Tales From a Tin Heart

Yvonne Wang
Simon Fraser University
Burnaby, Canada

Jack Weatherbe
Simon Fraser University
Burnaby, Canada

Thimira Wijepala
Simon Fraser University
Burnaby, Canada

Abstract

In the evolving landscape of human-robot interaction, the challenge lies in enabling robots to engage effectively with humans. Existing robots such as Pepper struggle to match human-like expressions. We wish to found out just how much this impacts their ability to connect with users. Our method makes use of two Pepper robots, one tells stories in a monotone voice while standing still, and the other tells stories in an animated voice while moving and gesturing; the participants' social signals and internal feelings are recorded during the storytelling. This research offers insight into the impact of expressiveness on human engagement, suggesting pathways for enhancing human-robot interaction.

1 Introduction

With the development of AI and the rise of autonomous robots, people are starting to imagine a world where humans live and interact with robots as a part of their daily lives. In fact, people already use AI often in their daily lives; for a few years now people have been using AI Assistants such as Apple Siri and Amazon Alexa to ask questions or even hold short conversations. Now with the rise of large language models like ChatGPT, people are now able to hold longer conversations, with ChatGPT even able to understand context clues from previously sent messages. However, this is still far from the world that people imagine, where physical robots are able to express and understand social cues when interacting with humans. The goal and focus of our research is to see how people react to more expressive robots specifically the Pepper robot [2] and whether there is a significant difference between their reaction and attention span versus a more monotone robot. With a better understanding of how people react to more expressive robots, we are hoping to unfold whether people are more engaged when interacting with expressive robots and if they are not, what is lacking in current robots that make it un-engaging.

2 Approach

2.1 Connecting to Pepper

The key idea of our experiment is having people listen to two stories from two different Pepper robots. One of these Pepper robots will tell a story in a monotone and robotic voice, while the other will tell a story in a more expressive and joyful voice with a lot of gestures to help present the illusion that the robot are more "human". To set up the experiment, we used two Pepper robots, multiple Python files and a bash script to run multiple commands in parallel. To connect to Pepper via Python, we followed a stack overflow page [3], where we used a qi Application to start a session connected to Pepper with the help of Authenticator and AuthticatorFactory objects. From there, we were able to run NAOqi commands on Python with the help of the Aldebaran documentation [1].

2.2 Story Telling

With the help of using NAOqi in Python, we used services like TextToSpeech and AnimatedTextToSpeech to help us convey the story in a monotone and expressive voice, AnimatedTextToSpeech also gives the robot gestures based on what is being said. However, while testing out the voices, we realized that the exaggeration between the two voices was not significant enough to notice, so we tuned the parameters of the speech to construct the two voices. For our monotone voice where we used TextToSpeech, we set the style to neutral, the pitch to 80 from 100 and the volume to 60 from 100; for the expressive voice using AnimatedTextToSpeech, we set the style to joyful, the pitch to 80 from 100 and the volume to 60 from 100.

2.3 Data Collection

To collect data about the person's valence and attention, we use another NAOqi service, ALMood. ALMood's command currentPersonState, is able to use Pepper's front camera to help extract data about the participant's valence and attention by giving a score and confidence percentage for each category at a given time. Using a separate Python file, we ran a while loop for the duration of the story and called the currentPersonState command once every second to get a discrete dataset which we stored in a Pandas DataFrame to use later for results.

We also collected video data for each participant study, using the NAOqi service, ALVideoDevice. Using the service, we subscribed to the front camera and in a similar fashion to collecting ALMood data, we ran a while loop for the duration of the story and captured frames of the video using the getImageRemote command, using these frames we attempted to create a video using the cv2 library that allowed us to create a video by writing images into the output file. However, the videos did not come out great and are a bit sped up so they can not be used as a method for analysis.

Finally, since cameras could only capture the social signals a person is exhibiting, we also asked participants to fill out surveys after each story in order to assess their internal feelings. On the survey, the participants were asked to score

their engagement and enjoyment levels for each Pepper robot, what felt lacking, and what did they enjoy about the experience.

2.4 Running the experiment

First, we had the participants "interact" with the robots. Similar to a Wizard of Oz experiment, we would have Pepper exchange greetings with the participants as if they were actually talking to them. This is done to lessen the novelty effect. We then sat them down and told them that two stories were going to be told, one by each of the Pepper robots, and started the storytelling.

To lessen the carryover effects, we created four run orders.

- Run order 0: story 1 with monotone Pepper, then story 2 with animated Pepper
- Run order 1: story 2 with monotone Pepper, then story 1 with animated Pepper
- Run order 2: story 1 with animated Pepper, then story 2 with monotone Pepper
- Run order 3: story 2 with animated Pepper, then story 1 with monotone Pepper

We alternated the run orders between each participant. Then to create consistency in timings, we made the bash scripts run three Python files at the same time, one for story, one for ALMood data and one for video data. Immediately after each story, the participant would complete a survey.

3 Dataset

From our experiments, we created two datasets, one containing data about external facial expressions and one containing data about participants' internal feelings. Due to some issues with saving data obtained from the Python file, our external data set only consists of 17 participants with two rows per participant, one for their data related to the monotone robot and one for their data related to the animated robot. Meanwhile, our internal feelings data that was collected from surveys, consists of 20 participants with two entries per participant, one for their data related to the monotone robot and one for the animated robot. It is also important to preface, that even though we collected the video data, the frame rate and quality of the video were not ideal, so we decided to scrap the data.

Our external facial expression data is saved as a JSON file, where the keys are the "mood_data_id_r", where id is the participants' id and r is the robot that was used. The values for each key is a list of key-value pairs that is saved as the following: {"Timestamp": time (s), "valence": score [-1, 1], "valence_confidence": score [0,1], "attention": score [-1, 1], "attention_confidence": score [0,1]}.

Our internal feelings data is saved as a CSV file, where the columns are as [timestamp, ID, Pepper, engagement, enjoyment, notes_lacking, notes_enjoyment]. Timestamp is just a string containing date and time, ID is a string containing the participant's ID, Pepper is a string containing which Pepper robot was used (M for monotone, A for animated), enjoyment and engagement are both integers rating their enjoyment and engagement on a scale of 1 to 5. notes_lacking is a string containing what participants taught was lacking in the robot (they were asked to enter N/A, if they thought nothing was lacking) and notes_enjoyment is a string containing what participants enjoyed about the robot (they were asked to enter N/A, if they did not enjoy it).

4 Results

4.1 Valence and Attention

From our external facial expression dataset, as shown in figures 1 and 2, we see that there is almost no difference between the valence and attention of the participants when interaction with the monotone robots vs the animated robot.

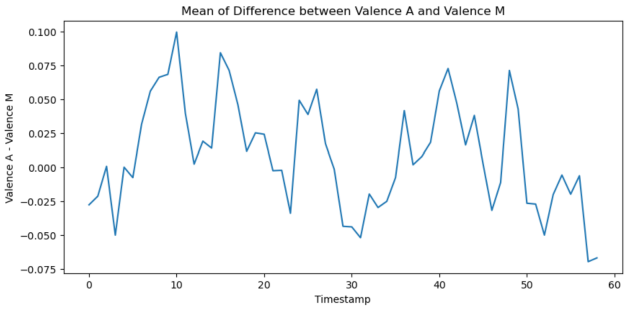

Figure 1. Mean of difference between valences

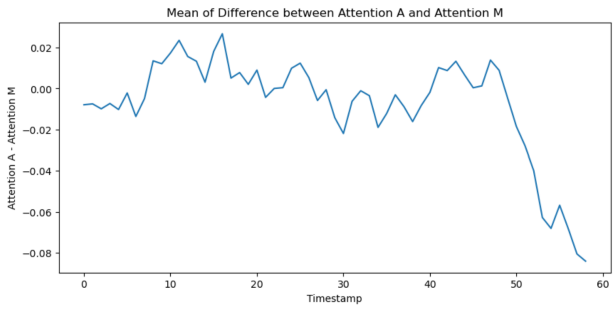

Figure 2. Mean of difference between attentions

Figure 1 shows that the max mean difference between the valences was around 0.1, where participants preferred the animated Pepper vs the monotone Pepper, however, the differences between the two valences were very inconsistent. Meanwhile, in figure 2, the differences between the attentions were more consistent and generally around 0.

4.2 Survey Data

From our internal feelings dataset, as shown in figure 3 and 4, we can see that there is definitely a difference between the engagement and enjoyment scores between the two robots.

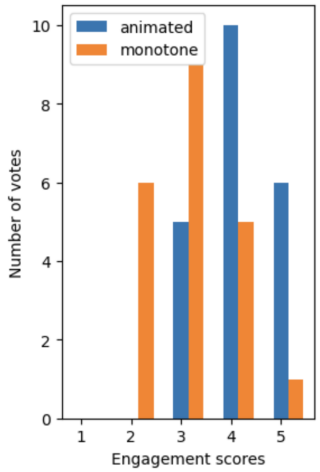

Figure 3. Plot of engagement scores between monotone and animated Pepper

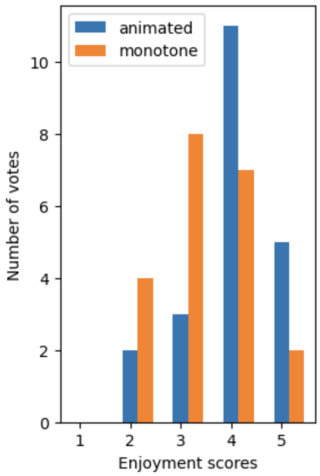

Figure 4. Bar plot of enjoyment scores between monotone and animated Pepper

Figure 3 and 4 shows the bar-plot of the count of engagement and enjoyment scores respectively. The scores for animated Pepper is in blue while the scores for monotone Pepper is in orange. As you can see, animated Pepper has much more higher scores (4 and 5) while monotone Pepper has more lower scores (2 and 3).

4.3 Statistical Analysis and Findings

For our external facial expression dataset, we used Levene's test to test for equal variance, with no significant difference between variance at the $p<0.05$ level. We obtained a p-value of 0.78 for valence data and a p-value of 0.56 for attention data, implying that we can assume equal variance for both the valence and attention data. Using the assumption of equal variance, we can use Student's t-tests to test if the difference in means for valence and attention between Pepper A and Pepper M are different. From the Student's t-test we obtained a $p=0.85$ for valence scores and $p=0.48$ for attention scores implying that there isn't a significant difference between the means between Pepper A and Pepper M for both valence and attention.

We also performed statistical tests to check our findings on the survey data. First, a Levene's test was performed to test for equality of variance. There was no significant difference in variance at the $p<0.05$ level for both engagement and enjoyment scores with $p=0.58$ and $p=0.49$ respectively. After assuming equal variance, we conducted Student's t-tests to compare the engagement and enjoyment scores between Pepper A and Pepper M. There was a significant difference in mean between the two groups at the $p<0.05$ level, with $p=0.00024$ for engagement scores between Pepper A and Pepper M and $p=0.046$ for enjoyment scores between Pepper A and Pepper M.

5 Discussion

From our valence and attention data, we see that there isn't a significant influence on the person's facial expressions based on the robot's expressiveness. This could be due to a few reasons such as, Pepper not being able to capture attention/valence well while the robot and participants are moving, participants not expressing their feelings externally, or maybe the robot's expressiveness is not significant enough for people to show a difference in valence/attention. However, from the survey data it was revealed that the participants did feel engaged and enjoyed the story told by animated Pepper more, even though their social signals did not convey the same.

We believe the most likely reason for this is that Pepper is not "human" enough, so the participants did not feel the need to express their internal feelings outwardly compared to if they were interacting with normal humans. Also due to the nature of the experimental design, people were probably more focused on observing Pepper and completing the survey rather than just acting naturally, which could have affected how reacted externally.

Some people also noted in the survey that Pepper is a bit hard to understand (both animated and monotone Pepper), which could have resulted in fewer outward social signals as they were focused on understanding what Pepper was saying.

We also recorded notes for what the participant felt was lacking and what they enjoyed about Pepper. The notes for what was lacking for Pepper A had 14 out of the 20 participants putting down 'N/A', and the notes for what was enjoyable for Pepper M had 12 out of the 20 participants putting down 'N/A'. Most people filled out the notes for what was enjoyable for Pepper A and what was lacking for Pepper M, which we used to make word frequency graphs

with to see the main reasons they thought that Pepper A was enjoyable while Pepper M was lacking, as shown in figures 5 and 6.

Figure 5. Word frequency graph for what was enjoyable about Pepper A

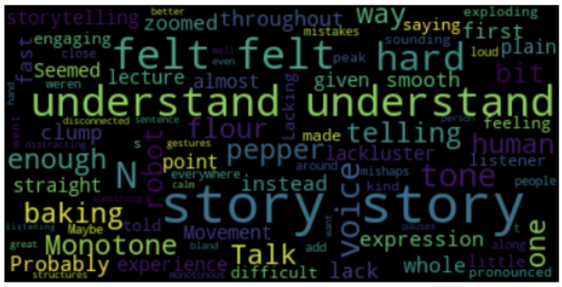

Figure 6. Word frequency graph for what was lacking about Pepper M

We can see that the reason most people liked Pepper A was due to her gestures, movement, and tone of voice, as those were some of the most common words in figure 5. While for why most people felt that Pepper M was lacking, it was mostly due to her being hard to understand and monotone, as those were some of the most common words in figure 6.

6 Conclusion

While our findings did not demonstrate a statistically significant difference in the way people's outward social signals responded to the Pepper robot's different modes of interaction, we did find a statistically significant different in people's internal feelings regarding the Pepper robot's different modes of interaction. The research also provides valuable insights into the challenges of creating truly engaging and emotionally resonant robots. Future research will expand on these findings, integrating a broader range of data to fully unravel the nuanced interplay between human users and robotic systems.

7 Appendix

7.1 Motivation

We created the datasets for the purpose of this research project. It was created by us, Jack, Yvonne, and Thimira.

7.2 Preprocessing, Cleaning, Labelling

For the mood data, we put the raw data into dataframes. For the survey data, some participants misspelled their IDs or wrote them in the wrong casing. We manually overwrote the misspelling and converted everything into uppercase format.

7.3 Uses

The survey and the mood tracking data can be used for further research into storytelling mood-ana in the future.

8 Contribution

- Jack: wrote code for connecting to Pepper, wrote code for Pepper storytelling, ran experiments, performed statistical tests on data
- Yvonne: wrote code for Pepper storytelling, ran experiments, preprocessed data, performed statistical tests on data, visualized data
- Thimira: wrote code for Pepper storytelling, ran experiments, preprocessed data, visualized data

References

[1] [n. d.]. Aldebaran Documentation. http://doc.aldebaran.com/2-5/index_dev_guide.html

[2] [n. d.]. Softbanks: Meet Pepper. https://us.softbankrobotics.com/pepper

[3] 2024. Stackoverflow: How can I connect to Pepper (NAOqi 2.9) via libqi-python. https://stackoverflow.com/questions/77987028/how-can-i-connect-to-pepper-naoqi-2-9-via-libqi-python