Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions

Martin Johansson, Gabriel Skantze, and Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
vhmj@kth.se, {gabriel,jocke}@speech.kth.se

Abstract. We present a data collection setup for exploring turn-taking in three-party human-robot interaction involving objects competing for attention. The collected corpus comprises 78 minutes in four interactions. Using automated techniques to record head pose and speech patterns, we analyze head pose patterns in turn-transitions. We find that introduction of objects makes addressee identification based on head pose more challenging. The symmetrical setup also allows us to compare human-human to human-robot behavior within the same interaction. We argue that this symmetry can be used to assess to what extent the system exhibits a human-like behavior.

Keywords: focus of attention, human-robot interaction, turn-taking.

1 Introduction

Robots of the future are envisioned to help people perform tasks, not only as mere tools, but as autonomous agents interacting and solving problems together with people. These future robots could interact with humans in a way similar to the way humans interact with each other, or they could make use of more simple, machine-like interaction patterns. Some situations could call for the robot to be a human-like artificial conversational partner capable of participating in a multiparty dialogue. One way of measuring human-likeness is by measuring the behavior of the human interlocutor[1]. The closer the behavior of the human interlocutor in the interaction with the robot to that of interaction with other humans, the more human-like this interaction is.

Aside from being human-like or not, there are some fundamental problems a robot will have to deal with to engage in dialogue with multiple humans. Two of these are to manage turn-taking in order to know when it is acceptable to say something, and to figure out to whom an utterance is directed. Previous studies have found head pose[2] to be a useful indicator to identify the addressee in three-party interaction, especially in combination with acoustic cues[3]. The problems future robots could help solve might however include visible objects competing for attention, possibly affecting head pose behavior.

Edlund et al.[1] proposed to evaluate human-likeness in human-machine interaction using a *two-way mimicry target*. In order for the machine's behavior to

G. Herrmann et al. (Eds.): ICSR 2013, LNAI 8239, pp. 351–360, 2013.
© Springer International Publishing Switzerland 2013

be deemed human-like, it should behave like a human interlocutor in a human-human conversation, and the human speaking to it should behave like when speaking to another human. Any difference in the human's behavior is seen as a result of the machine's behavior. In a multiparty setting with two humans and a robot in similar roles, a two-way mimicry target can be evaluated using, for example, symmetry in behavior between the human-human and human-robot interactions.

This paper presents a data collection setup for three-party conversations with two humans and one robot, designed to allow comparisons between the robot and human participants through symmetry. The purpose of this study is to test the setup by exploring head pose patterns surrounding turn changes when we involve more targets for visual attention than just the participants. We also evaluate the human-likeness of the robot by comparing head pose patterns between human-human and human-robot interactions.

2 Background

The function of gaze in interaction has been found to serve multiple functions, one of them being turn-taking control. Kendon[4] found that speakers look away at beginning of turns, and look back at their interlocutors towards end of turns. Gaze behavior has also been found to provide information about the target of attention. Vertegaal et al.[5] found eye-gaze to be a good predictor of conversational attention in multiparty conversations, while Katzenmaier et al.[3] found head pose to be a cue for identifying addressee in human-human-robot interaction. Stiefelhagen and Zhu[6] showed that head pose is a reliable cue to estimate focus of attention in a small meeting scenario. Ba and Odobez[7] expanded the small meeting scenario to include more targets for attention, and concluded that good separation of targets is essential for accuracy. Automated means of recording gaze in conversational settings are available, using, for example, eye trackers[8], or head pose tracking[7] for estimated gaze. Both methods could conceivably be used by a robot to monitor interlocutors. Estimating gaze through head pose instead of tracking eye-gaze, however, has the advantage of being more robust in regard to head movement and blinking.

Argyle and Graham[9] studied dyadic interactions involving additional targets for visual attention. Objects relevant to the task at hand were found to attract visual attention at the expense of the other subject.

3 Method

We conducted a laboratory experiment in a human-human-robot setup for this exploratory study. The intention was to gain an understanding of symmetry in collaborative human-robot problem solving when objects relating to the task that is discussed are present. In order to elicit engagement in interaction, the task needs to be fun and interesting. One way of achieving this is to use games. The Speech group at KTH has initiated collections of a series of multi-party

games for interactional corpus collections, the KTH games corpora[10], of which the current study is the first human-human-robot corpus.

We selected an adaption of the "Desert Survivor" team-building game[11] as the task in this study. The team is to collaboratively prioritize a set of items based on their usefulness for survival in a desert after a hypothetical plane crash. We chose to use the Desert Survival game since it is an engaging social task that elicits intra-group communication, thus allowing us to study group dynamics and multi-party interaction phenomena.

A *Wizard-of-Oz* setting was selected for initial data collection to give the robot some sense of intelligence in the dialogue via the wizard, while maintaining the physical appearance of the robot and limiting its range of actions compared to a human.

3.1 Experiment Setup

The experiment setup, overviewed in Fig. 1, was designed around a round table at which the two human subjects and the robot were placed. Subjects were seated on static chairs at fixed locations to have the triad placed in an equilateral triangle pattern, with equal distances and angles between all participants. The objective of this placement was to allow for conclusions regarding attention behavior between human-human and human-robot interactions. This setup also has the advantage that the predicted target areas for visual attention are as separated as possible from a head pose rotation frame of reference, as suggested by Ba and Odobez[7]. The complete setup in action is shown in Fig. 2 with a snapshot from one of the sessions.

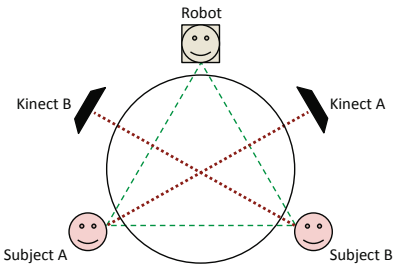

Fig. 1. Spatial configuration

Fig. 2. The robot and two subjects

Each subject was monitored by a Microsoft Kinect sensor, responsible for tracking the head pose of its subject as well as the angle to the most prominent audio source. The use of one sensor per subject limits the expected head pose yaw to around $\pm 30°$ from the Kinect's point of view, avoiding problematic extreme angles yielding low accuracy. Another benefit is the symmetry of the recorded data for both subjects.

High quality audio was recorded using headsets worn by the subjects, keeping the table available for use by the subjects as they saw fit when discussing the items in the scenario. In this adapted exercise, the items to discuss were given physical presence in the form of stylized pictures, each on an individual sheet of paper. Video was recorded using two high-definition video cameras. The first camera oversaw the interaction from behind the human participants, whereas the second one recorded the interaction from the robot's point of view.

Furhat[12] was chosen as the robot participant in the experiment. Furhat is a back-projected human-like robot head using speech synthesis[1] and state-of-the-art facial animation, mounted on a robotic neck. This makes it capable of combining head pose and eye gaze to direct attention[13]. The facial animation architecture allows for speech with accurate synchronized lip movements, as well as for control and generation of non-verbal gestures, eye movement and facial expressions.

3.2 Methodology and Experimental Design

The robot in this experiment was partly automated and partly controlled by a wizard. The gaze of the robot was automated, whereas the speech was controlled by a wizard. The wizard decided when the robot should say something, and what it should say, by selecting one of several predefined utterances from an interface on a networked computer. When speaking, the gaze was directed either towards a subject or, while speaking about an item, towards the table. When not speaking, the robot's gaze was directed towards the participant coinciding with the most prominent audio source detected by the Kinect sensors.

Each session started with the robot giving instructions about the task, placing emphasis on its collaborative nature, before proceeding to the first iteration of items to rank. No instructions concerning the roles of the robot and the subjects were given, other than that it was a team-building exercise, and that they were to discuss and reach a unanimous decision about the ranking of the items. The instructions regarding collaboration and consensus were intended to encourage the subjects to affiliate with the robot as a team[14].

The adapted exercise comprised three iterations of five unique items to discuss and rank, each iteration starting with the robot asking the subjects to open a numbered envelope containing the items. The robot could then, during the discussion that followed, have opinions about the relative importance of two items, say one of two predefined positive or negative things about an item, ask the subjects for their opinions, answer questions with a yes or a no or confess that it did not know.

During the course of the three-party conversation, the goal was to have all interlocutors actively involved in order to collect comparable data for both human-human and human-robot interactions. The wizard's strategy to keep the robot involved in the discussion and, if necessary, both human subjects involved, was to try to answer directed or open questions and to either make a statement or pose a question when given opportunity.

[1] CereProc ltd: `http://www.cereproc.com/`

3.3 Participants

Eight subjects aged 23-41, divided into four pairs, were used in the data collection. The mean age was 30.5, with a standard deviation of 5.42. Three of the participants were female and five male. All subjects were employees at the Department of Speech, Music and Hearing, KTH.

3.4 Measurement

We captured the subjects' behavior using an audio recorder with headsets, Kinect sensors and high-definition cameras. Two Kinect sensors were used to record head positions and rotations in 3D space, as well as a set of face expression parameters and the angle to the most prominent sound source. The video recordings could be used when annotating the dialogues in the future, and the audio recordings can be processed with, for example, automatic speech recognition or prosody extraction.

For this initial analysis, we employed an automated approach where Kinect sensor data was used to estimate target of visual attention based on head position and rotation. The regions of interest as potential target of attention for a subject were the robot, the other subject and the table. No attempts were made to distinguish between different individual items on the table.

We employed automated segmentation using a voice activity detector to extract utterances from the recorded audio. The extracted utterances were then used to locate instances of turn changes. A change of turn was defined as having two consecutive non-overlapping utterances, not shorter than one second each, belonging to two different speakers. Utterances shorter than one second were not included in this analysis of turn changes for robustness reasons. In our case, this means that no change of turn took place; the current speaker continued to speak as the interlocutor with the very short utterance was not claiming the floor.

We recorded twelve iterations, three per pair of subjects. The recorded interactions, excluding instructions, lasted a total of 78 minutes with an average iteration length of 6.5 minutes (standard deviation 1.37). Next, we segmented the recorded audio through automated means, resulting in a total of 1312 segments of human subject speech encompassing 38 minutes.

4 Results

4.1 Observations

All subjects interacted with the items on the table during the dialogues, even though the only instruction given related to the physical items was to open the envelope containing them. The interaction ranged from active spatial organization of the items to signaling which items were considered, for example by picking them up or pointing at them.

4.2 Target of Attention Around Turn Changes

For each change of turn, the participants were labeled with one of the following roles: *current speaker*, the speaker who finished the utterance in the ending turn; *next speaker*, the one who was taking the turn; *other*, the one not speaking. The targets of attention for each subject in a time frame of three seconds before and after the end of the last utterance in a turn were estimated. Head pose data was split into intervals of fifty milliseconds, each with one single target defined by majority classification. The possible targets were either one of the other participants in their current role or the *table*, harboring the items.

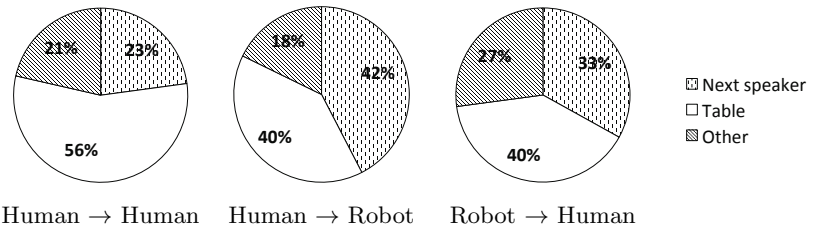

Fig. 3. Visual attention of the current speaker near the end of turn, all instances

We evaluated the overall distribution of the speakers' visual attention in the specified window at end of turns for the different combinations of robot and human speakers (Fig. 3). The robot as the current speaker exhibited an overall distribution of attention resembling the humans', suggesting that the robot did not behave completely different as a speaker. There was no dominant visual attention towards the next speaker in any of the three combinations of interlocutors. The largest share of looking at the next speaker was when the next speaker was the robot.

4.3 Speakers Looking at the Next Speaker

Many turn changes occurred without the current speaker looking at the next speaker. To compare symmetry in the situation where the current speaker looked at the next speaker, we analyzed the instances where the current speaker did look at the next speaker at least once in a time frame spanning one second before to one second after the end of the turn.

First we investigated the estimated target of visual attention for a human ending a turn while looking at the next speaker (Fig. 4). Results indicate that humans address the robot clearly. However, since the robot's decision on when to speak was made by a wizard, the human-to-robot patterns are likely affected by the turn-taking strategy employed by the wizard. Due to the criterion used to select instances, the peaks in attention towards the next speaker around the end of turn were expected.

Next we investigated the estimated target of visual attention for the human taking the turn, when looked at by the speaker who previously had the

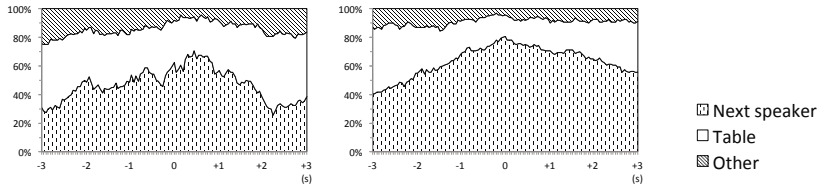

Fig. 4. Visual attention of the current speaker near the end of turn. Instances where the current speaker was looking at the next speaker.

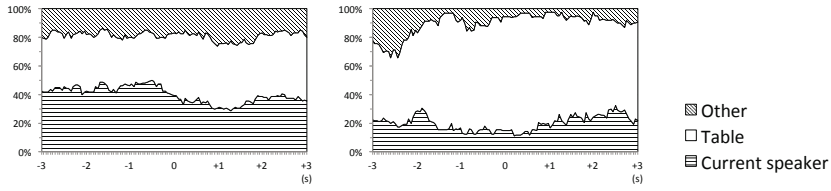

Fig. 5. Visual attention of human taking the turn near the end of preceding turn. Instances where the current speaker was looking at the next speaker.

turn (Fig. 5). The majority of the turn taker's visual attention was directed towards either the current speaker or the table. More visual attention was directed towards the speaking robot than when the speaker was human.

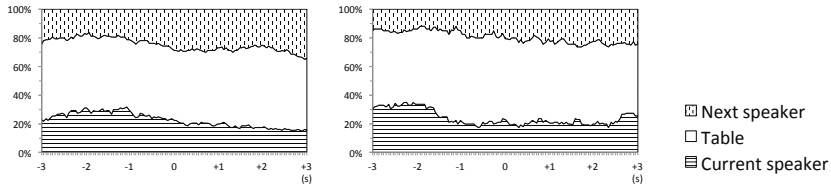

Fig. 6. Visual attention of human neither ending nor taking the turn, near the end of turn. Instances where the current speaker was looking at the next speaker.

Finally we investigated the estimated target of visual attention for the human neither ending nor taking the turn (Fig. 6), when the next speaker was looked at by the current speaker. The distributions appeared more similar in shape, with more attention to the first speaker towards the end of turn and more attention to the next speaker afterwards.

5 Discussion

We analyzed turn changes in the data collected during the pilot experiment, working from the idea that head orientations could be used for estimation of focus of attention as was concluded by Stiefelhagen and Zhu[6].

Introduction of objects makes visual attention tracking less useful for addressee detection and turn change prediction. Like Argyle and Graham[9], we found that objects relevant to the task attract a great deal of attention. Visual attention close to turn changes (Fig. 3) was placed on the table in a large portion of the collected instances. The head pose being directed towards the table in many cases, at the expense of the other participants, indicates that detection of addressee based on the speaker's head pose is more complicated in this situated dialogue than in, for example, the one evaluated by Skantze and Gustafson[2]. When the human speaker's visual attention actually was directed towards the next speaker at some point close to the end of turn (Fig. 4), our data show an increased amount of head orientation towards the next speaker around the end of turn, no matter who the next speaker was. The human speaker looking at the next speaker increasingly towards the end of turn in these situations is in line with the findings of Kendon[4], but the trend may be predisposed by the employed selection criterion.

Human-robot and robot-human turn changes were clearer than human-human turn changes. Comparing the head pose distribution patterns between human-human and human-robot turn changes (Fig. 4), the human-robot distribution was more consistent with fewer peaks, and had the main peak located at the end of turn. The difference could be related to the wizard, as the robot's decisions on when to speak were made by the wizard. It could also be related to the robot, due to, for example, the subjects' expectations on a robot dialogue system, or the robot not making use of visual cues. When the speaker's visual attention was directed towards the next speaker at some point close to the end of turn (Fig. 5), we observed a transfer of the next speaker's visual attention away from the current speaker to the table near the end of turn. This was the case for both human and robot current speakers, albeit an earlier transfer in the case of a human speaker and differing overall proportions. The disparity could be related to differences between the humans and the robot in signaling intentions with, for example, prosody or visual cues, or to differing dialogue strategies or types of utterances.

The behavior of the human speakers was not symmetrical between robot and human interlocutors (Fig. 4 and 5). The human not involved as one of the speakers, on the other hand, exhibited a similar distribution of attention (Fig. 6) close to turn changes regardless of the robot being the first or the second speaker. Visual attention towards the table remained fairly constant, while a transfer of attention from the current speaker to the next speaker took place. Using the mimicry target[1] to define human-likeness, we employ the distributions of visual attention as a measure of human-likeness. This gives us a measurable target to work towards in order to make the robot more human-like in the aspect of visual attention by a human interlocutor. As mentioned, we found differences in the

human speakers' head pose patterns surrounding turn changes when the turn taker was another human, compared to when it was the robot. Thus, we need to modify the robot's behavior in order to bring the distribution of visual attention of its human interlocutor closer to the one exhibited for two human interlocutors. Matching distributions is however only a first step towards human-likeness. The distributions provide an overall view, they do not reveal if any individual actions were human-like or not.

Our observation that all subjects interacted with the objects on the table, combined with a large part of the interlocutors' attention directed towards the table, suggests that exploration of more detailed targets of attention could be worthwhile. A dialogue system could conceivably make use of estimations about which objects the interlocutors are paying attention to. Additionally, comparing, for example, the dialogue acts leading to the change of turn for different targets of attention might also be useful when designing a dialogue system for this setting.

6 Future Work

With the continued goal of exploring the symmetry of turn-taking, the next step is to adjust the robot's behavior to see if we can get it to trigger more human-like turn-taking behavior from humans talking to it, compared to when the humans talk to each other. The long-term goal is to build an automated system capable of improving the symmetry of an ongoing dialogue by adjusting the robot's behavior.

Another interesting goal is to replace the wizard with an autonomous system that appears to be intelligent. One first step in that direction is to improve the robot's ability to deduce where the visual attentions of the interlocutors are. Having a fine-grained sense of attention to task-related object could help provide the robot with valuable insights on the intentions of its interlocutors.

7 Conclusions

In this paper we presented a setup for collecting multimodal data from three-party human-robot interaction, and used the setup to create an initial corpus collected in a Wizard-of-Oz setting. We explored the symmetry of head pose patterns in turn-taking between human and robot participants, finding both similarities and differences. The robot attracted different head pose patterns from humans during turn changes than what humans used between each other. The differences can be useful when implementing a dialogue system, but also indicate a disparity between robot and human interlocutors. In other words, the robot's behavior needs to be changed to make it more human-like in this aspect. We can use symmetry in behavior as a target to evaluate progress. We also investigated the question about implications of introducing task-related objects into the dialogue. Objects attract visual attention at the expense of interlocutors, possibly affecting the usefulness of head pose as an indicator for addressee identification.

Acknowledgments. This work is supported by the Swedish research council (VR) project "Incremental processing in multimodal conversational systems" (2011-6237), as well as the SAVIR project (Situated Audio Visual Interaction with Robots) within the Strategic Research Area ICT - The Next Generation, funded by the Swedish Government.

References

1. Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. Speech Communication 50(8-9), 630–645 (2008)
2. Skantze, G., Gustafson, J.: Attention and interaction control in a human-human-computer dialogue setting. In: Proceedings of SigDial, London, UK, pp. 310–313 (2009)
3. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the addressee in human-human-robot interactions based on head pose and speech. In: Proceedings of International Conference on Multimodal Interfaces, pp. 144–151 (2004)
4. Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychologica 26, 22–63 (1967)
5. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 301–308 (2001)
6. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: Conference on Human Factors in Computing Systems, pp. 858–859 (2002)
7. Ba, S.O., Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(1), 16–33 (2009)
8. Jokinen, K., Nishida, M., Yamamoto, S.: Collecting and annotating conversational eye-gaze data. In: Proceedings of Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Language Resources and Evaluation Conference, Malta (2010)
9. Argyle, M., Graham, J.A.: The central Europe experiment: Looking at persons and looking at objects. Environmental Psychology and Nonverbal Behavior 1(1), 6–16 (1976)
10. Al Moubayed, S., Edlund, J., Gustafson, J.: Analysis of gaze and speech patterns in three-party quiz game interaction. In: Interspeech, Lyon, France (to appear, 2013)
11. Burgoon, J.K., Bonito, J.A., Bengtsson, B., Cederberg, C., Lundeberg, M., Allspach, L.: Interactivity in human-computer interaction: a study of credibility, understanding, and influence. Computers in Human Behavior 16(6), 553–574 (2000)
12. Al Moubayed, S., Skantze, G., Beskow, J.: The Furhat back-projected humanoid head - Lip reading, gaze and multiparty interaction. International Journal of Humanoid Robotics 10(1) (2013)
13. Al Moubayed, S., Skantze, G.: Perception of gaze direction for situated interaction. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. The 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA (2012)
14. Nass, C., Fogg, B.J., Moon, Y.: Can computers be teammates? International Journal of Human-Computer Studies 45(6), 669–678 (1996)