The Power of a Glance: Evaluating Embodiment and Turn-Tracking Strategies of an Active Robotic Overhearer

Spyros Kousidis and David Schlangen
Dialogue Systems Group, Bielefeld University
Universitaetstr 23, 33615
Bielefeld, NRW, Germany

Abstract

Side-participants (SPs) in multiparty dialogue establish and maintain their status as currently non-contributing, but integrated partners of the conversation by continuing to track, and be seen to be tracking, the conversation. To investigate strategies for realising such 'active side-participant' behaviour, we constructed an experimental setting where a humanoid robot appeared to track (overhear) a two-party conversation coming out of loudspeakers. We equipped the robot with 'eyes' (small displays) with movable pupils, to be able to separately control head-turning and gaze. Using information from the pre-processed conversations, we tested various strategies (random, reactive, predictive) for controlling gaze and head-turning. We asked human raters to judge videos of such tracking behaviour of the robot, and found that strategies making use of independent control of gaze and head direction were significantly preferred. Moreover, the 'sensible' strategies (reactive, predictive) were reliably distinguished from the baseline (random turning). We take this as indication that gaze is an important, semi-independent modality, and that our paradigm of off-line evaluation of overhearer behaviour using recorded interactions is a promising one for cost-effective study of more sophisticated tracking models, and can stand as a proxy for testing models of actual side-participants (whose presence would be known, and would influence, the conversation they are part of).

1 Introduction

Modeling the behaviour of embodied agents (ECAs or robots) in multiparty interactions is a currently on-going endeavour (see e.g. (Bohus and Horvitz 2010; Matsuyama et al. 2010; Pappu et al. 2013; Al Moubayed et al. 2014)), with challenges that go beyond those posed by dyadic interactions. One important aspect of multiparty interaction is that of conversational *roles* (Goffman 1981), which need to be both recognized and performed by a participating agent. Both tasks present challenges: assigning roles to participants dynamically in a shared space requires fusion of multimodal information, e.g. body and head posture, as well as audio source localization (Bohus and Horvitz 2010), while implementing proper behaviour for the role assummed by the agent requires models based on the mulitmodal behaviour of human participants in similar scenarios. Both functions are also constrained by the capabilities of the system: for example, a 'tilt' head gesture may express interest and moderate agreement to what a speaker is currently saying, but a robot that cannot tilt its head could not perform this behaviour and needs to find other ways to express these functions.

Compared to robots, Embodied Conversational Agents (ECAs) are currently richer in terms of behavioural repertoire: their bodies, free of the mechanical constraints of the robots, are much more flexible, as is the case with their facial expressions and gaze behaviour. The latter in particular is one of the most neglected robotic actuators, although robots with slow moving mechanical eyes exist, (e.g. (Lutkebohle et al. 2010)).The promising strategy of projecting a computer-graphics face with eyes and lips onto a robotic head aims at bringing the flexibility of ECA faces to humanoid robots (Al Moubayed, Edlund, and Beskow 2012).[1]

It is not uncommon in Human-Robot Interaction (HRI) studies to refer to the turning of the robot head as "gaze" (e.g. (Mutlu et al. 2012) and (Huang and Mutlu 2013)), despite evidence that gazing and turning the head have different pragmatic effects in human interactions (Jokinen, Nishida, and Yamamoto 2009). ECAs have much more detailed gaze behaviour, but suffer from the Mona Lisa effect (Al Moubayed, Edlund, and Beskow 2012), which, as explained in the next section, makes it impossible to realise true 'gazing-at'.

As part of our on-going effort to develop a conversational agent capable in participating in multiparty, multimodal, situated interaction, we describe our recent work in addressing this problem. We have equipped a humanoid NAO robot[2] with a pair of 'eyes' (small graphical displays), thus allowing it to 'look at' conversational partners. In order to test how much this improves the embodiment, we have designed an evaluation method in which human raters observe videos of NAO performing the role of an overhearer of a dyadic conversation between human participants. There are several versions of the same conversation with NAO exhibiting different behaviours in each one. In some his eyes are fixed, while

[1] https://www.engineeredarts.co.uk/socibot/, http://www.speech.kth.se/furhat/

[2] http://www.aldebaran.com/en/humanoid-robot/nao-robot

in others they can move. We have also varied the timing of the robot's switching its gaze towards one of the speakers, while they exchange turns. Therefore, our study has two goals: (a) test how well the robot can perform this particular role, depending on the behaviour it exhibits, and (b) test our paradigm of off-line evaluation of overhear behaviour, in order to consider using it for future exploration of more complex conversational roles and behaviours.

Some background on the role of gaze under different roles in multiparty interactions and the Mona Lisa effect is given in Section 2. We present the design of the models for the moving pupils, as well as some technical information on the implementation in Section 3. In Section 4 we describe our experimental design. We present our results with some discussion in Section 5 and our conclusions and future work in Section 6.

2 Background

The role of gaze in dyadic interaction has already been studied by Kendon (1967) and Argyle (1976), with some of their findings summarized in Vertegraal et al (2001): gaze serves several purposes in conversations, such as to convey emotions, provide visual feedback, manage turn-taking, and increase concentration to the speaker. In addition, it is more the case that listeneres look at speakers, rather than the opposite. Vertegaal et al (2001) presented evidence from literature that the opposite is true in multiparty interactions, as speakers need to display who their addressee is. Some important results from (Vertegaal et al. 2001) were that gaze of listeners in multiparty interactions is a good predictor of the current speaker, while gaze of the speaker is a good predictor for the addressee.

In a related finding reported in (Foulsham and Sanderson 2013), observers of videos of multi-party human-human interactions tended to look more at the speakers, as their gaze changed from the current speaker to the next more timely and in a more synchronized manner across observers when the videos had audio, rather than with muted audio. Edlund et al. (2012) explored the utility of this function of gaze, by using third party observers to annotate the turns of prerecorded interactions. The method proved effective when the gaze behaviours of many observers coincided, which was the case more often than not. Kawahara, Iwatate and Takanashi (2012) were able to predict the next speaker in multi-party interactions from the gaze and prosodic features of the current speaker, with some success (best accuracy 70 % of selecting the speaker but low F-score in predicting speaker changes). Using gaze features, Ishii et al (2014) were able to predict the next speaker change with a (lowest) error of 335 ms and average of 775 ms in multi-party meetings.

Jokinen, Nishida and Yamamoto (2009) differentiated between head turns and gaze and suggested that the former are more likely to be used for managing conversation flow in multi-party interactions, as they are more visible than glances of the eyes, which are faster and can be used to scan the other participants and try to predict who the next speaker might be. Bednarik, Eivazi and Hradis (2012) used gaze as a predictor of engagement level of participants in multi-party interactions and trained an SVM (74 % precision) to automate the task. Finally, Inoue et al (2014) combined gaze and acoustic features to perform speaker diarization.

A classification of roles in multi-party interactions (beyond speaker and listener) and their embodiment is presented in (Wang, Lee, and Marsella 2013). The study collected evidence from literature about the behaviour – including gaze – of participants in multi-party interactions according to their specific *roles* (following the schema of Goffman (1981)), and implemented such behaviour in virtual agents. Goffman's schema distiniguishes between *addressees, side-participants* and *unofficial participants or bystanders*. The latter are divided into *eavesdroppers* and *overhearers*, depending on whether they want to conceal that they are overhearing or not, respectively. Different behaviours need to be realised by artificial agents in order to embody these roles. For example, eavesdroppers will have their heads turned away, while overhearers will turn their heads towards the speakers, but avoid eye contact.

The importance of gaze in interaction has drawn attention in the field of human-robot interaction. In (Mutlu et al. 2012), a robot used its gaze to successfully convey the roles of addressee and overhearer to two human subjects. In (Huang and Mutlu 2013), a robot used a repetoire of social behavours, including mutual gaze and deictic gaze, in order to elicit engagement and improve performance in collaborative tasks. However, in both studies, the robots were only able to establish mutual gaze by moving their heads (their pupils were fixed). An example of a robot that participates in multi-party interactions and can direct its gaze is the one used in the SCHEMA project (Matsuyama et al. 2010). It has mechanical eyes that can glance around, altough their motions do not appear human-like. However, it demonstrates abilities of conversation flow management such as those that we envisage to realise in our project.

A crucial difference between ECAs and robots is that the latter can more readily avoid the so-called Mona Lisa effect, which owes its name to the common example of the eyes on the famous portrait appearing to stare at the observers regardless of the viewing distance and angle. As discussed in (Al Moubayed, Edlund, and Beskow 2012), this is just one manifestation of much more general phenomena, that have do with the distinction between the *physical space* that the observer is in, and the *virtual space* as shown on a perspective drawing or a computer monitor. As a result, ECAs drawn on computer screens, such as that presented in (Bohus and Horvitz 2010) cannot look towards different participants standing in front of them during multi-party interactions. According to Al Mouyabed et al (2012), it is crucial to make the eyes perceived to be part of the *physical* space, which they accomplish by drawing them on a 3D mask using the afore-mentioned back-projection technique. Our own approach to adding true gaze capabilities to a robot is described in the next section.

3 Building the Eyes

We have designed a pair of eyes for our NAO robot, using two small displays (1.77″ diagonally), with a paper mask covering both monitors and presenting a 3D nose (see Figure 1. As the mask itself is 3D and the square edges of the flat TFTs are not visible, the pupils are perceived as part of the physical space, rather than the virtual space, thus eliminating the Mona Lisa effect: viewers of the eyes in the physical space do not all have the same impression of where the eyes are looking *relative* to themselves. Rather, the eyes are perceived as looking towards *absolute* targets in the room (such as conversation partners), regardless of the observer's position. This only fails for extremely sideways viewpoints, but this is the case also for observing human gaze direction. In addition, the pupil positions dynamically adjust for the position/rotation of the robot head. Below we describe both the modeling of the eye movements as well as the technical implementation in our experiment.

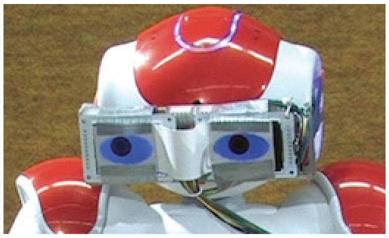

Figure 1: Eye assembly mounted on NAO robot

Modeling Pupil Movement

When human eyes look at targets around them, they *rotate* so that the pupil is turned towards the target. This is perceived by observers as a *displacement* of the pupil along two axes: the horizontal (X) axis when the target is to the side, and the vertical (Y) axis when the target is higher or lower than the eye level of the viewer. Therefore, the rotation of the eyes equates visually to a displacement of the pupils. The amount of displacement is a function of the angle of a vector originating from each eye and pointing towards the target (see Figure 2). The angle of the vector is measured relative to the plane of the face, and has thus two components: Pitch and Yaw, which are the angles around the X and Y axis, respectively. The head posture is also described with a pair of pitch and yaw angles. We do not model the pupil displacement in case of a tilted head, because the NAO head has only two degrees of freedom (yaw and pitch) and cannot be tilted.

We have built our model using data collected with a Facelab eye, head and gaze tracking device.[3] Two volunteers were tracked while fixating their eyes on targets at short to medium distances (0.5 - 3m), while simultaneously rotating their heads. At each sampled frame (60 Hz) the tracker gives us both the gaze (target) vector, the head posture, as well as the pupil displacement relative to the centre of each eye. Therefore, we can directly compute a model using linear regression, using the formula $P = A \times T + B$, where P is the vector $[X\ Y]$ of the pupil displacement from center, T is the $[Yaw\ Pitch]$ target vector, and A, B are the coefficients of the model. A is a 2×2 matrix and B is the vector of intercepts. This is a bivariate linear regression, but we observed that by setting some of the coefficents to zero (including the intercepts), we arrive at a much simpler univariate model for each axis: $X = 0.01 \times Yaw$ and $Y = 0.01 \times Pitch$ (the displacements are in meters and the angles in radians), without losing any goodness of fit ($R^2 \approx 0.85$ for all models).

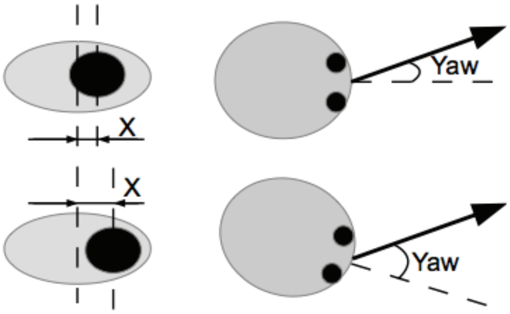

Figure 2: Relationship between head rotation, target vector and pupil displacement

Similarly to (Al Moubayed, Edlund, and Beskow 2012), we have not found significant improvements when trying to make the eyes converge to a point: the differences in the computed displacements between models that support vergence and parallel models is non-observable (1-2 mm). As a result, we currently use a design in which both eyes show identical pupil displacement. Note however, that our model is based on a linear regression of about 10000 points per speaker (after filtering the data and dropping low confidence points), and the matching of pupil displacement is based on many head rotations, rather than keeping the head in place and moving only the eyes. In the future, we plan to further improve the pupil models by introducing also changes in size, and exploring vergence of pupils more extensively.

The Eye Assembly

We have built the eyes using two Arduino 1.77" TFT LCD screens driven by an Arduino Uno (rev 3) microprocessor.[4] Figure 1 shows the robot head with the eye assembly mounted on it. The processor draws the eyes on the screen based on the coordinates it receives from a serial connection.

In order to make the robot turn its head and eyes towards the target, we control these two actuators separately, using two different components, the *head controller* and the *eye controller*. The head controller receives a signal to turn towards a specific target, which is defined as a pair of pitch and yaw angles relative to the robot's torso. Before beginning to move the head, the head controller sends this information immediately to the eye controller. The latter computes the pupil displacement based on the model described above, and thus the eyes of the robot look at the new target before the head starts to turn towards it. In addition, the eye controller is connected to the robot and reads the head posture param-

[3] http://www.eyetracking.com/Hardware/Eye-Tracker-List

[4] http://arduino.cc/en/Main/GTFT, http://arduino.cc/en/Main/ArduinoBoardUno

eters in real time. While the head controller turns the head of the robot towards the target, the eye controller continually adjusts the eyes so that they are always looking at the target during the head motion. The movements of both the eyes and head are fast, which means that the eye controller has to adjust for the *velocity* of the head, in addition to its position.

While the implementation allows the robot to track targets in the shared physical space convincingly with its head and eyes, this capability is not used to its full potential in the work presented here. In our experiment which is described below, the robot has to alternate its attention between only two targets, whenever it receives a signal to do so, as described in the next section.

4 Experiment Design

In order to test our gaze-capable NAO robot as a participant in a multi-party interaction we designed a behaviour evaluation experiment that we describe in this section. A pre-recorded dyadic human/human interaction is replayed through loudspeakers (where each human participant in the interaction is assigned their own loudspeaker). The robot then takes the role of what we call an 'active overhearer', and can show its 'engagement' in the interaction for example by turning its head and/or eyes towards one or the other loudspeaker.

In our study, the participants then watch videos of the robot performing this function (see Figure 3), and are asked to rate the quality along two dimensions, expressed as follows: [5]

1. "What do you believe, how well was the robot able to follow the discussion?"
2. "How close was the behaviour of the robot to what you would expect a human to do in the same situation?"

Thus, the robot's role requires it to behave as an overhearer: not speaking, but turning its head towards the audio source to improve hearing conditions, and following the surface structure (turn-taking) of the interaction, so that it signals to other observers that it is attentive. In pilots performed with human volunteers in NAO's place, we have observed similar behaviour, i.e. turning of the head and eyes mostly towards the currently active speaker.

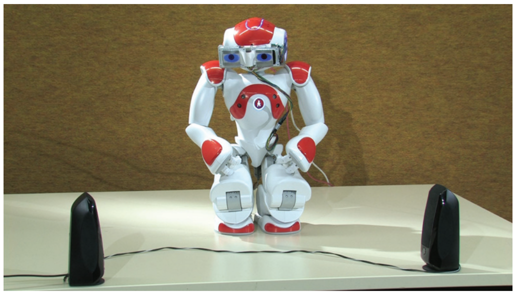

Figure 3: Snapshot from video stimulus used in experiment

Overhearer Models

We use this setting to evaluate three different models for controlling the head/gaze behaviour of the robot. Note that in this setting, since we are using pre-recorded conversations, we have available at once information about the timing of all speaker turns. Clearly, this is not information one has in a live system. We are using this setting to systematically explore which kinds of information a system *should* have, and we can systematically reduce the look-ahead that we grant to our model.

- Random Model (RND): This is our baseline model. The robot turns towards one of the two audio sources (speakers) and remains there for a random amount of time, then turns towards the other speaker, and continues in this fashion. The random amount of time is sampled from a Gaussian distribution with the same mean and standard deviation as that of the turn durations for the particular dialogue from which the given interaction is taken.
- Reactive Model (RCT): The robot turns towards the *current* speaker 300 ms *after* the turn onset. This model represents the (best-case) behaviour achievable in current dialogue systems relying on *endpointing*: The amount of time after turn onset coincides with the minimum endpointing threshold that has been achieved in a practical system, when using the best optimizations (Raux and Eskenazi 2012).
- Predictive Model (PRD): The robot turns towards the *next* speaker 300 ms *before* the turn onset. This model represents the desired performance of a system that can predict the end of turn (Schlangen 2006), and acts accordingly. Although currently there are no systems that can actually do this, the framework on which to build such systems, *incremental processing*, has long been proposed, and toolkits with which to develop such systems also exist (Baumann and Schlangen 2012). We can test this behaviour here, because we are replaying recorded interactions in which the turn-transition points are known.

For each of the model classes above, we implemented two versions. In the first version, the eyes of the robot do not move, simulating the gaze capabilities of many robots currently in use (including NAO). In the second, the eyes can move. As described in Section 3, when the robot switches targets, the gaze moves rapidly to the new target, while the head follows more slowly. The pupil positions are updated in real time while the head moves, so that when the head finally turns towards the target, the pupils are once again centered (all of this occurs very quickly). The models with moving eyes are further differentiated in that, if a turn is going to be shorter than 1000 ms, the robot will not turn its head, but rather glance at the other speaker with only the eyes and then return to looking back at the current speaker. (Again, we can do this, because we know how long the turn is going to be. We take this information to be a proxy for information about whether a turn is going to be a backchannel or an independent contribution, which is information that we assume might be possible to predict in a live system.)

Therefore, we have six models in total (3 model classes x 2 eye versions) which we label by combining the model

[5] The exact wording of the questions is shown in Appendix A

class with the letter E if the eyes are moving, e.g. RND_E means a random model with moving eyes, and PRD means a predictive model with fixed eyes.

Generation of the Stimuli

In order to construct the stimuli for our experiments we extracted audio parts of three dyadic interactions from the *Dream APartment Corpus* (Kousidis, Pfeiffer, and Schlangen 2013). In these interactions, participants discuss the layout of an apartment they design for themselves, given an extraordinary amount of available funds. The interactions in the DAP corpus are characterized by spontaneous speech, disfluencies and laughter, as well as a multitude of short turns and backchannels. Hence, it models ordinary spontaneous interaction among humans, rather than human-computer interaction, which typically progresses more slowly and is more structured with respect to turn taking (excluding the errors). Each interaction part extracted is about 60-65 seconds in duration. The parts were selected mostly randomly, only controlling for an approximately equal speaking time between the two speakers, and the presence of at least of a few instances of turn-taking.

Using the available transcriptions of the audio parts, we extracted the turn onset times and turn lengths for each turn, which we used to inform the robot when to change its gaze target based on the overhearer models described above. For each audio part and model, we recorded a short video of NAO overhearing the interaction and performing the behaviour. For this, we used two components from the *mint.tools* collection that we have developed (Kousidis, Pfeiffer, and Schlangen 2013): *ELANmod*, a modified version of ELAN (Brugman, Russel, and Nijmegen 2004) that can send play/pause commands to our event logger/replayer, *venice.hub* (Kennington, Kousidis, and Schlangen 2014). The audio is replayed by ELANmod, while venice.hub synchronously replays the turn onset events, sending at each event the following information to the head controller component (described in Section 3): (a) the target vector to the position of the speaker initating a new turn, and (b) the duration of that turn. In case of the random model, the head controller does not listen for events, but decides when to turn based on the random sample of the Gaussian distribution, while the audio is just replayed.

For each audio part, we recorded six videos (one per model). The videos were trimmed to exactly the same start and end using the audio as reference. In total, 18 videos (6 per audio part) were presented to the participants for rating.

Rating the Robot's Behaviour

As we mentioned in the beginning of this section, participants are asked (after watching each video) to provide ratings for two questions. The questions are always asked in the same order.

For the ratings, we avoided using fixed scales, such as 5-point or 7-point Likert scales. Such scales may occasionaly introduce bias when a participant uses one of the extreme points of the scale for one of the early stimuli, and as a result cannot give a yet more extreme score to a stimulus that occurs later. Instead we asked participants to use unbouded ratings: they could assign any number of their choice in order to rate each stimulus. Thus, they were free to use any higher/lower number in order to rate the next stimulus, and then a yet higher/lower number for the one after that, and so on in that fashion. We only asked them to try to be consistent in how they assign the numbers. This type of rating scheme results in each participant devising their own scale of ratings, although most used scales such as 0-5, 0-10, or 0-100, as implied by the numbers they used. Most participants did not use the full range of their implied scale, while some went over the theoretical maximum or below the theoretical minimum of these implied scales.

Each participant was presented with the full set of 18 videos in one of six different possible presentation orders, designed to counterbalance the effect of presentation order bias. First we permuted the presentation order of the six versions of one audio part, using a *Latin Square* of order 6 (Box et al. 1978), getting six presentation orders. As each participant had to rate three groups of six videos (one group for each audio part), we divided this square into two halves (rows 1,3,5 and 2,4,6). In addition, we permuted the presentation order of the groups with a Latin Square of order 3, yielding three presentation orders. Combining the two, we got six unique combined presentation orders (2 halves of first square x 3 rows of second square). The video files were automatically renamed as required, so that participants only saw a sequence of 18 videos that were named simply by their sequential number (e.g. 1.mp4, 2.mp4 etc).

Participants and Procedure

In total, 29 participants — students or otherwise affiliated with Bielefeld University, and all native German speakers — participated in the study. The vast majority were female (only six were male), aged 17-55 years (mean: 24, median: 23). Two participants had previously participated in Human-Robot interaction studies and three had other previous experience with robots. After reading the instructions and signing a consent form, the participant sat in front of a computer screen showing a GUI with a single "play" button. They wore a set of headphones, and were given a scoresheet on which to write their ratings. After any clarifications required by the participant were offered by the experimenter, the session commenced. By clicking the play button, the participant could watch one video (on a separate window, using the freely available VLC player). After watching the video, the participant recorded their ratings on the scoresheet. They then moved on to the next video, performing the same sequence of actions. Each video could be watched only once. After each participant completed the 18-video sequence, they were given a questionnaire to fill out and offered a small compensation. All 29 participants completed the full sequence successfully.

5 Results and Discussion

Here we present the analysis and results of our study. We have collected 18 ratings per participant for each of the two questions. In order to compare the ratings of the participants against each other we need to normalize them. We use two forms of normalization, namely ranking (with no assumptions about the underying distribution of ratings) and transforming the ratings into z scores (assuming a normal distribution). The z scores are computed per group of six videos rather than per participant, as some participants changed scales between groups (each participant rated three groups of six). Our first observation is that the answers of the two questions are highly correlated ($R^2 = 0.97$), as shown in Figure 4. This is not surprising, as the only way the participants can assess how well the robot understands the conversation is by its behaviour. It follows that when the rating to the *second* question is high, i.e. the robot behaves as a human would, then the robot projects an intelligent image, and thus the answer to the first question is also high. However, participants tended to give slightly lower ratings to the second question. The slope of the linear regression model fitted is 0.9, which can be interpreted as a sign that the robot's naturalness is on average rated slightly lower in comparison to its intelligence.

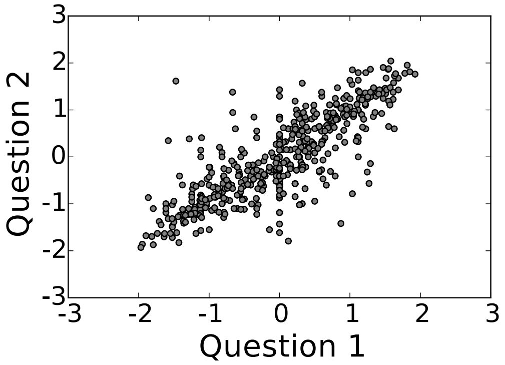

Figure 4: Scatter plot of (z-scored) ratings (question 1 vs question 2)

The distributions of the ranks of ratings for questions 1 and 2 are shown in Figure 5 and Figure 6, respectively. The significance of these results was tested using the Friedman's test for repeated measures ($x^2 = 75.88$, $p < 0.001$) followed by pairwise comparisons between all pairs using the Wilcoxon Signed Rank test as the post-hoc test. Table 1 shows all the pairwise comparisons and their results. Results marked with * denote a significant difference between the ratings for the two conditions in question, at the Bonferoni-adjusted p-value of 0.0033. In order to satisfy the assumption of the Friedman test that each row is independent, each participant's ranks of scores for the three groups of clips were first aggregated into one row (mean rank per condition for that participant), yielding 29 sets of 6 repeated measures.

As expected, the moving eyes make a big difference in both ratings, as behaviour models with eyes are rated consistently higher than the respective models with fixed eyes (the difference is significant for all these pairs). In the post-experiment questionnaire, some of the participants explicitly stated that the moving eyes make the robot "look more natural". The effect is so strong that it is common for the RND_E model (randomly timed behaviour with moving eyes) to be rated *higher* than either PRD or RCT (the predictive and reactive models with fixed eyes), especially in the "naturalness" question (Figure 6).

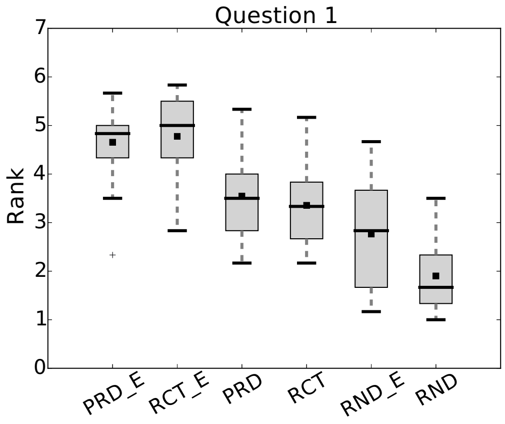

Figure 5: Box plot of ranks of ratings for question 1. Lines show the medians and squares show the means

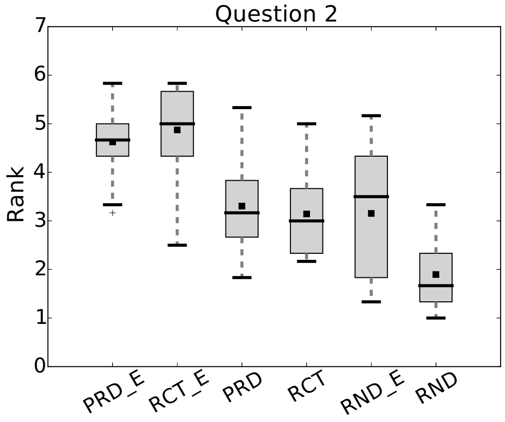

Figure 6: Box plot of ranks of ratings for question 2. Lines show the medians and squares show the means

Apart from the effect of the eyes, the similarity of ratings for RND_E with those of the "Oracle" models with fixed eyes, PRD and RCT, could be explained by the fact that the latter models are somewhat naive: they either predict or react, without any specific reason to do so. But there are many different situations in interaction that warrant prediction of (e.g., one of the interactants asks a question) or reaction to (e.g., one of the interactants unexpectedly barges in during the other's turn) turn-taking events. Therefore, neither of the current Oracle models does the right thing always.

This might also explain why we find no significant difference in the ratings *between* the PRD and RCT classes of models, either with fixed or moving eyes. A difference between the models could perhaps be found if the interaction was specifically tailored to one of them. For example a Question/Answer type of interaction would probably be in favour

Cond	Cond	Q 1		Q 2	
1	2	T	p	T	p
PRD_E	RCT_E	177	0.381	149	0.139
PRD_E	PRD	57	< 0.001 *	39	< 0.001 *
PRD_E	RCT	54	< 0.001 *	31	< 0.001 *
PRD_E	RND_E	8	< 0.001 *	26	< 0.001 *
PRD_E	RND	2	< 0.001 *	0	< 0.001 *
RCT_E	PRD	36	< 0.001 *	14	< 0.001 *
RCT_E	RCT	47	< 0.001 *	22	< 0.001 *
RCT_E	RND_E	25	< 0.001 *	27	< 0.001 *
RCT_E	RND	0	< 0.001 *	1	< 0.001 *
PRD	RCT	196	0.642	184	0.469
PRD	RND_E	130	0.058	199	0.689
PRD	RND	11	< 0.001 *	19	< 0.001 *
RCT	RND_E	131	0.061	211	0.888
RCT	RND	7	< 0.001 *	17	< 0.001 *
RND_E	RND	70	0.001 *	48	< 0.001 *

Table 1: Multiple comparison test results per condition pair for questions Q1 and Q2. The Wilcoxon Signed Rank test result T shows the positive or negative rank sum (whichever is smaller) per pairwise comparison. The differences are significant if $p < 0.0033$.

of PRD (because questions make speaker changes highly predictable), while a multi-party interaction or an interaction with many unexpected turn onsets would elicit higher ratings for RCT. However, given the high number of conditions to be tested, we decided to not vary yet another factor, but keep interaction content (largely) random.

In any case, rather than tailoring the interactions to the models, the way forward is to improve the models themselves. One would expect a model that could exhibit the appropriate behaviour (predict or react), depending on a deeper (but still shallow) monitoring of the discourse, to be rated better than the current "monolithic" approaches we have used. We defer the testing of such a model to future work, in which our current best rated models PRD_E and RCT_E would be the baseline at best.

Finally, both Oracle models clearly outperform RND, and the difference becomes even more pronounced in the presence of moving eyes. While this is an expected outcome, we interprete this as a "sanity check": it indicates that our experiment methodology is yielding sensible results. We have used a somewhat peculiar setup, in which we ask participants who were not part of an interaction to judge an overhearer (i.e., also non-participant) of that interaction, thus losing some ecological validity. On the other hand, we gain a number of advantages:

- Cost-effectiveness: we can ask many participants to watch the same interaction as observers and do not have to set up many interactions with different participants;
- Control: as these are not live interactions, we can use the same "base-material" to produce and compare many different variants of robot behaviour, as we have done here;
- Robustness: the participants cannot attempt to speak to the robot or do anything else that is unexpected, which might "break cover" and bring about an uneasy (and unwanted) situation;
- Unbiased ratings: participants are not part of the interactions, and are thus not biased too much about how complicated the content of the interaction is. Such bias could be detrimental to the robot's intelligence rating;
- More testing options: We can design Oracle models to test the performance of not yet developed behaviours, as we have done here, based on our knowledge of the recorded interaction. This setup is similar to a Wizard-of-Oz study.

In the post-experiment survey, a few participants noted that it was difficult to give fair judgements in comparison to each and every single video, due to the high number of videos. However, when asked to rate the overall difficulty of the task, on a scale of 0-5 (the higher the number the easier the task), participants responded with an average rating of 3.17 (std: 0.83). The full distribution of difficulty ratings is shown in Figure 7.

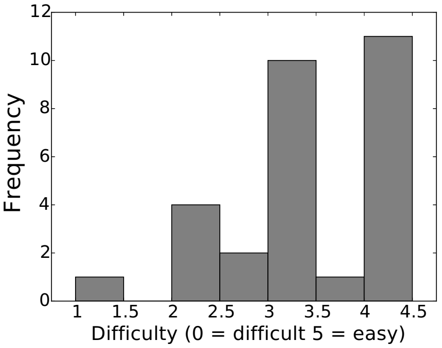

Figure 7: Participant's ratings of the task difficulty

6 Conclusions and Future Work

We have presented in this paper the current state of our ongoing work on developing a robotic agent capable of participating in multi-party interactions. We have developed and tested an eye assembly to give our robot true gazing capabilities, and we have tested the effectiveness of this using a – to our knowledge – novel evaluation paradigm, by asking participants to "overhear the overhearer". We have found that the eyes add considerably to the impression that the robot is an attentive overhearer of an interaction and can understand it, especially when combined with Oracle models that guide the timing of the robot's gaze and head turns towards the speakers. The same temporal models with fixed eyes did not perform significantly better than randomly looking at one of the two speakers but with moving eyes. The same results were found for the perceived naturalness of the robot's behaviour.

In the future, we will continue to explore this route by making our overhearer behaviour models incremental (i.e. really predicting the turn onsets rather than using Oracle models), as well as implementing behaviour models for other conversational roles in multi-party interactions.

7 Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG), grant SCHL 845/5-1, project DUEL.

We would like to thank Gerdis Anderson, Michael Bartholdt, and Oliver Eickmeyer, for assisting with various aspects of this work, from developing the technical infrastructure to conducting the experiments. We also would like to thank Robert Eickhaus for filming the robot.

References

Al Moubayed, S.; Beskow, J.; Bollepalli, B.; Gustafson, J.; Hussen-Abdelaziz, A.; Johansson, M.; Koutsombogera, M.; Lopes, J. D.; Novikova, J.; Oertel, C.; et al. 2014. Human-robot collaborative tutoring using multiparty multimodal spoken dialogue. In *HRI 2014*, 112–113. ACM.

Al Moubayed, S.; Edlund, J.; and Beskow, J. 2012. Taming mona lisa: communicating gaze faithfully in 2d and 3d facial projections. *ACM Transactions on Interactive Intelligent Systems* 1(2):25.

Argyle, M., and Cook, M. 1976. Gaze and mutual gaze.

Baumann, T., and Schlangen, D. 2012. The InproTK 2012 Release. In *NAACL 2012*.

Bednarik, R.; Eivazi, S.; and Hradis, M. 2012. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, 10. ACM.

Bohus, D., and Horvitz, E. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *ICMI-MLMI 2010*, 1. Beijing, China: ACM Press.

Box, G. E.; Hunter, W. G.; Hunter, J. S.; et al. 1978. Statistics for experimenters.

Brugman, H.; Russel, A.; and Nijmegen, X. 2004. Annotating multi-media/multi-modal resources with elan. In *LREC 2004*.

Edlund, J.; Alexandersson, S.; Beskow, J.; Gustavsson, L.; Heldner, M.; Hjalmarsson, A.; Kallionen, P.; and Marklund, E. 2012. 3rd party observer gaze as a continuous measure of dialogue flow. In *LREC 2012*. Istanbul, Turkey: LREC.

Foulsham, T., and Sanderson, L. A. 2013. Look who's talking? sound changes gaze behaviour in a dynamic social scene. *Visual Cognition* 21(7):922–944.

Goffman, E. 1981. *Forms of talk*. University of Pennsylvania Press.

Huang, C.-M., and Mutlu, B. 2013. The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *Journal of Human-Robot Interaction* 2(2):80–102.

Inoue, K.; Wakabayashi, Y.; Yoshimoto, H.; and Kawahara, T. 2014. Speaker diarization using eye-gaze information in multi-party conversations. In *Interspeech 2014*.

Ishii, R.; Otsuka, K.; Kumano, S.; and Yamato, J. 2014. Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings. In *ICASSP 2014 IEEE*, 694–698. IEEE.

Jokinen, K.; Nishida, M.; and Yamamoto, S. 2009. Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd International Universal Communication Symposium*, 303–308. ACM.

Kawahara, T.; Iwatate, T.; and Takanashi, K. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Proceedings of Interspeech 2012*, 727–730. ISCA.

Kendon, A. 1967. Some Function of Gaze-Direction in Social Interaction. *Acta Psychologica* 26:22–47.

Kennington, C.; Kousidis, S.; and Schlangen, D. 2014. Inprotks: A toolkit for incremental situated processing. *Proceedings of SIGdial 2014: Short Papers*.

Kousidis, S.; Pfeiffer, T.; and Schlangen, D. 2013. Mint. tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. *Proceedings of Interspeech 2013*.

Lutkebohle, I.; Hegel, F.; Schulz, S.; Hackel, M.; Wrede, B.; Wachsmuth, S.; and Sagerer, G. 2010. The bielefeld anthropomorphic robot head flobi. In *ICRA 2010*, 3384–3391. IEEE.

Matsuyama, Y.; Taniyama, H.; Fujie, S.; and Kobayashi, T. 2010. Framework of communication activation robot participating in multiparty conversation. In *AAAI Fall Symposium: Dialog with Robots*.

Mutlu, B.; Kanda, T.; Forlizzi, J.; Hodgins, J.; and Ishiguro, H. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.* 1(2):12:1–12:33.

Pappu, A.; Sun, M.; Sridharan, S.; and Rudnicky, A. 2013. Situated multiparty interaction between humans and agents. In *Human-Computer Interaction. Interaction Modalities and Techniques*. Springer. 107–116.

Raux, A., and Eskenazi, M. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing* 9(1):1–23.

Schlangen, D. 2006. From reaction to prediction: Experiments with computational models of turn-taking. In *Interspeech 2006, Pittsburgh, Pennsylvania*.

Vertegaal, R.; Slagter, R.; van der Veer, G.; and Nijholt, A. 2001. Eye Gaze Patterns in Conversations : There is More to Conversational Agents Than Meets the Eyes. In *SIGCHI 2001*.

Wang, Z.; Lee, J.; and Marsella, S. 2013. Multi-party, multi-role comprehensive listening behavior. *Autonomous Agents and Multi-Agent Systems* 27(2):218–234.

A Original Wording of Rating Questions

The original wording (in German) of the questions on which the participants provided the ratings is as follows:

- Was glauben Sie, wie gut der Roboter der Unterhaltung folgen konnte?
- Wie nah war das Verhalten des Roboters daran, was Sie erwarten wurden was ein Mensch machen wurde in derselben Situation?