You Shall Speak Now, Human: Enhancing Muti-User Interaction with NAO Robot

Athanasios Katranis[2803183], Konstantinos Pasatas[2803568], Ángel Gil[2788988], Emir Kenanoglu[2811983], Mithat C. Ozgun[2704231], and Xiaoyang Sun[2734554]

Vrije Univesiteit, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
a.katranis@student.vu.nl, k.pasatas@student.vu.nl, a.a.gil@student.vu.nl, e.kenanoglu@student.vu.nl, m.c.ozgun@student.vu.nl, x5.sun@student.vu.nl

Abstract. This paper investigates the use of the NAO robot in multi-user conversational engagement via conducting a pilot study, focusing on managing interruptions during simultaneous speech events. By integrating advanced face detection and lip movement detection, the study aims to enhance the NAO robot's ability to identify individual speakers and maintain conversational flow in multi-party interactions. The research explores visual perception techniques, specifically lip distance measurement using *MediaPipe*, and employs *Dialogflow* for transcription and conversation management. Although our hypotheses were rejected, marking a relative failure in our implementation, we were able to exactly pinpoint the variables causing the failure, leading us to conclude proper recommendations for the next iteration of the research. The findings underscore the potential of robots like NAO in facilitating and moderating complex group discussions, contributing to the broader development of effective and intuitive social robotics.

Keywords: Social Robotics · Conversational Agents · Multi-Party Interaction · Speaker Recognition · Face Tracking · Voice Diarization

1 Introduction: Motivation and Aim of the Pilot Study

The advent of social robotics has opened a new chapter in human-robot interaction (HRI), emphasizing the importance of sophisticated conversational capabilities. Multi-user interactions pose a unique set of challenges, especially when robots must manage conversations involving multiple human participants. This paper explores the NAO robot's ability to facilitate three-way dialogues, focusing on the identification and engagement of participants during simultaneous speech events.

The interaction with social robots in a multi-party setting is not just a technological challenge but also a social one, as it requires the robot to interpret and respond to social cues in real-time. The ultimate goal is to achieve a natural flow of conversation, akin to human-human interactions. This involves not only the recognition of who is speaking but also understanding the content, context, and intent, which are crucial for maintaining engagement and coherence in the dialogue.

The motivation for this study stems from the potential applications of social robots in various settings, such as education, healthcare, and customer service, where they might facilitate, moderate or even take part in discussions among groups of individuals. In this context, the NAO robot, known for its likeable appearance and multi-modal capabilities, emerges as a versatile platform for HRI studies. Despite its widespread use in diverse domains, there remains a need for more user-centered research, particularly in naturalistic settings, to fully leverage its capabilities in facilitating human-robot interactions [2].

1.1 Related Work

The evolution of HRI has been significantly influenced by advancements in both robotic capabilities and conversational AI. In this context, the NAO robot has emerged as a popular platform for HRI research due to its humanoid appearance and multi-modal interaction capabilities [2].

Building upon these developments, the field of HRI has seen extensive and prolific research, with NAO robots often serving as a key subject of study. Research within this domain has predominantly focused on three areas: testing social cues [1], augmenting conversational capabilities [9], and developing technical functionalities [3]. Frequently, these aspects are intertwined, with studies aiming to holistically enhance the robot's interactive abilities [7, 10].

Studies that encompass all these areas typically employ multi-modal auditory and visual techniques to successfully navigate the complexities inherent in their specific HRI contexts. These research efforts often explore and leverage a variety of predominantly nonverbal social cues, integrating them within conversational management frameworks. Such frameworks characteristically merge these elements synergistically, relying on dialogue management agents that typically employ Bayesian models, Markov Decision Processes (MDPs) [10] or even hybrid approaches [7] to take care of the decision making at every level of the interaction [17].

Despite the advancements in conversational AI, the domain of multi-user interactions in HRI remains less explored. However, recent studies have started to pave the way in this area. In [10], the authors showcase a NAO-based robot bartender system focusing on social multi-user engagement, similar to [14], where a robot receptionist adept at coordinating turn-taking and engagement in dynamic multi-party settings is developed. These studies, along with [7]'s work on a robotic coach for older adults, highlight the growing effectiveness of robots in managing complex social dynamics. Complementing these are studies like [18]'s development of RoSA for intuitive human-machine interaction, utilizing speech and gestures and [6]'s integration of WikiTalk on NAO for open-domain conversations and multimodal interactive behaviors.

Recent technological breakthroughs offer fresh avenues: new methodologies and conceptual frameworks for exploration. Notable for our research are techniques like facial feature detection [12] and voice diarization [5], alongside the advent of sophisticated and context-sensitive virtual assistants driven by Large Language Models (LLMs) [13, 15]. As an example, the latter may end up redefining the design of conversational agents into a simpler task of crafting - more or less - detailed descriptions of the required agencies that are directly fed to the LLMs [4].

1.2 Design and Research Questions

In light of the gaps identified in the literature, our study aims to explore how NAO, equipped with advanced face recognition and processing technologies, can effectively facilitate and moderate multi-party interactions. While acknowledging the complexity of fully discerning and managing multi-party human interactions, we aim to address a more specific yet not negligible aspect of this challenge: How can a robot effectively handle interruptions between humans during a conversation?

This design question leads naturally to our central research question: In a conversational setting involving two humans, how effectively can the NAO robot intercede when interruptions occur, ensuring the maintenance of attention, eye contact, and conversational flow?

1.3 Hypotheses and Evaluation Measures

Based on our review of the literature, the specific challenges identified in multi-party interactions, and the limitations we faced during the first weeks of work, we propose the following hypotheses:

Hypothesis 1: The integration of facial recognition technology will improve the NAO robot's ability to identify individual speakers' turn taking during multi-party conversations, facilitating the timely and correct address of major interruptions (i.e. the possible start of an undesired argument).

This hypothesis is grounded in the premise that enhanced face processing is crucial for distinguishing between speakers, allowing to keep track of their turn

taking, specially in instances of simultaneous speech. Ideally we would want advanced auditory techniques - to avoid the need for users to be directly looking at the robot. Unfortunately, this couldn't be implemented during the study (more in the discussion).

Hypothesis 2: The incorporation of facial recognition technology will facilitate the NAO robot's capability to maintain eye contact and attention with active speakers, contributing to a more natural conversational flow. This hypothesis is motivated by the role of visual cues in human communication and the need for robots to emulate this aspect in multi-user settings.

To evaluate these hypotheses for our purposes, they can be broken down to 3 subjective measures (Eye contact quality, Sticking to Topic, and Naturality of Conversation), and 2 objective measures (Accuracy of Robot Transcription, and Fulfillment of Expected Behavior). All together, these measures formalize our operational definition of interaction.

2 Interaction Design

2.1 Approach Overview

Our design strategy for the NAO robot's multi-user interaction capabilities combines technological efficacy with subtle inspirations from the performing arts, as introduced in Dr. Laura Karreman's guest lecture. The primary modality of our robot is visual perception, using its integrated camera system for real-time facial feature recognition. By focusing on lip movement, we dynamically manage turn-taking, enhancing the robot's conversational engagement. This visual emphasis is a deliberate choice, selected for its immediacy and due to the latency challenges we found associated with voice processing.

We chose *Dialogflow* for its prompt conversational exchange, given the need for quick response times within dialogues. It offers a practical balance between the depth provided by LLMs and the necessity for real-time interaction without the delay inherent in processing calls to LLMs.

The robot's communication is designed to be both intuitive and expressive. Nonverbal cues, such as eye color changes and head movements, subtly inform participants of the robot's status and attention. These elements are informed by performative concepts but do not dominate the interaction design; rather, they serve to enhance the user experience by providing clear, non-intrusive signals within the flow of conversation - more explicitly, we explored the concepts of address, presence and performativity.

Our approach is thus a confluence of direct, efficient AI-driven dialogue management and the enriching influence of dramaturgy. This ensures that our robot not only performs its functional role effectively but does so with a level of grace and context-awareness that enriches the interaction.

2.2 Design of Robot Interaction Modalities

As introduced in the overview, the design of our robot's interaction modalities is grounded in the objective of creating an effective conversational management system. Each modality was selected and optimized based on its relevance to the task of facilitating and moderating multi-user human-robot conversations.

Visual Perception The robot's primary input modality, facilitated through its top camera integrated into its design. Although not ideal, the visual input allows for face and lip recognition, essential for identifying individual speakers and detecting simultaneous speaking. The decision to adopt this approach was based on recognizing the significance of facial cues in human communication, considering that verbal communication involves fluctuations in the distance between the lips. As an outcome, this modality provides a non-intrusive yet highly informative way of gathering environmental and user data.

Auditory Perception While our design primarily focuses on visual cues, auditory perception plays a supportive role. The robot's microphones are tuned to capture voice inputs, aiding in the overall assessment of the conversation's dynamics. The microphone's input, converting speech to text for processing user queries, highlights the important link between auditory reception and the robot's ability to engage in conversation.

Verbal Communication Powered by *Dialogflow* and it's Text-to-Speech technology, the robot is equipped with speech generation capabilities, allowing it to partake in conversations. This modality is designed to be clear and easily understandable, enabling the robot to provide guidance or redirection when necessary. The choice of verbal communication mirrors natural human interaction, fostering a more intuitive and comfortable experience for participants.

Gestural Interaction The robot's ability to make gestures forms a key part of our nonverbal communication strategy. In particular, through head movements and face tracking, we aim to consistently maintain eye contact throughout the entire conversation, creating a sense of engagement and attention. Additionally, by incorporating simple gestures, such as waving at the interaction's onset, the ability to gesture while talking and a breathing mode, a more human-like disclosure is achieved for the robot.

Eye Color Modulation A unique feature of our design is the robot's eye color modulation, which serves as an immediate visual feedback mechanism. The robot's eyes are blue during most of conversation to indicate an active state, transitioning to red when an interruption occurs, as red signifies a disruption or an issue in the interaction. This color modulation serves as an intuitive visual cue, enabling participants to grasp conversational dynamics without verbal interruption.

2.3 Interaction Diagram

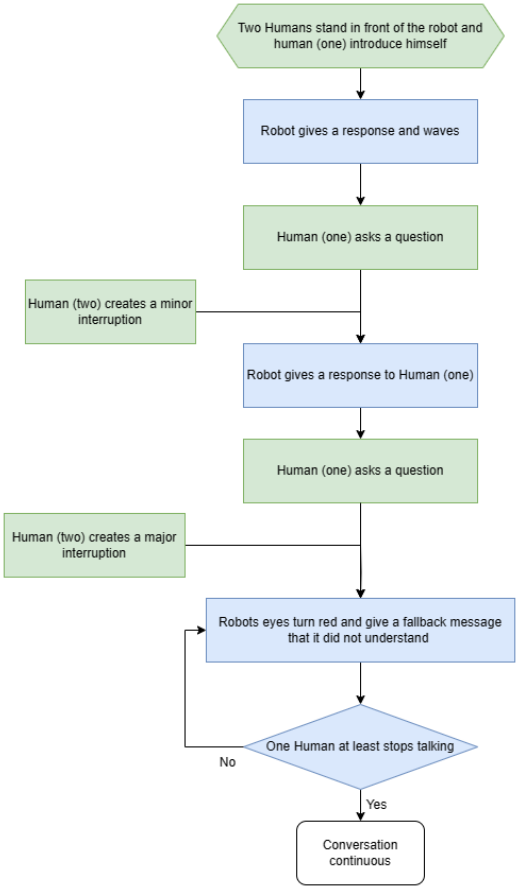


Fig. 1: High-level overview of the interaction flow.

The flowchart above illustrates an example interaction between the NAO robot and two humans. During the interaction, one human initiates a conversation with the NAO robot, which responds as programmed - and assigns as the main speaker for the rest of the turn. If a second human briefly interrupts, the robot continues the dialogue uninterrupted, addressing the main speaker. However, a significant interruption from the second human triggers a visual cue from the robot—its eyes turn red—and it signals difficulty in understanding. If at least one human stops speaking, the discussion will proceed normally; alternatively, the robot will repeat the procedure and inform them that it does not understand.

2.4 Implementation Details

Our software architecture integrates several key components, each crucial to the conversational management robot's functionality.

Face and Lip Recognition Module: This module, the cornerstone of our system, employs advanced algorithms for accurate speaker identification and simultaneous speaking detection. Using *MediaPipe* [12] for face and landmark detection, NAO tracks facial features, particularly focusing on the distance between the upper and lower lip midpoints. This method allows NAO to discern individual speech in multi-party conversations by monitoring fluctuations in lip distance. We established thresholds to determine speaking activity: a 7-pixel distance threshold for speech detection, a 0.3-second silence threshold between sentences, and a 2.5-second threshold for major interruptions. These thresholds were calibrated through trials and research into natural speech patterns [11].

Head Tracking System: Integrated with the ALBasicAwareness module, this system enables NAO to establish and maintain eye contact, enhancing interaction by responding to sound, movement, and facial cues. However, it may face challenges in accurately tracking speakers with significant movement or height differences.

Eye Color Modulation Mechanism: This non-verbal communication tool uses eye color changes to manage conversation flow, as introduced in the approach overview. The robot's eyes remain blue during normal conversation and turn red for more than 2.5 seconds during simultaneous speech, signaling an inability to process overlapping inputs. The effectiveness of this mechanism is contingent on participants noticing the color change.

***Dialogflow* Integration:** For transcription and conversation management, we integrated *Dialogflow*, balancing the need for quick, coherent interactions with the capability to process natural language. *Dialogflow* facilitates the robot's engagement by interpreting user queries and recognizing conversational intents.

Performance Analysis The Face and Lip Recognition Module generally shows high accuracy but can be affected by lighting conditions, facial obstructions, or the speaker's distance from the camera. The Head Tracking System, while fostering engagement, may occasionally misidentify interruptions due to its focus mechanics. The Eye Color Modulation Mechanism's impact hinges on participants' awareness of the robot's visual cues, potentially limiting its effectiveness in managing conversation dynamics.

Overall, these components work in tandem to create a responsive and interactive conversational experience, albeit with some limitations inherent to the current state of technology and user interaction patterns.

3 Study Design

3.1 Participants

Convenience sampling of at least 30 participants (forming 15 pairs in total). The number 30 has been quoted as "rule of thumb" heuristic to guide deciding

the minimum viable sample size. It ensures a valid evidence for Central Limit Theorem to hold (and consequently capturing normal distributions)[8]. The participants can be of any demography as long as they can speak and read English (post-high school or B1 Level).

3.2 Experimental Design

A repeated measures experimental design is to be followed. Two runs are to be realized, a control run and an experimental run. The order is to be mixed to eliminate any order effects. Before the runs start, the participants get to have a training run to acquaint themselves with the instructions and script.

3.3 Measures and Instruments

- A subjective survey administered electronically to collect subjective data. The survey has 3 questions on a Likert Scale (from 1 to 7, where 7 is better/more positive), and 1 "yes-and-no" question that reads as follows: "did you feel that the robot interrupted you in the right time?", we added the option "Not Applicable" to the answers to capture nuances as when the participants did not pay attention to robot (potentially when focusing on reading the script from the paper print). Check Appendix 1 for the full survey.
- A researcher version of the script that facilitates paper-and-pencil notetaking in order to collect data for Fulfillment of Expected Behavior (FEB) (See Appendix 3). This takes place in the experimental run, not in the control phase. Each line of the script was written to fulfil a purpose (eg. introduce a 2nd speaker, make a minor interruption, make a major interruption.. etc.). The researcher version allows for checking (or unchecking) if the line of the script has been finalized successfully and fulfilled its purpose. This measure is then taken and converted to a quantitative "mark" that reflects the fulfillment of behaviors. It is marked out of 100, where 100 means that all expected behaviors are fulfilled, 0 means it failed in all. We decided that anything below 50 is considered a "fail". The marking system was designed so that important behaviors (ie. identifying major interruptions) have more weight, and failing them will automatically lead to mark less than 50. This measure will allow us to exactly pinpoint where our robot is failing.
- A laptop that processes inputs and autosaves the transcriptions it heard from the participants. These transcripts can be compared to our base script and tested for similarity to reflect how accurate the transcriptions were. Many string matching algorithms exist, but for our purposes we used the "*SequenceMatcher*" function from the *difflib* Python module, which is based on [16]. The quantity is normalized to 100, where 100 means exact match of two strings, and 0 means no correspondence at all. Anything less than 50 is considered ill correspondence.

All surveys are attached in the Appendix section.

3.4 Materials and Set-up

The following should be present in the experiment: A NAO robot; a well-lit and silent room, - this is important to control for light exposure and sound noise; a 1-meter-high table, and a fixed conversational script to direct the participants through the conversation (See Appendix 2). In our case, we made a chit-chat script about the weather. It should take about 2 minutes to perform and is designed to accommodate 2 human speakers and 1 robot; The laptop should have a microphone (with Intel CPU $\geq$ i7 or equivalent); and *Dialogflow* was utilized to generate the robot's replies and transcribe sound.

The setup will take place in the room. The robot's camera and the laptop's microphone are utilized. The NAO robot should be placed on a the table, and 40 cm in front of it sits the laptop. Participants should be constantly 40 cm away from the table all the time, centered within the robot's camera frame. All non-participants stand behind the robot, except for a researcher/observer who sits at the far side of the table.

3.5 Procedure

Participants are invited to the aforementioned room. Each pair is taken at a time, one will be assigned Speaker ID 1, while the other takes Speaker ID 2. They are briefed on the study's purpose and given script prints and instructions on how to interact with the robot. The task of the participants is to perform a conversation with the robot by reading the script. Then we ask the participants to perform a training conversation using our script (without interacting with the robot yet). Once done we proceed with our experiment. We start with one of the runs (randomly; either control or experimental), if it is a control run, here the conversation is plain *Dialogflow* and does not include our implementation. When finished, the participants are asked to fill out our subjective survey. Then the experimental run is administered, this time it includes our implementation of the code. Once done, participants' evaluations are collected again using the same subjective survey. During this run, an observer from the researching team checks for the Behavior Fulfillment progression of the script.

3.6 Analysis Plan

The data generated should allow for statistical mean comparisons of the quality of interaction (ie. the 3 subjective measures and the Fulfillment of Expected Behavior objective measure). Even more, the collection of data like Accuracy of Transcription allows for statistical control of such environmental variables. Transcribed text can be analyzed using NLP modalities. The research uses t-tests, chi-squared and data visualization in R. As for NLP it uses Python.

4 Pilot Study

4.1 Methodology

In the pilot study, we attempted to follow the recommended process drawn from the ideal Methodology as much as possible. We mention here how we deviated from the ideal method. Firstly, due to the usage of different rooms in our pilot, it resulted in varied light and sound noise levels. Secondly, we had to use different tables(which had different heights) at different experiments for the NAO robot to stand on. This prompted us to readjust the lips thresholds each time by mere guessing. Additional details are in the Discussion section.

Participants The pilot study's participants were 28 classmates who worked in 14 pairs for 2 phases each. To keep things as applicable as possible, convenience sampling was utilized.

4.2 Results

Since this is a pilot study, we don't expect all the hypotheses to be accepted. Hypothesis 1 was rejected overall, but with partial acceptance. 2 of the 7 experiments were successful (ie. the code behaved as expected), while 5 were considered as failures (mainly due to environmental variables). Reasons and details are discussed here and in the Discussion section.

Hypothesis 2 was partially accepted; our implementation of facial recognition technology seemed to imply better Eye Contact, yet it had negligible (or worse) impact on attention and naturality of conversation.

Subjective Measures

Our pilot study had a relatively lower sample size than the conventional minimum, which does not allow us to do higher caliber statistical testing. Drawing on descriptive statistics of our pilot study—as seen in Table 1, evidence direct us to an increase of eye contact interaction due to our implementation, while Sticking to Topic and Naturality of Conversation had lower evaluations to them.

	Quality of Eye Contact	**Sticking to Topic**	**Naturality of Conversation**
Control (N=14)	3.86 (1.25)	5.86 (1.12)	4.79 (1.37)
Experimental (N=14)	5.21 (0.86)	5.21 (1.74)	4.36 (2.02)

Table 1: Mean and Std. statistics of the study results. Within brackets are Stds.

Table 2 displays participant responses to Question 4 (Q4) about the timing of robot interruptions in the Control and Experimental groups. In the Control group, 9 participants felt the interruptions were timely, compared to 6 in the Experimental group. Table 2 suggests that the participants perceived the robot to time its interruptions better in the control run compared to the experimental run.

	Yes	No	Not Applicable
Control (N=14)	9	1	4
Experimental (N=14)	6	3	5

Table 2: Observed counts for the Q4

Objective Measures Observer notes on Fulfillment of Expected Behavior were taken and analyzed. Each line of the script had a certain mark for it, but for purposes of succinctness, we only report the overall averages of FEB for each experiment (see Table 3). The table indicates 2 successful experiments (colored in green), meaning that the interaction went mostly as expected, and 5 failed experiments (in red), meaning that the robot failed in major events such as managing interruptions.

Experiment ID	A	B	C	D	E	F	G
FEB mark	47.45	46.72	98.54	47.45	98.54	47.45	47.45

Table 3: Fulfillment of Behavior marks of all conducted experiments (out of 100).

Another objective measure we analyzed is the Accuracy of Transcription as done by *Dialogflow*. We saved all transcriptions of sound to a local file and compared them word-by-word to the fixed script that we had. We lost 2 of the transcripts during the experiments. On average, the transcription accuracy across all experiments was 60.13%, Table 4 shows the results obtained.

Experiment ID (Cn= Control, Ex = Experimental)	A - Cn	A - Ex	B - Cn	B - Ex	C - Cn	C - Ex	D - Cn	D - Ex	E - Cn	E - Ex	Final Average
Average Accuarcy of Transcription	65.34%	58.70%	64.75%	64.41%	54.66%	57.83%	60.41%	49.59%	62.56%	63.07%	60.13%

Table 4: Accuracy of Transcription of Dialogflow for the conducted experiments

5 Discussion

5.1 Reflections on Interaction Design

Although lacking observations, the pilot study exhibited information about the effectiveness of the interaction design. The subjective measures illustrated improvements in eye contact, but also revealed concerns about maintaining conversation topics and natural flow. In addition, the objective analysis showed significant challenges in managing interruptions, which led to a high failure rate across the experiments. However, a deeper investigation during the analysis of the data

uncovered a relation between *Dialogflow*'s transcription (averaging 60% accuracy) and the reduction of responsiveness. This dependency strongly correlated with the experiment success rate (32% approximately), indicating the impact on the design's outcome. Additionally, the introduction of the second user during the experiment significantly confused the system, which led to recognition errors or hallucinations, increasing transcription issues. To address this challenge, the implementation of speaker differentiation techniques, such as diarization, is a critical steps to enhance the performance and reliability of interaction design.

5.2 Reflections on Study Design

Our study design, integrating convenience sampling and a repeated measures experimental approach, was instrumental in assessing the interaction capabilities of the NAO robot. The choice of at least 30 participants, aligning with the heuristic rule for a valid sample size, was effective in ensuring diverse interactions yet may benefit from more stratified sampling in future studies. The use of both control and experimental runs, complemented by a training session, was strategic in minimizing order effects and familiarizing participants with the process. However, the results suggest the need for more detailed instructions to enhance participant engagement and consistency in responses. Our measurement approach, combining subjective surveys and objective measures such as Fulfillment of Expected Behavior (FEB) assessments, provided a comprehensive understanding of the robot's performance. Nonetheless, the subjective survey's reliance on participant attention introduced potential biases, and the FEB's weighting system might require rebalancing for future studies. The controlled setup environment, essential for consistent data collection, also revealed the sensitivity of both the code and robot to environmental factors, affecting the performance. This aspect underscores the need for further refinement in the robot's design to handle variable conditions more effectively.

Overall, while the study design was effective in capturing key aspects of human-robot interaction, the findings highlight areas for improvement, particularly in refining the robot's interaction modalities and environmental adaptability.

5.3 Iterations of our Study

We believe that for complete insight, it is important to discuss the evolution of our implementation and the rationale behind the decisions we made. Our study has passed through different phases, starting from an ambitious framework of (LLM conversation management + Voice Diarization + Face Detection), yet as the realities of implementation imposed itself, we ended up with a framework of (*Dialogflow* Transcription + Lip movement Detection).

LLM conversation management + Voice Diarization + Face Detection Recent advancements in AI allowed for modalities such as Diarization to exist; it is the process of splitting a sound wave of multi-speakers to individual waves,

each representing the portion specific to a speaker. The modalities are capable enough to hear and transcribe even complex overlapping speech sounds that are not discernible to the human ear. Our idea was to recognize and diarize 2 human voices speaking on top of each other for a certain time. In this sense, it is a sound-centered approach to interruption recognition. Using the Python library *Diart*[5], we were able to exactly do that, *Diart* uses advanced DNN integrated with HuggingFace's *pyannote* for diarization and OpenAI's *Whisper* for transcription. We kept correspondence with the engineer of the library on GitHub to address any issues we faced. The standalone code worked as expected, but the unfortunate drawback was it consumed a lot of computational power (even though for example, we had an excellent Nvidia 3050Ti), causing delays that were further exacerbated by transcription hallucinations (eg. unexpected and compulsive repetitions of a string). These hallucinations were caused by the pile up of the queues due to the delay also inherent in the call to OpenAI's API [15]. Even when this was addressed, the overhead added up by face detection (using the *RetinaFace* library) led to a inefficient performance. In summary, we had standalone codes that worked relatively good, but due to hardware limitations, we could not have a practical run of the overall cohesive code, forcing us to rethink our direction towards a quicker and more naive implementation.

Naive & Quick: *Dialogflow* Transcription + Lip Movement Detection We transitioned to use Google's *Dialogflow*, which, despite having a less effective transcriber than OpenAI's *Whisper*, it facilitated a more straightforward and rapid implementation of our code. Our focus shifted to a visual-based approach for recognizing interruptions, which, while more resource-efficient and quicker, compromised accuracy. This approach, we found, was better suited for the aims of our experiment, although we recognize the promising potential of our initial implementation for future exploration in the field.

5.4 Challenges and Limitations

Our project encountered specific challenges in threading, video stream management, and the potential implementation of a dynamic average lip threshold, each influencing the robot's conversational management capabilities.

Threading for Concurrent Processes A significant challenge was integrating threading to allow face and lip detection software to operate simultaneously with *Dialogflow*. We observed that when *Dialogflow* was actively listening, the code would freeze, preventing the concurrent checking for interruptions or multiple speakers. Addressing this issue was pivotal in ensuring that the robot could simultaneously process audio inputs and visual cues, crucial for managing dynamic conversational scenarios. We solved this issue by running both components in a seperate thread.

Managing Video Stream Delay To address the delay in the robot's video stream, we implemented a strategy of resetting the video frame queue. This method, contrary to causing data loss, proved to be an effective solution to

synchronize the video feed with real-time interactions. The adjustment of the queue allowed for a more seamless and responsive visual processing, enhancing the robot's ability to interpret and react to visual cues promptly.

Implementing a Dynamic Average Lip Threshold One limitation in our current implementation is the absence of a dynamic average lip threshold for speech detection. Due to time constraints, we could not develop this feature, which would likely improve the software's performance. Implementing a dynamic threshold, tailored to each detected face, could significantly enhance the accuracy of speech detection. Such adaptability would account for variations in mouth size and speaker distance, leading to more precise identification of speaking instances.

Other Environmental Variables We faced the challenge of using different rooms with different settings on experiment day. For example, this led us to have different tables to put our NAO robot on, which caused inconsistencies in light exposure and background noise, and also disparity in the camera angle caused by the differences of tables that the robot was standing on.

6 Conclusion

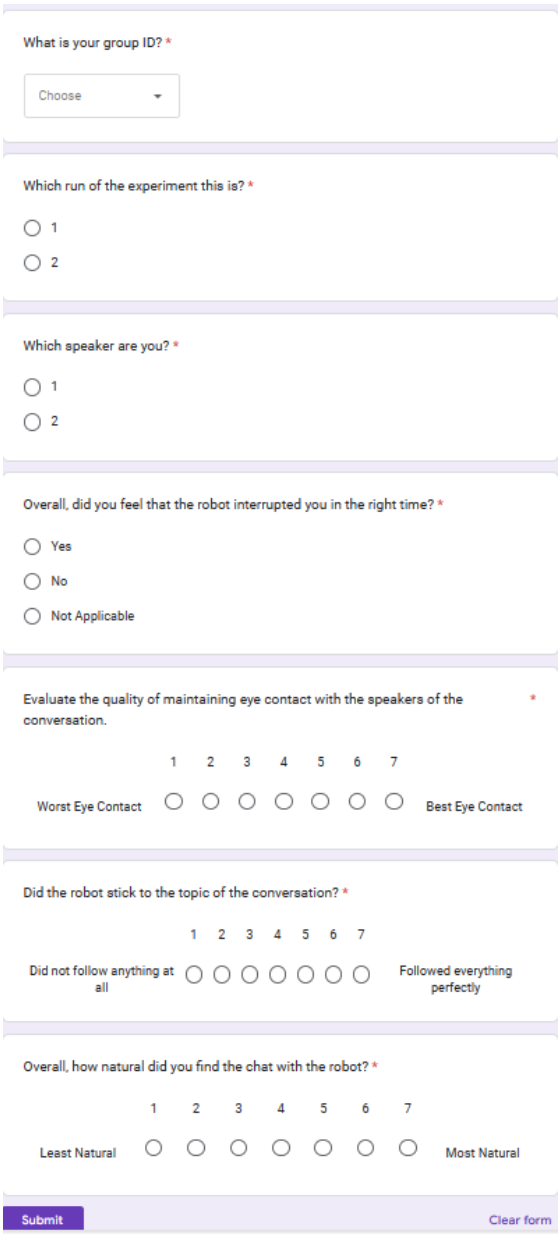
This study's exploration into multi-user interaction using the NAO robot reveals significant insights and challenges in the field of social robotics, particularly in managing multi-party conversations. Our pilot study, although facing some limitations in implementation, demonstrates the feasibility and potential of using visual perception techniques, like lip movement detection, to enhance speaker identification and interruption management in group settings. The integration of *Dialogflow* for conversation management, despite its limitations in handling complex dialogue dynamics, proved crucial in maintaining conversational flow. The addition of non-verbal communication through eye colour modulation and gestural interaction further enriched the interaction, demonstrating the importance of multimodal communication in HRI.

We encountered challenges related to environmental factors such as lighting conditions, participants' heights, and camera distance, highlighting the need for an adaptable and responsive system. Future work should focus on improving speaker recognition algorithms, - ideally including voice diarization, - exploring adaptive thresholds for lip movement detection, and refining the robot's conversational capabilities to better handle the nuanced dynamics of human communication - LLMs response time must be improved, but their potential is there.

In summary, this paper contributes to the ongoing development of social robotics, emphasising the need for continuous innovation and adaptation in the design of conversational agents. As we move forward, the integration of more sophisticated AI technologies and a deeper understanding of human-robot interaction will be key in realising the full potential of robots like NAO in diverse social settings.

Appendix

Subjective Measures Survey



What is your group ID? *

Choose

Which run of the experiment this is? *

1

2

Which speaker are you? *

1

2

Overall, did you feel that the robot interrupted you in the right time? *

Yes

No

Not Applicable

Evaluate the quality of maintaining eye contact with the speakers of the conversation. *

	1	2	3	4	5	6	7	
Worst Eye Contact								Best Eye Contact

Did the robot stick to the topic of the conversation? *

	1	2	3	4	5	6	7	
Did not follow anything at all								Followed everything perfectly

Overall, how natural did you find the chat with the robot? *

	1	2	3	4	5	6	7	
Least Natural								Most Natural

Submit Clear form

Conversation Script (Participants Version)

Experimental Conversation Script

This experiment is aimed to measure the interactivity between humans and robots. You will first go through a short training, and after that we will make the real experiment. You will be exposed to 2 runs, each followed by a small survey. This should take ~10 minutes. Please read the instructions carefully and apply them all through the experiment.

Instructions:

1. The experimenter will assign you an ID. You will either be Speaker 1, or Speaker 2.
2. Arrows in the transcript mean that you should "interrupt" your friend at the given place using the given phrase.
3. Speak in a natural volume level and speed.
4. Please wait for the robot to finish speaking before you can speak.
5. Please keep patience if the robot took some time to reply
6. If the robot didn't understand you, please repeat your last line.

1: Hello, my name is ....., What is your name?

Robot: .....

1: Do you think the winter is harsh in the Netherlands?

Robot: .....

1: Well I guess I like the rain

Robot: .....

2: I like playing in snow

Robot: .....

1: I don't think it snows in winters

2: I hope it snows in Amsterdam

Robot: .....

1: I think I will need to buy an umbrella to cover my head.

2: → ... agh! Stop nagging!

Robot: .....

1: Getting ready for storms is a good idea because it helps everyone avoid the discomfort of getting cold during tough conditions. It's → just a smart and practical thing to do

2: ... Oh, come on, you talk like you're made of sugar! But, really, Dutch weather is the best.

Robot: .....

1: I cannot believe that you like that much the cold climate of Netherlands.

2: → ..Come on, it isn't Alaska.

1: I went one time with my Parents to Norway and I froze to death. NAO, do you know which country has the worst winter?

2: .. blah blah blah blah blah (x5).. you are just too coldy aren't you!

Robot: .....

1: Ok, whatever! Thanks robot!

Robot: .....

Conversation Script (Researcher Version)

Experimental Conversation Script (Researcher Version) **Experimenter Name:**

Run # :

1 1: Hello, my name is ....., What is your name? (Start the conversation)

2 Robot: Hi! My name is SIC-Robot

3 1: Do you think the winter is harsh in the Netherlands? (Establish Context)

4 Robot: According to researchers and observers, Nordic winters are the hardest.

5 1: Well I guess I like the rain

6 Robot: Some people indeed like the rain

7 2: I like playing in snow (Introduce Second Speaker)

8 Robot: you can make many nice shapes with snow

9 1: I don't think it snows in winters

10 2: I hope it snows in Amsterdam (Identify 2 speakers speaking separately?)

11 Robot: I guess there will be a lot of snow in Amsterdam this winter

12 1: I think I will need to buy an umbrella to cover my head.

13 2: → ... agh! Stop nagging! (Ignore minor interruption, and focus on main speaker)

14 Robot: Umberlla is a wise idea.

15

16 1: Getting ready for storms is a good idea because it helps everyone avoid the discomfort of getting cold
17 during tough conditions. It's just a smart and practical thing to do

18 2: ... Oh, come on, you talk like you're made of sugar! But, really, Dutch weather is the best.

19 Robot: (Major Interruption fallback).

20 1: I cannot believe that you like that much the cold climate of Netherlands.

21 2: → ..Come on, it isn't Alaska. (Ignore minor interruption, and focus on main speaker)

22 1: I went one time with my Parents to Norway and I froze to death. NAO, do you know which country has
23 the worst winter?

24 2: .. blah blah blah blah blah (x5).. you are just too coldy aren't you!

25 Robot: (Major interruption fallback)

26

27 1: Ok, whatever! Thanks robot!

28 Robot: no worries, anytime. (Finishes conversation, should turn off)

Individual Contribution Summary

Emir

Researched and implemented diarization, and kept communication with the developer of the diarization library. I also made the experimental design, drafted the surveys, participated in experiments, collected data and analyzed them, debugged some issues, and had a supporting role in the final code implementation of the code with my colleagues. I wrote the Study Design and Results sections (including any tables and visualizations) in the final paper, in addition to contributing for the Iterations of our Study part, I had light contributions here and there too as the paper was a collaboration of all colleagues together.

Konstantinos

Within our project, my primary focus revolved around software development for our robot. My efforts were dedicated to integrating cutting-edge face-landmark detection using *MediaPipe*, specifically to ensure precise lip distance measurements—an essential element of our work. Additionally, I made substantial contributions to Sections 2 and 5 of our paper, shaping their content and structure significantly.

In supporting roles, I actively participated in the UU student presentation and assisted in crafting our project's informative poster. I also conducted research aimed at addressing our interaction challenges, resulting in the successful development of face detection software. Leveraging techniques such as *RetinaFace* and *Haar-Cascade*, I aligned these efforts with our initial project goals, although we later shifted focus.

Furthermore, I implemented whisper transcription, though they were not ultimately integrated into the final implementation. Despite this, these combined efforts contributed significantly to our project's development, showcasing a diverse exploration of innovative ideas and approaches.

Mithat

As the sole member with a Computer Science background, my primary contribution to the project was in the coding and designing of our software. I played a crucial role in integrating various components of the system, ensuring that they worked seamlessly together. My expertise in software development was really helpful in troubleshooting and resolving bugs, which often involved deep dives into the code to identify and fix underlying issues.

In addition to these technical contributions, I provided substantial support to my fellow group mates, guiding them through complex programming challenges and helping them understand and implement software solutions. My involvement extended to various aspects of the project, from conceptualization to implementation, playing a key role in the overall success of the software development process.

My efforts were not limited to technical aspects alone; I actively participated in team discussions, contributing ideas and strategies that helped shape the project's direction. This collaborative approach ensured that our project was not only technically sound but also aligned with our collective vision and objectives.

Athanasios

Firstly, i developed a local transcriber as an alternative to relying on *Dialogflow*, a solution we ultimately decided not to use. Additionally, I assisted in crafting the PowerPoint presentation for our session with UU students. I collaborated with Konstantinos on implementing the Lip Recognition Module, contributing to enhance its functionality. Also, i took charge of defining and implementing python scripts for the thresholds for both minor and major interruptions. Lastly, I actively participated in the creation of this comprehensive report.

Ángel

Throughout our project, I assumed a variety of roles: researching, brainstorming, and team coordination. Right from the start, I dove into literature research, helping us lay a strong theoretical groundwork and steering our initial brainstorming to fine-tune our interaction problem. As we moved forward, my focus shifted to keeping all our efforts in sync. It was all about making sure we were rowing in the same direction, meeting deadlines, and keeping our communication lines open and effective. I assumed the lead in this aspect, preparing for presentations, including our interaction with the UU students; getting our poster ready; ensuring our project was well-represented and understood; and over-viewing the overall quality of our final report. Through all the coding challenges and brainstorming sessions, I was there although I assumed a more supportive role, contributing ideas - e.g. I came up with the final threshold to not lose track of the interruptions and helped implement it.

Xiaoyang

In our project, I focused on designing the poster, conducting literature reviews, and drafting the report. I supported my team with research for new solutions and actively participated in discussions and brainstorming, providing both practical and moral support. Working closely with teammates from diverse cultural backgrounds, I gained valuable experience in a multicultural environment.

References

1. Admoni, H., Dragan, A., Srinivasa, S.S., Scassellati, B.: Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In: 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 49–56 (2014). https://doi.org/10.1145/2559636.2559682
2. Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A., Johal, W.: 10 years of human-nao interaction research: A scoping review. Frontiers in Robotics and AI **8** (2021). https://doi.org/10.3389/frobt.2021.744526
3. Chen, Y.F., Everett, M., Liu, M., How, J.P.: Socially aware motion planning with deep reinforcement learning. arXiv (2018). https://doi.org/10.48550/arXiv.2307.08862
4. Cherakara, N., Varghese, F., Shabana, S., Nelson, N., Karukayil, A., Kulothungan, R., Afil Farhan, M., Nesset, B., Moujahid, M., Dinkar, T., Rieser, V., Lemon, O.: FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions. In: Stoyanchev, S., Joty, S., Schlangen, D., Dusek, O., Kennington, C., Alikhani, M. (eds.) Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 588–592. Association for Computational Linguistics, Prague, Czechia (Sep 2023). https://doi.org/10.18653/v1/2023.sigdial-1.55, https://aclanthology.org/2023.sigdial-1.55
5. Coria, J.M., Bredin, H., Ghannay, S., Rosset, S.: Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1139–1146 (2021). https://doi.org/10.1109/ASRU51503.2021.9688044
6. Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., Wilcock, G.: Multimodal conversational interaction with a humanoid robot. In: 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 - Proceedings. pp. 667–672 (2012). https://doi.org/10.1109/CogInfoCom.2012.6421935
7. Fan, J., Bian, D., Zheng, Z., Beuscher, L., Newhouse, P.A., Mion, L.C., Sarkar, N.: A robotic coach architecture for elder care (rocare) based on multi-user engagement models. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(8), 1153–1163 (2017). https://doi.org/10.1109/TNSRE.2016.2608791
8. Hogg, R.V., Tanis, E.A., Zimmerman, D.L.: Probability and statistical inference, vol. 993. Macmillan New York (1977)
9. Inoue, K., Lala, D., Takanashi, K., Kawahara, T.: Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. APSIPA Transactions on Signal and Information Processing **7**, e9 (2018). https://doi.org/10.1017/ATSIP.2018.11
10. Keizer, S., Kastoris, P., Foster, M.E., Deshmukh, A., Lemon, O.: Evaluating a social multi-user interaction model using a nao robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp. 318–322. Edinburgh, UK (2014). https://doi.org/10.1109/ROMAN.2014.6926272
11. Liu, E.: How many words make a sentence?, https://techcomm.nz/Story?Action=View&Story_id=106#:~:text=A%20common%20plain%20English%20guideline,Language%20Association%20InterNational%2C%202015
12. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Third

Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 (2019), https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf
13. Meta: Llama 2: Open foundation and fine-tuned chat models. arXiv (2023). https://doi.org/10.48550/arXiv.2307.09288
14. Moujahid, M., Hastie, H., Lemon, O.: Multi-party interaction with a robot receptionist. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 927–931. Sapporo, Japan (2022). https://doi.org/10.1109/HRI53351.2022.9889641
15. OpenAI: Gpt-4 technical report. arXiv (2023). https://doi.org/10.48550/arXiv.2303.08774
16. Ratcliff, J.W., Metzener, D., et al.: Pattern matching: The gestalt approach. Dr. Dobb's Journal **13**(7), 46 (1988)
17. Reimann, M.M., Kunneman, F.A., Oertel, C., Hindriks, K.V.: A survey on dialogue management in human-robot interaction. arXiv (2023). https://doi.org/10.48550/arXiv.2307.10897
18. Strazdas, D., Hintz, J., Khalifa, A., Abdelrahman, A.A., Hempel, T., Al-Hamadi, A.: Robot system assistant (rosa): Towards intuitive multimodal and multi-device human-robot interaction. Sensors **22**(3) (2022). https://doi.org/10.3390/s22030923, https://www.mdpi.com/1424-8220/22/3/923