Continuous Multi-Modal Interaction Causes Human-Robot Alignment

Sebastian Wallkötter
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
sebastian.wallkotter@postgrad.plymouth.ac.uk

Michael Joannou
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
michael.joannou@postgrad.plymouth.ac.uk

Samuel Westlake
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
samuel.westlake@postgrad.plymouth.ac.uk

Tony Belphaeme
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
tony.belpaeme@plymouth.ac.uk

ABSTRACT

This study explores the effect of continuous interaction with a multi-modal robot on alignment in user dialogue. A game application of '20 Questions' was developed for a SoftBank Robotics NAO robot with supporting gestures, and a study was carried out in which subjects played a number of games. The robot's confidence of speech comprehension was logged and used to analyse the similarity between application legal dialogue and user speech. It was found that subjects significantly aligned their dialogue to the robot throughout continuous, multi-modal interaction.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces; I.2.9. Artificial Intelligence: Robotics; I.2.6. Artificial Intelligence: Learning

Author Keywords

human-robot alignment; multi-modal interaction; SoftBank NAO Robot

INTRODUCTION

Whether interacting with a child, a colleague or a stranger, it is widely accepted that humans adapt their communication in accordance with their understanding of the listener's knowledge and capability [1]. This unconscious process occurs across multiple channels, and greatly simplifies dialogue production and comprehension [6]. In contrast, many modern domestic robots still use just a single modality which can result in the loss of information, such as context, and less effective communication and irritation for the user.

With emergence of less computationally-expensive computer vision techniques and advances in the field of HRI, modern social robots can utilise some of the many non-verbal forms of communication that come naturally to humans. For example, gesture and gaze comprehension are of particular importance when resolving context in dialogue, as humans often refer to objects and events using these channels. This report explores the extent to which humans naturally align to multi-modal, human-like robot communication.

The strength of this alignment will have implications in the design of future HRI systems. In this report, a study was conducted to understand the strength of human verbal alignment, i.e. adaptation of grammar, vocabulary and speaking style, to a multi-modal social robot by continuous interaction. A SoftBank Robotics, NAO robot was programmed to play games of '20 Questions'. Users would think of an animal and the NAO would work out what the animal was by asking a series of questions. In addition to verbal communication, the robot was capable of relaying information through LEDs in its eyes and ears, and via gestures. Interactions between the user and the robot were then logged over a series of games in order to evaluate if humans automatically adapt to robots even when the robot utilises multiple communication channels.

RELATED WORK

Entertainment is one of the most promising applications of social robots [4]. However, the consequences of fragile, error-prone communication systems in HRI include degraded performance and limited commercial potential [5]. Dialogue performance can be greatly improved through the additional utilisation of non-verbal modalities, as demonstrated by a study involving a storytelling robot [4].

Another application of social-robots is their use as classroom assistants. When designing a robot to aid children in their learning, one would readily assume that the robot, like human tutors, should have social and adaptive behaviour. However, experiments by Kennedy, Baxter and Belpaeme [2] demonstrated that this is not necessarily the case, and it was hypoth-

HAI '17, October 17–20, 2017, Bielefeld, Germany

DOI: https://doi.org/10.1145/3125739.3132599

esised that social behaviour of robots may distract children from their learning tasks.

An important factor to consider in HRI is the way in which users align their verbal communication style to the robot, particularly in error-resolution situations. Oviatt, Bernard and Levow [5] analysed the type and magnitude of linguistic adoption that occurred during human-computer error resolution. They discovered that users adapt to the system in three distinct ways: increasing linguistic contrast, increasing hyper-articulation and suppression of linguistic variability. Further, the researchers also found that the feedback given by the robot had a significant effect on the users' behaviour [3].

IMPLEMENTATION

Hardware

The NAO robot was chosen for this experiment as it has broad functionality and an infant-like appearance that helps to limit any preconceived expectations of its capabilities. It has 25 degrees of freedom to allow the development of a range of physical gestures to complement verbal communication. There are 11 degrees of freedom in the lower body (pelvis and legs) and 14 in the upper body (head and arms). Low level control is updated every millisecond while high level control and sensor data is updated every 20 *ms*. Additional features include two loudspeakers to allow the robot to play audio and speech, as well as four microphones (two at the front of the head and two at the back) to allow the robot to capture the user's speech. Captured utterances are processed using NAO's built-in speech recognition engine (Nuance VoCon 4.7).

Software

Gestures

Blinking is a subconscious form of non-verbal communication in human-human interaction and consequently, prolonged staring throughout an interaction will result in alienation. The NAO has a number of LEDs embedded in its head, with a total of 24 LEDs dedicated to each eye. To maximise agency, LEDs were turned off and on in sequence to imitate human blinking which occurred at a constant base frequency with added random noise. Throughout normal interaction, the NAO's eyes shone white; however, upon comprehending speech with high confidence, the eyes flashed green for one second. Conversely, if detected speech was not confidently comprehended, the eyes flashed red. This modality was designed to play a major role in informing the user how to adapt for alignment. Additionally, LEDs in the NAO's ears were turned on upon detection of sounds above a certain volume threshold and otherwise, turned off. Given that verbal feedback can be invasive to conversation, these LED controls were implemented to provide an intuitive alternative.

Further gestures were implemented by manipulating the robot's joints. Upon receiving an answer from the user, a motion to suggest that the NAO was thinking was selected at random, initiated and coupled with a verbal response. Question-specific gestures were also implemented as well as end-game gestures that represented the NAO's reaction to either losing or winning the game. The advantages of this approach were twofold. Firstly, this approach forced breaks in the conversation and gave the dialogue a more natural pace, closer to that of human-human interactions. Secondly, these motions gave the user some indication of what the robot is doing, namely, processing the answer of the previous question, and indicated that the robot will give a response in a moment.

QiChat

The corpus was outlined within QiChat topic files using a bootstrap method. The resulting system was context-based grammar, and consequently, only a restricted portion of the grammar was available at any particular point depending on the flow of the conversation. This was achieved by dynamically loading and unloading portions of the corpus.

The overall dialogue flow was system initiative but could be switched to short user initiative dialogues upon particular user requests. To encourage the user to stay within the grammar it was decided that in-corpus grammar would be used when the robot was talking. Possible questions the NAO could ask were specified in YAML files along with the grammar for the expected answers. Once the user response to the question was received and recognised, it was passed to a Python script running the game engine.

Each question topic contained a concept for 'yes' and 'no' that allowed the question to be answered in a variety of ways specific to that question. Additional topics were added to handle uncertainty in sentences (e.g. 'I think so'). This was to ensure that such an answer does not result in the disqualification of a possible animal due to gaps in the user's animal knowledge or wrong answers. For instance: NAO: 'Does it fly?' User: 'Maybe'. This would not disqualify the animal 'bird' from the list of potential candidates.

Although animals are not necessarily disqualified due to user responses, the nature of the animal YAML file definitions ensured that some uncertainty is accounted for. Each animal definition was outlined in its own YAML file. The file contained the name of the animal, a short question to be asked when the robot wished to guess the animal (e.g. 'is it a bear?') and a frequency value for each label of each question the robot may ask. This frequency represents the number of times a label, animal pair has been observed in the past. For instance a cat may have a frequency value of '100' for 'it has fur' and a value of '5' for 'it does not have fur'. This adds robustness to the system given an instance when the user says 'no' as the cat is not completely disqualified.

Game Engine

This section describes the robot's internal representation of the game. First, answers to questions were clustered into a finite amount of categories called labels, *L*. A question such as 'Does it fly?' would have two: 'yes' and 'no'. Elements of the robot's corpus were then be mapped onto the according label. Questions were subdivided into two categories: differentiating questions and guesses. Differentiating questions, *Q*, as the name implies, help the robot to differentiate between animals. An example would be: 'Does it have legs?'. Guesses *G*, are yes/no questions, specifically asking for an animal, e.g. 'Is it a cat?'. To win the game, the robot has to ask a guess and detect the label "yes". There was a total of $Q = 17$ differentiating questions and $G = 32$ guesses, one for each animal, making a total of 49 questions. All question labels were combined into a feature space, where each dimension represented the frequency of a label's observation. Each animal was then represented as an element of this space.

To find the animal, it was assumed that the user's animal was in the set of animals, A, known to the robot. This allowed the robot to create a probability distribution over animals, modelling the user's belief. The optimal distribution, P^* would assign 0 probability to every animal except the user's, which would have probability 1. This distribution had to be found by asking questions. Initially however, the robot had no information about the user's animal, thus its prior, $P(A)$, was a uniform distribution over all animals. Given the label of the user's response to a question, this prior could be improved in a Bayesian fashion:

$$P(A|L=l) = \frac{P(L=l|A)P(A)}{P(L=l)} \;, \tag{1}$$

where $P(L=l)$ was the total probability of observing label l as an answer and $P(L=l|A)$ was the probability of observing l, given the currently asked question. Further, $P(A)$ was the robot's current prior and $P(A|L=l)$ was the new, better prior. As an alternative to computing $P(L=l)$ and marginalising over it, one can normalise the result of $P(L=l|A)P(A)$ after computation.

One challenge in using this method alone was that the probability of an animal could never reach 0 exactly. This decreased robustness if the robot's model of an animal differed from the user's. An example would be the user thinking of a squirrel and being asked: 'Does it have two or four legs?'. While the robot may think that a squirrel has two legs, the user may think it has four and thus answer accordingly. This would decrease the squirrel's probability and other animals, i.e. dog, cat, and so on would become more likely. To solve this, the robot's belief was thresholded after each Bayesian update:

$$P_{\text{thresh}}(A) = \begin{cases} P(A) \;, \text{ if } P(A) \geq \frac{0.05}{\sum_A [P(A)>0]} \\ 0 \;, \text{ otherwise} \end{cases} \;. \tag{2}$$

Here, $\sum_A [P(A) > 0]$ counted the number of animals with a non-zero likelihood, scaling the threshold dynamically. This can be viewed as 'discarding' an animal, if enough information had been gathered suggesting another. Not only did this solve the problem of a potential difference between the robot's and the user's model of an animal; tests also showed that this creates robustness against deliberately-supplied false information. For example, if the only remaining animals are a cat or a dog, both of which are equally unlikely to fly, then if the user told the robot that the animal does fly, each animal's probability would decrease initially, but reset after normalisation. This means the robot is mostly unaffected by false information, if enough correct information has been specified beforehand.

Finally, the robot chose its next question depending on how many animals could be discarded on average when asking. It was done by simulating each label as a reply for each question, using above inference method. However, when computing the updated prior, P_{thresh}, the number of times a probability was set to 0 was counted. As the likelihood for a label is known, the expected number of discarded animals per question could be calculated. Consequentially, the question with the highest expected value was chosen as next question.

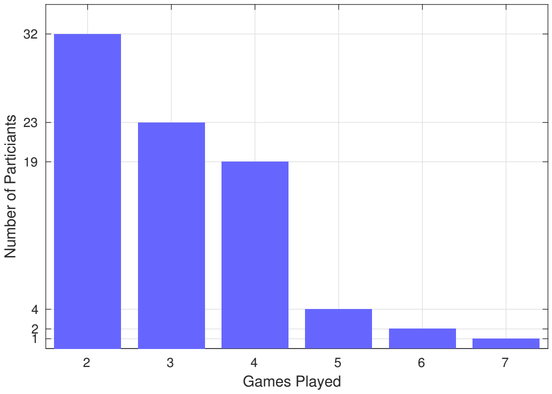

Figure 1. The figure shows the distribution of games played by participants. There was a total of 32 participants. The graph shows how many participants played at least N games.

This setup scales well into the case of an unknown animal. The robot would assign high probability to a known animal sharing the most features with with user's animal. However, since the user will answer the corresponding guess question with label 'no', the robot runs out of animals to consider and concedes.

STUDY DESIGN

The goal of this paper is to answer the hypothesis: 'Does continuous multi-modal interaction cause human-robot alignment?'. A within-subject study was conducted, asking a number of subjects to play a sequence of four games. For each game the subject's verbal alignment to the robot was measured.

In the beginning, the robot would offer an explanation of the game and then start the experiment. This allowed a controlled and repeatable introduction. During the experiment the robot would record all detections of the speech recognition engine together with their confidence. This capture happened automatically and in the background, minimising influence on the subject.

The way the study was set up allowed minimal interaction between the researchers and the subject. This provided consistency across all experiments and minimised the Hawthorne effect as neither observers nor clear recording equipment (camera or microphone) were present. This facilitated authentic or near authentic behaviour throughout the interaction.

RESULTS AND ANALYSIS

Carrying out the study, a total of 32 subjects were asked to play initially. However, many participants could not play four subsequent games, due to time constraints. The distribution over how many games were played by all participants is shown in figure 1. Each of the 32 participants played at least two games,

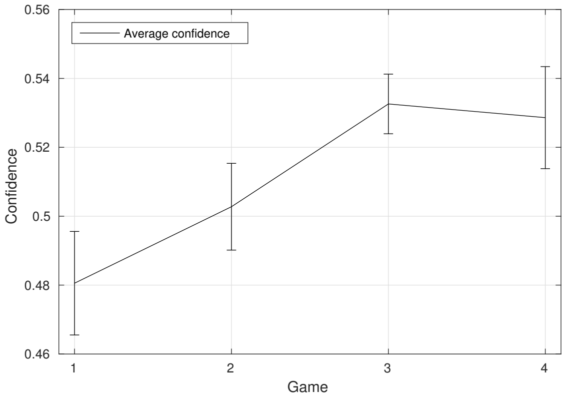

Figure 2. The graph shows the average confidence of the speech recognition system over the number of games played. The error bars visualize the standard error. The confidence increases significantly ($p = 0.018$) over the course of multiple games.

however only a total of 19 played four or more consecutive games.

The group of participants that played four or more games was analysed using ANOVA with repeated measurements. For this, the first four games of each participant were considered. Figure 2 shows the average confidence in each game as well as the standard error. The result shows that the confidence increases significantly over time ($F(3,54) = 3.651$, $p = 0.018$).

To measure the speech recognition's confidence the ASR's (Nuance VoCon 4.7) confidence value was used. This result suggests alignment between the human and the robot.

DISCUSSION

Throughout these interactions, the principal method of indicating if detected speech had been matched to a phrase in the corpus was non-verbal and expressed via the colour of the robot's eyes. In addition, only implicit verbal feedback was given by the NAO as to whether the subject's answer was correctly categorised by the dialogue system. This ensured that no information was given as to the specific content of the robot's grammar. Users acquired knowledge of the NAO's grammar through trial and error only, and therefore, alignment occurred entirely naturally, without explicit instruction from the robot or a researcher.

As seen in Figure 2, the NAO's confidence of comprehension initially averaged at a level below the speech recognition confidence threshold of 0.5. As interaction continued, the average confidence of comprehension increased significantly, and eventually peaked in game number three. This experiment demonstrated that subjects significantly aligned their spoken communication during these multi-modal interactions to maximise the NAO's confidence of comprehension.

A slight, but insignificant, decline in the confidence of comprehension was observed in game number four, seen in Figure 2. The reason behind this is unknown. However, the answer may lie in an underlying compromise between effective dialogue, and speech that is natural to the user. Significant alignment occurred throughout the first three games, at which point the conversation may be considered effective, however, it is likely that subjects suppressed their natural linguistic variability to achieve this. Once sufficiently effective dialogue had been achieved, the users may have begun to slip back into more natural linguistic habits.

The results of this study should be leveraged by designers of social robots. The strong degree of alignment that was observed indicates that subjects quickly built a belief of the robot's capability in order to predict what the robot will understand and, subsequently, tailor their grammar accordingly. Consequently, this implies that small corpora can still result in efficient dialogue, whilst reducing development time. The occurrence of significant alignment implies that the NAO was 'over-promised', a situation that can lead to disappointment for the user. Consequently, this report hypothesises that gradient of alignment can be used as proxy for measuring the degree to which a robot has been over-promised.

CONCLUSION

This study found that subjects automatically strayed from their natural style of verbal communication in order to align their dialogue with that of the NAO robot throughout continuous, multi-modal interaction. This adaptation occurred in the presence of communication through multiple channels, with the NAO relaying information through speech, gestures and LEDs in its eyes and ears. In addition, this alignment occurred in the absence of explicit instruction from the robot or researchers.

The study observed some degree of overshoot when subjects simplified their speech to align with the robot. However, this was not observed with statistical significance. If true, it would highlight the compromise that users make between effective interaction and natural speech.

It is clear that the phenomenon of alignment has positive and significant effects on the effectiveness of dialogue in HRI. This paper proposes that the gradient of alignment could also be used as a proxy to measure the degree to which a robot is over-promised by its appearance. Future study into the possible interaction between rate of alignment and over-promising is recommended.

REFERENCES

1. Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355–2368.

2. James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. (2015), 67–74.

3. Manja Lohse, Katharina J Rohlfing, Britta Wrede, and Gerhard Sagerer. 2008. "Try something else!"–When users change their discursive behavior in human-robot interaction. (2008), 3481–3486.

4. Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. (2006), 518–523.

5. S Oviatt, J Bernard, and G Levow. 1998. Linguistic adaptations during spoken and multimodal error resolution. *Language and speech* 41, 3-4 (1998), 419–442.
6. Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences* 27, 2 (2004), 169–190.