of facial recognition, I moved on to incorporate pose detection. This involved a research phase to discover existing frameworks and implementations for pose detection such as OpenPose and MediaPipe. We eventually opted for the first variant, MediaPipe, as this was geared more towards lower-end hardware and was more suitable for HRI tasks. Thus, for the pose recognition I build a script using MediaPipe to recognize pose landscapes in images. Eventually, MediaPipe was dropped in favor of analysing the picture through an LLM, which Raj handled. I focused on implementing temporal features, such that we could detect gestures. Due to time constraints, our advisor advised us to drop this feature and focus on what we already had, thus I focused on the experiment. Together with Diana and Oskar, we conducted research to find papers and databases revolving around conducting experiments with multi-modal HRI. Meanwhile, when the majority of the pipeline was built, the team and I discussed the flow of the interaction. I then built an interaction diagram to present the high-level visualization of the interaction flow the user would have with the robot and how the robot would process incoming cues and create a response for the user. Afterwards, I collaborated with Oskar to develop the experiment environment script, such that the user would have a graphical interface and would be properly guided throughout the experiment. The script we developed integrated the robot interaction script so that it also ensured that the robot interacted with the participants at the appropriate times. During the experiment, Oskar and I acted as intercoders to observe and note the emotions of the user during the experiment. Furthermore, prior to the experiment, I informed the participants of their participant number and provided details on how the experiment would be conducted to ensure a smooth trial. After we conducted the experiment, Oskar and I performed a data analysis of the results collected during the experiment. I created a script to filter for important data we could use for the result section, such as mapping users to the correct group, indicating to which conversation the data belonged and combining collected conversation data with the questionnaire answers. With the collected data, I extended the script to allow for the visualization of the recorded results. For this task I collaborated with Oskar to ensure that the results I provided could be used for the result section and gave clear insight in the data we had collected.

Anesa: Throughout this course and project, I have focused on several key aspects of the project. The first role I undertook, involved that of a project leader where I created a roadmap with building blocks that we were supposed to work together to tackle our assigned interaction problem alongside with the various roles all of the 7 members of this group can take. Aside from assigning tasks, I have made sure throughout the weekly milestones to stay on track with our objectives and ensured each of us were optimally focusing on tackling a building block and therefore serving as helping hand and task-coordinator. When focusing on the actual codebase framework, I have initially worked alongside Lisann on the speech model integration pipeline, where I have researched and set up emotional speech models (EmotiVoice, OpenVoice, CoquitTTS) and ultimately configured and optimised speech synthesis for integration with NAO via the OpenVoice library. After each of my teammates finished their assigned roles/building block they were focusing on, I worked alongside Raj to ultimately enable a smooth, fast, efficient interaction between the User and Robot. This enhancement was a mini-project itself which lasted until the very end of the course and therefore required me to engage and focus on several existing methods and approaches and carefully integrate them into our pipeline as well as make all of the necessary adjustments and fixes to ensure compatibility with the NAO framework. After intensive weeks of testing and adjustments to reach our final desired interaction with the user for our experimental trial, I finally aided each participant in the experimental trial during their interaction with the robot by inspecting the log output, robot movements and participant behaviour. Lastly, the clean-up of our repository was partially enabled by me through the adaptation of the readme file and the several

merging of branches/codebases to our main branch. In addition, I aided in this report as well were I provided the necessary insights in our Interaction Design Section.

Raj: Throughout the project, I played a crucial role in developing and implementing the technical infrastructure, particularly focusing on emotion detection and system architecture. I began by establishing the foundational setup, including Redis, virtual environments, and the necessary libraries for the SIC framework, while also working on dialogue flow and robot motion. My most significant contribution was developing a comprehensive real-time emotional AI system that could process and understand human emotions through multiple channels. This included creating a custom emotion detection pipeline, deploying vision models on GPU machines, and containerising the application using Docker. I established a serverless architecture using RunPod and integrated various cloud services, ensuring cross-platform compatibility and robust error handling. I had essentially offloaded the much computationally expensive multimodal pipeline of the models as well as text to speech models on cloud gpu. Another challenge was real time transcription for which i had to implement a real time streaming transcription model using google cloud. Since understanding the text was first step in the whole emotion pipeline , we had to make sure this step was achieved in limited time. Dealing with multimodal models would mean to make sure the datatype of each input is maintained as we send it to / from gpu on cloud , which means we had to accurately deal with audio,image as base64 formats. Syncing the whole process was a fun and interesting challenge , especially being able to stich all the emotions from different modalities and allow the NAO to feel emotions felt quite interesting. The final layer to conclude emotional state of the person was done by an llama 3.2 model which was given all the emotional states from different modalities. The challenge was also to make sure that we take care of defaulting behaviour at required points to take care of unpredictibility in LLMs and keep the system stable. Throughout the project , all of the team members showed great enthusiasm and work ethic that resulted in an end to end working pipeline as we had intended while achieving the required performance metrics.

My: During this project, I contributed mostly to the technical part of the project. During the first few weeks, I worked together with Oskar to get to know the SIC framework. This involved installing libraries, connecting to the NAO robot, creating Dialogflow intents, recording motions, and enabling the robot to replay saved motions. While we encountered minor challenges, such as a flipped camera and sparse documentation on motion recording, we resolved these issues by collaborating with teammates. In the following weeks, Oskar and I focused on the conversational flow. Incorporating Whisper for speech-to-text functionality, and experimenting with Gemini for generating responses. We fixed problems like Whisper's early cut-off issues and the timeout errors by increasing timeout settings and handling errors programmatically. Additionally, we worked on text and audio emotion detection to use them in the responses and adjusting prompts for a more casual tone. Although Gemini and Whisper were not the final models we used, I gained valuable experience and knowledge from working with them. Throughout the project, I worked on improving real-time interaction. For instance, when Whisper struggled with the NAO mic, I experimented with alternative microphones and models, ultimately opting for a desktop mic to ensure reliability. I also explored other speech emotion recognition models, but ultimately fell back the to the original model (Speechbrain). Furthermore, I integrated visual feedback using the NAO's LED eye colors to signal turn-taking during conversations. In the final weeks, I focused on refining the experimental setup and fixing critical issues, such as the NAO's microphone loop and response timing. During the experiment I made sure that the participant questionnaires went smoothly.

Oscar: During the SIR project, I have contributed in all aspects of the process: team coordination, coding, conducting the experiment, result analysis, the interaction video creation, and the writing of the report. In the early stages of the project, I made sure that everybody was on the same page when it comes to the coding setup and connection with the NAO robot. During this stage, I worked closely with My on exploring the robot's capabilities and the Dialogflow integration. As dialogflow presented us with limited options, we thought that using whisper from OpenAI would be more appropriate for such a task. Together with My, we constructed the code skeleton for the process of detecting voice and text input in spoken form and integrating it with the robot. We were able to use different models (including gemini llm) and present a working version. When it came time to integrate the different flows of the model, My took initiative and was the main person of the code created to assist others with the integration process and I took a

back seat on the coding aspect. Instead of coding, I moved to setting up the experimental process together with Diana and later Michael as well. I worked on writing the study design and creating a flow that is efficient and would not present potential errors. I kept close contact with the coding team as we required certain aspects implemented for the experiments. Together with Michael, we created a python script as a base for the experimental process which would allow the subjects to interact consistently and it would also allow for the human-robot interaction to be integrated directly into the script and the experiment environment. Together with Diana, I took some footage on the day of the experiment with my camera to showcase our human-robot interaction process with the professors and while presenting our project poster. Michael and I took it upon ourselves to compile the results from the experiments and create insights from them that would shine light onto our research problem. I took initiative in writing the study design and the pilot study results. I worked closely with Michael to curate visualizations for the paper, making sure the paper and the graphs/tables are consistent and good insights are being taken. Overall I am very happy with my contribution as I was able to tap into many different aspects of the project.

Diana: Throughout this course I was involved in a set of different tasks, such as coding, research, organizational tasks, and experimental design. In the early stages of the project, my focus was on completing the weekly coding milestones. This involved ensuring that all necessary software and tools were installed and configured correctly. As the project progressed, my focus shifted towards more research-oriented tasks. I took the lead in conducting the literature review and writing the introduction of our final paper. This involved selecting relevant academic papers, narrowing down our hypotheses, and defining our research question. In addition to the literature review and introduction, I contributed by writing the "Materials and Setup" and "Analysis Plan" sections. One of my key contributions was designing the final experiment. To do this, I investigated relevant sources to understand how Human-Robot Interaction (HRI) experiments are typically performed. I documented this process thoroughly, and the resulting document, titled "Experimental Setup," is available on our shared drive. As part of the experimental team, I collaborated closely with Michael and Oscar. Together, we built a Python framework to conduct the experiment. I provided the initial skeleton for this framework, which was further enhanced by Michael and Oscar. During the final experiment with participants, I played an active role in assisting with the trials. I researched and compiled appropriate inter-coder questionnaires and metrics, which I then printed and distributed among Michael and Oscar. Once the experiment was completed, I processed the data and computed Cohen's Kappa to assess inter-coder reliability. The materials used in this process are also available on our shared drive. When it comes to more organizational activities, throughout the project, I maintained regular communication with our assigned Ph.D. student, Sander. This involved discussing meetings and exchanging emails as needed to ensure that we were on track and aligned with our project goals.

Lisann: Throughout the project, I contributed to both technical tasks and team coordination. In the first weeks, I set up the NAO robot, completed tutorials, and helped select our project topic. I also learned how to use the SIC framework and Dialogflow to run speech-related applications. While I continued working on the tutorials to understand how to play voice files, I also began integrating a speech model, specifically exploring EmotiVoice. I also brainstormed project ideas with the team, noting down my suggestions.
Next, Anesa assigned different parts of the multi-model speech model to us and I continued implementing and testing different text-to-speech models like CoquiTTS and OpenVoice with her. When Coqui didn't fully meet our requirements, I worked with Anesa to switch to OpenVoice. I created the scripts for the TTS model of our final system, and prepared a script integrating all TTS subscripts.
Toward the end, I tested our implementation, added features like conversation transcription, however Raj and I had to resolve these as he had implemented the same. During final experiments, I recruited participants and wrote the discussion and conclusion sections of the report.

7.2 Experiment material

Group	Emotion Trigger	Subject-Selected Option	NAO-Predicted Emotion
1	excitement	Disgust	fear
1	disgust	Neutral	neutral
1	excitement	Fear	fear
1	disgust	Disgust	fear
1	excitement	Fear	anger
1	disgust	Fear	anger
1	excitement	Disgust	fear
1	disgust	Fear	fear
1	excitement	Neutral	neutral
1	disgust	Disgust	fear
2	excitement	Neutral	neutral
2	disgust	Disgust	disgust
2	excitement	Disgust	disgust
2	disgust	Fear	fear
2	excitement	Happy	neutral
2	disgust	Disgust	disgust
2	excitement	Neutral	neutral
2	disgust	Fear	fear
2	excitement	Neutral	fear
2	disgust	Fear	fear

Table 1. Emotion detection comparison table

Group	Text Emotion	Voice Emotion	Visual Emotion	Body Language
1	fear	cheerful	bored	arms crossed/forearms resting on knees
1	joy	angry	anxiety	nervous and fidgety
1	fear	angry	confused	unknown
1	fear	default	anxious	brow furrow
1	neutral	angry	unknown	unknown
1	fear	angry	confused	hand gestures
1	neutral	cheerful	unknown	unknown
1	fear	cheerful	confident	hand gestures associated with expression of ideas
1	neutral	cheerful	confused	casual
1	fear	cheerful	anxious	fingers together in a relaxed fashion
2	neutral	default	unknown	unknown
2	disgust	cheerful	happiness	nervous
2	disgust	default	neutral	smiling
2	fear	cheerful	unsettled	gripping the armrests of the chair
2	neutral	angry	neutral	indifferent
2	disgust	angry	neutral	monotone facial expression
2	neutral	cheerful	anxious	casual
2	fear	angry	bored	head turned away
2	fear	default	nonchalant	eye contact to camera, arms at side
2	fear	cheerful	confused	confident

Table 2. Comprehensive table of Emotion Analysis Across Modalities

Group	NAO-Predicted Emotion	Intercoder 1 - Emotion	Intercoder 2 - Emotion
1	fear	neutral	happiness
1	neutral	neutral	happiness
1	fear	happiness	neutral
1	fear	neutral	neutral
1	anger	neutral	neutral
1	anger	neutral	neutral
1	fear	happiness	neutral
1	fear	neutral	neutral
1	neutral	surprise	neutral
1	fear	surprise	neutral
2	neutral	happiness	happiness
2	disgust	surprise	happiness
2	disgust	neutral	surprise
2	fear	surprise	neutral
2	neutral	happiness	neutral
2	disgust	surprise	neutral
2	neutral	happiness	surprise
2	fear	surprise	surprise
2	fear	happiness	neutral
2	fear	neutral	neutral

Table 3. Comparison of NAO-predicted emotions and intercoder annotations

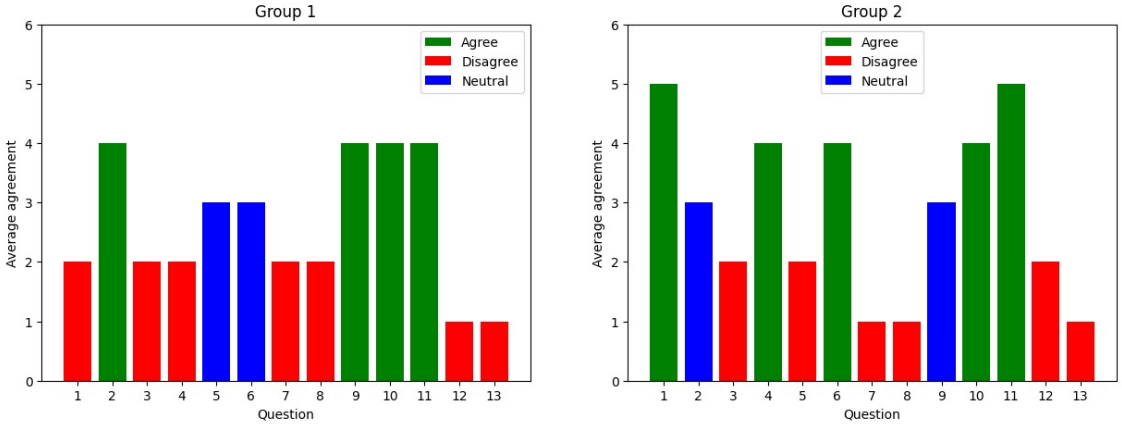

Fig. 4. Results from questions from final survey on robot emotion satisfaction.

References

1. Baraka, M., Velupillai, S.: Emotion recognition from speech: A review. arXiv preprint arXiv:2003.03909 (2020)
2. Barde, A., Chang, C.H., Horvath, R., Oktay, E., Patel, A.: Social robotics: A detailed background and prudential development (2024), https://www.oxjournal.org/social-robotics-a-detailed-background-and-prudential-development, retrieved from OxJournal
3. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
4. Breazeal, C.: Toward sociable robots. Robotics and Autonomous Systems **42**(3-4), 163–175 (2003)
5. Cavallo, F., Semeraro, Q., Sgorbissa, S., Zollo, F.: Emotion recognition in human-robot interaction: A review. IEEE Access **6**, 3287–3306 (2018)
6. Dautenhahn, K.: Socially intelligent robots: dimensions of human-robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences **362**(1480), 679–704 (2007)

7. Hartmann, J.: Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ (2022)
8. Kleinsmith, A., Bianchi-Berthouze, N.: Automated recognition of affect: A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **36**(6), 772–787 (2006)
9. Kumar, P., Malik, S., Raman, B.: Interpretable multimodal emotion recognition using hybrid fusion of speech and image data (2023), https://arxiv.org/abs/2208.11868
10. Leite, I., Martinho, C., Pereira, A.P., Silva, P.: Social robots for long-term interaction: A review. Interaction Studies **14**(3), 375–403 (2013)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
12. Paiva, A., Dias, J., Silva, P., Ribeiro, T.: Artificial empathy: A review. IEEE Transactions on Affective Computing **9**(2), 187–200 (2018)
13. Picard, R.W.: Affective computing. MIT press (1997)
14. Qin, Z., Zhao, W., Yu, X., Sun, X.: Openvoice: Versatile instant voice cloning (2024), https://arxiv.org/abs/2312.01479
15. Riek, L.D.: Empathy in human-robot interaction: A review. International Journal of Social Robotics **9**(3), 333–346 (2017)
16. Rogers, S., Tannen, D.: Conversational pauses and perceived empathy in interpersonal communication. Journal of Language and Social Psychology **30**(2), 177–193 (2011)
17. SpeechBrain: Open-source conversational ai for everyone. https://speechbrain.github.io/ (2023), accessed: 2024-12-05
18. Udahemuka, G., Djouani, K., Kurien, A.M.: Multimodal emotion recognition using visual, vocal and physiological signals: A review. Applied Sciences **14**(17) (2024). https://doi.org/10.3390/app14178071, https://www.mdpi.com/2076-3417/14/17/8071
19. Zeng, Z., Pantic, M., Roisman, G., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(12), 2109–2125 (2009)