Are You Talking to Me? Improving the Robustness of Dialogue Systems in a Multi Party HRI Scenario by Incorporating Gaze Direction and Lip Movement of Attendees

Viktor Richter[1] **Birte Carlmeyer**[1] **Florian Lier**[1]
Sebastian Meyer zu Borgsen[1] **David Schlangen**[1] **Franz Kummert**[1]
Sven Wachsmuth[1] **Britta Wrede**[1]

ABSTRACT

In this paper, we present our humanoid robot *Meka*, participating in a multi party human robot dialogue scenario. Active arbitration of the robot's attention based on multi-modal stimuli is utilised to observe persons which are outside of the robots field of view. We investigate the impact of this attention management and addressee recognition on the robot's capability to distinguish utterances directed at it from communication between humans. Based on the results of a user study, we show that mutual gaze at the end of an utterance, as a means of yielding a turn, is a substantial cue for addressee recognition. Verification of a speaker through the detection of lip movements can be used to further increase precision. Furthermore, we show that even a rather simplistic fusion of gaze and lip movement cues allows a considerable enhancement in addressee estimation, and can be altered to adapt to the requirements of a particular scenario.

ACM Classification Keywords

I.2.9 Robotics; I.5.5 Implementation: Interactive systems; I.2.11 Distributed Artificial Intelligence: Intelligent agents; I.2.7 Natural Language Processing: Speech recognition and synthesis; I.4.8 Scene Analysis: Motion, Shape, Color; I.5.2 Design Methodology: Feature evaluation and selection; I.2.10 Vision and Scene Understanding: Modeling and recovery of physical attributes

Author Keywords

dialogue systems; multi-party; multi-modal; interaction; autonomous robot; attention management; speaker; addressee

[1]Bielefeld University (CITEC), 33615 Bielefeld, Germany
[vrichter,bcarlmey,flier,semeyerz]@techfak.uni-bielefeld.de
d.schlangen@uni-bielefeld.de
[franz,swachsmu,bwrede]@techfak.uni-bielefeld.de

INTRODUCTION

In the context of Human Robot Interaction (HRI), it has become increasingly apparent that social and interactive skills are indispensable in order to build an intuitive, natural communication via speech, gestures, and facial expressions [4][10]. Moreover, *socially correct* interaction is desired. Therefore, Dautenhahn [8] already proposed a *"robotiquette"*, a set of "social rules for robot behaviour that is comfortable and acceptable to humans" in 2007. In a series of HRI studies [33], it was classified as socially interactive, in contrast to socially ignorant, that the robot took an interest in the humans activity and that it was actively *looking* at the human. Dautenhan further argues, that "a robot that serves as a companion in the home [...] needs to possess a wide range of social skills which will make it acceptable for humans. Without these skills, such robots might not be 'used' and thus fail in their role as an assistant." This finding is also confirmed in [9][12][15] – to only mention a few. To this end, the robotics community puts considerable effort into the development of "attentive systems" capable of interactively directing the robot's attention towards the human and vice versa [22][5][8][13].

While it is already a complex task to correctly direct the robot's attention in a 1:1 interaction between a human and a robot, this complexity significantly increases in a 1:N scenario where a robot needs to participate in a mixed interaction with, and between, multiple persons at the same time. Hence, single user HRI allows for multiple design simplifications for a situated robot. To give an example, it is sufficient to direct the robot's attention, via gaze for instance, towards its *sole* interaction partner or a potential focus of discourse. Moreover, the conversational roles in a single user interaction are limited to *speaker* and *addressee* [32]. Thus, usually all dialogue acts produced by the human interaction partner are targeted at the robot and therefore do not need to be verified or tested.

In contrast, in case of a multi user interaction, these simplifications may become a hindrance to the interaction dynamic. Additional conversational roles, like *ratified* (intended to listen) by the speaker or *side participant* (not part of the present dialogue act) emerge [32]. These additional roles will have a negative impact on the human-robot interaction dynamic if not considered and designed correctly in the robot's behaviour capabilities.

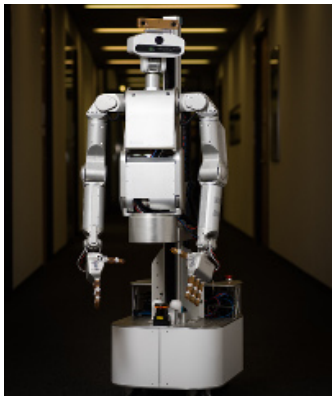

Figure 1. The Meka M1 Mobile Manipulator robotic-platform. ***Image ©Johannes Wienke***

The assumption that all dialogue acts are directed towards the robot does not hold in a multi user scenario. Users may also direct their gaze and/or speech towards another human instead of the robot. Thus, in the worst case scenario, the robot will react to every speech recognition result – even if it is not addressed. This may lead to refusal or even exclusion of the robot from the interaction. Subsequently, due to the fact that users recognize and negotiate conversational roles among each other, a robot will negatively influence the interaction whether or not this is intended. Moreover, Mutlu et al. [24] already showed that the gaze behaviour of a robot has a significant impact on conversational roles and participation of people in a multi person interaction. This effect was also confirmed in [31]. Therefore, if a robot is not capable of distinguishing multiple conversation partners and also exhibits this distinction, via directed attention for example, the human-robot interaction will become less natural, unintuitive and in the worst case – unacceptable. Finally, in case of a mobile robot, it is not safe to assume that all interaction partners are always located in front of the robot or any other "visible" location. It is also not safe to expect that a human interaction partner is always willing to move to the area where the robot is able to recognize them. Therefore, a robot must possess capabilities that allow for spacial localization and recognition of potential interaction partners in a dynamic environment, not only using vision but also other modalities. Moreover it needs to recognize if this potential partner is expressing an intention to communicate.

In this contribution we present our humanoid *Meka* (Figure 1), participating in a multi party human robot dialogue scenario. We investigate the impact of attention management and addressee recognition on the robot's performance to distinguish utterances directed at it from communication between humans. Taking into consideration the aforementioned issues and requirements, we formulate the following hypotheses:

1. Mutual gaze, as a means of yielding a turn to someone, can be used to facilitate the decision to whom an utterance was directed.

2. Recognition of lip movements can be used to validate an interaction partner as producer of perceived speech, and thus further increases the accuracy of addressee recognition.

3. Mutual gaze and lip movement recognition complement each other, and therefore can be combined in a simple, logical manner to further enhance the addressee recognition performance.

To test our hypotheses we conducted a proof of concept evaluation where the robot participated in a multi party human-robot interaction. During the interaction the robot was occasionally addressed, e.g., to provide information or execute a simple command. In this scenario the robot is required to direct its attention towards spatially distributed interaction partners. At the same time it needs to be able to react to robot-directed speech (addressee recognition) while ignoring interpersonal dialogue. To this end, we evaluate approaches to addressee recognition utilizing different types of visual cues, i.e. mutual gaze at the end of utterances, detection of lip movements, and logical combinations of these.

RELATED WORK

With respect to control of a robot's attention Ruesch [28] et al. presented a bottom up approach using the iCub robot based on audio-visual saliency. However, in their work face recognition as a social cue was not considered. Moreover, the evaluation was not carried out in the context of an human robot interaction (HRI) scenario. Breazeal [5] et al. introduced an attention system using the Kismet robot which implements bottom-up saliency and top-down habituation capabilities only using visual features. By changing weights between features, they generated different behaviours of their robot with respect to gaze preferences in the scene. Lang [21] et al. introduced a person tracking and anchoring system using the mobile robot BIRON. The presented system allowed the robot to identify and follow speakers based on the fusion of face detection results, sound source localization and laser scan data. However, the system is not able to instantly switch its communication partner due to a fixed interaction decay period and was only evaluated for speaker localization but not addressee recognition.

A further important question for our work is the effect of a robot's behaviour on the participants of an ongoing interaction. Bruce [6] et al. showed that actively turning to human interaction partners significantly increases their willingness to interact. Moreover, Mutlu [24] showed that shifting a robot's attention via looking behaviour during an interaction, can impose a conversational role on participants.

With regard to addressee recognition, Li [23] et al. fused features from upper body posture, face, gaze and lip-movement detection and emotion recognitions to calculate which person would most likely want to interact with their system. The calculated results were used to interact with the presumably most attentive person in the systems field of view. However, they do not take into consideration towards which participant an utterance was directed. In [3], Bohus et al. use a virtual avatar on a screen. They evaluate their system's turn taking performance in multi party interaction and observe that errors in addressee recognition have a negative impact on the quality of their turn taking model. [2] further elaborate on this model. They use sound source localization for speaker detection and classify the speaker'e visual focus of attention (vfoa) (based

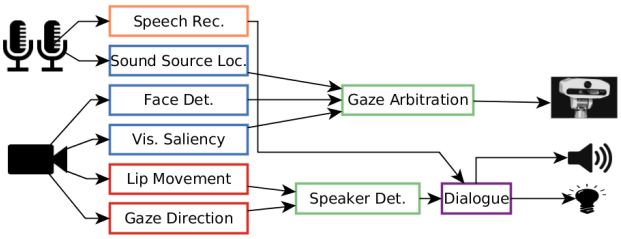

Figure 2. Overview of the robot's components for speaker/addressee recognition, gaze arbitration and dialogue management. ©*Sebastian Meyer zu Borgsen*

on head poses) as addressee of the speech. The work by [31] eliminates the problem of speaker recognition by utilizing close talking microphones. Their robotic head *Furhat* considers itself addressee, when a speaker looked at it at some time during speech production. They further evaluate different turn taking cues produced by the robot in multi party interaction. Using the resulting data [17] creates a data driven classification after which utterance the robot should take the turn, utilizing voice activity, syntax, prosody, head pose of both persons, movement of cards, and dialogue context. Another data driven approach is realized in [16] where proportions of time a speaker/partner is looking at the robot/partner during utterances and contextual features are used to classify whether a *Nao* was addressed.

The presented work above assumes that all participants are visible throughout the whole interaction. Additionally, [3] and [31] (and therefore [17] too) used an external static camera not affected by the agent's actions. [23] and [16] implicitly require all participants of the interaction to reside within the robots field of view. Thus, while there exists a large body of research on robot attention modelling and addressee recognition, we go beyond these approaches by (1) taking into account more realistic settings (i.e. that it is not always possible to see all present interaction partners at once) and (2) by providing an evaluation of turn taking cues within a multi party setting with respect to correct attention management and addressee recognition.

SYSTEM DESCRIPTION

The humanoid robot platform *Meka* is part of the Cognitive Service Robotics Apartment (CSRA)[1] research project. It is used to explore research questions related to human-robot-interaction in smart-home environments. It's core software system (Figure 2) running on the robot consists of a speaker/addressee recognition, a gaze arbitration component and a dialogue system. These components include further sub-modules to process camera and stereo microphone input streams. The system controls the robot and is capable of producing speech output and triggering external actuators in the apartment, such as switching the light on/off.

[1] https://www.cit-ec.de/de/content/cognitive-service-robotics-apartment-ambient-host

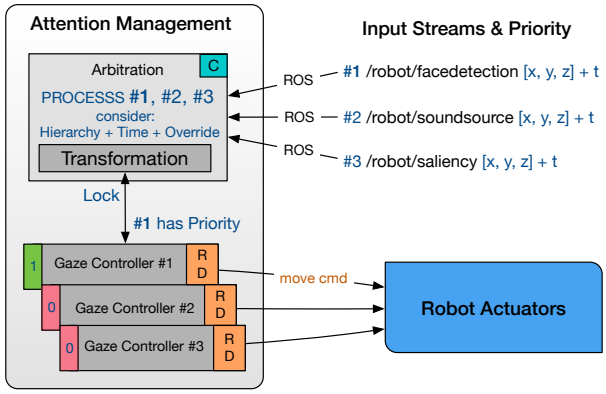

Figure 3. Overview of the attention management system. ©*Florian Lier*

Robot Platform

Amongst many other sensors, the M1 Mobile Manipulator robotic-platform *Meka* (Figure 1) features a *PrimeSense* Carmine RGBD camera to receive RGB images. Furthermore, a laser range finder allows to gather spacial information about the environment. Two microphones allow to retrieve audio in stereo. A real-time-enabled computer controls the robot hardware. With the compliant force controlled actuators including four-fingered hands the robot can grasp objects and execute human-interpretable gestures. An omni-directional base and lift-controlled torso enables navigating in complex environments. In total, the robot is equipped with 37 motor-powered joints. It has 7 per arm, 5 per hand, 2 in the head, 2 in the torso and 9 joints actuate the base including the z-lift. The motors in the arms and hands are Series Elastic Actuators (SEAs)[26], which enable fine force sensing.

Attention Management

The robot's attention management system is based on hierarchical prioritization of multi-modal sensor input streams. We realized this subsystem, which is openly available on github[2], as follows (Figure 3).

In general, the attention management system is sensor- and robot-independent. This is achieved by a) abstracting multi-modal sensor input via middleware data streams and b) an easily exchangeable robot control interface. Supported middleware implementations are Robot Operating System (ROS)[27] and Robotics Service Bus (RSB)[34].

Input data streams are a series of typed sensor messages consisting of a global target position [x,y,z] and a time stamp (t). Usually, these streams contain the position of a face detection result, a sound source or an interesting location in the robot's environment. The arbitration component is set up using a global configuration file (C). In this file topic names of *N* desired input streams, e.g., */robot/facedetection*, their associated data type, priority, timeout, control strategy, control mode, and override values are defined. Based on this configuration, the arbitration component continuously reads sensor messages – starting with the stream that has the highest configured priority. If the time stamp of the current message is *not* older than its

[2] https://github.com/CentralLabFacilities/simple_robot_gaze

pre-defined maximum timeout value (100 milliseconds for example), the current position is transformed from the global target position into the robot's field of view. If the time stamp is "too old" or if there is no new message at all, the next priority (stream) is evaluated. Besides the arbitration component, the attention management system implements a so-called gaze controller per configured input stream. A gaze controller holds a reference to the robot driver (RD) and an activation flag.

Essentially, the robot driver implements the interface to the robot's hardware, its hardware abstraction layer or control API. The robot driver can be easily exchanged for a desired target platform, e.g., for NAO or iCub. However, after the current input message has been verified and the position has been transformed, the corresponding gaze controller is activated by toggling its activation flag and a movement command is sent to the robot. The control strategy of a gaze controller can be either configured to open- or closed loop. In case of an open control strategy, the command is issued and no further processing is required. In case of a closed-loop strategy, the command is issued and the arbitration component is locked until the desired position is reached. Moreover, the control mode can be set to relative or absolute positioning. These modes are required for sensors that move along with the robot (relative), a camera in the robot's head for instance, or fixed sensor setups (absolute). The attention management component features an override mode. If a preconfigured threshold is exceeded, the default hierarchy is temporarily disabled and the highest priority is instantly shifted to the input stream which triggered the override. In our setup we configured the attention management system as follows.

The highest priority were face detection results, the second highest priority were sound source localization results and the third highest priority were results produced by the visual saliency component. We activated the override feature for all three input streams which enabled the robot to initially look at a person and instantly shift it's attention towards another location where a loud sound or a visually salient spot was detected. This made it possible to dynamically and spontaneously shift the attention of the robot towards potential communication partners, even if they where not in the robot's field of view.

Addressee Recognition

To assess the gaze direction of currently observable persons the gaze detector from Schillingmann et al. [29] is used. The implementation was extended to be able to receive video data via ROS and publish its results, containing relative gaze directions and face landmarks [19] for all observed persons via RSB for further processing.

Based on this data, the addressee recognition component classifies its current speaking state for each person observed and whether the person maintains mutual gaze with the robot or not. Mutual gaze is assumed when a person keeps its gazing direction within a range of α around the robots head. To classify a person's speaking state we inspect its facial landmarks over a predetermined time period. When the variance of the distance between the horizontally centred points of the inner lips exceeds a threshold, the person is classified as currently speaking:

$$Speaking(p_n) = \begin{cases} yes & \text{when } \mathrm{Var}(X_n) > d \\ no & \text{else} \end{cases}$$

where p_n is the n-th observed person and X_n is the set of its inner lip distances during the last Δ_t. The constants $\Delta_t = 600ms$, $\alpha = 12°$ and $d = 1.5$ were estimated in advance to produce reasonable results. Finally, the robot is classified as addressee if the person is speaking while maintaining mutual gaze with it.

Figure 4. The experimental set up as used in a pre-study.

Dialogue Management

For verbal communication we use a combination of the incremental natural language processing toolkit InproTK[1] and the human-robot dialogue manager Pamini[25] that have been integrated in [7]. MaryTTS[30] is used for speech synthesis and Sphinx[20] for speech recognition. A simple dialogue act generation module produces human dialogue acts based on keyword-spotting on the incremental ASR results, e.g., *action requests* such as "turn on/off the light", *information requests* such as "What time is it?" or *confirmations/negations*. The dialogue manager receives these results and processes them in sequence based on the current state of the interaction and the results of the addressee recognition.

STUDY SETUP & METHOD

In this section, the study setup and applied evaluation methods will be described in detail by elaborating on the experimental setup and the data recording and annotation.

Experimental set up

Figure 4 depicts the experimental set up. Three participants are sitting around the table in the CSRA, one participant is sitting on the sofa and two in the armchairs. On the table are 15 small slips of paper available with the following set of possible tasks or questions for the robot: (i) *turn on/off the light*, (ii) ask for the *current time*, (iii) whether a *call* or (iv) *delivery* has been missed, (v) request about *possible experiments* or (vi) which *data is getting recorded*, and (vii) ask for more *information about the Zen-garden* in the apartment.

The multi party interaction consists of two parts: interpersonal communication and human robot interaction. In the first phase a participant picks one of the tasks from the table, chooses another member of the group and tell him/her to issue the current task/question to the robot. In the second part of the interaction the chosen participant has to gain the attention of the robot and repeat the request addresing the robot. The participants

were told to repeat their utterance a maximum of three times if the robot did not react. During the experiment, results of the speech recognition were evaluated and – if possible – executed, only if the robot was recognized as addressee of the utterance. This was only the case if mutual gaze at the end of the utterance and lip movements were detected, and allows us to evaluate other, more permissive, strategies later on (using the recorded interactions). The grammar chosen for speech recognition was relatively small because it was not subject of the evaluation.

Data Recording and Annotation

All interactions were recorded via two network-enabled Basler cameras and one Rode NT55 omni-directional microphone mounted at the ceiling of the apartment to cover the whole interaction area. Additionally, the robot's internal PrimeSense camera video stream has been recorded. Moreover, we collected various system events such as speech recognition results, generated dialogue acts and detailed information of the addressee recognition component. These consist of facial landmarks, gaze recognition results, and classification results for mouth movement, mutual gaze , and addressee.

For annotation purposes, the two top-down videos, the audio track and system events were merged into one ELAN[35] file (for further information about this process cf. [14]).

The study has been carried out with German native speakers. In total, we recorded approximately 53 minutes of interaction in 5 trials with 2 female and 13 male participants in total. A typical trial takes approximately 10 minutes. Altogether the dialogue system detected 874 human dialogue acts, 152 of these would have triggered a verbal response or a corresponding system action (light on/off). In order to evaluate the means of different approaches to addressee recognition, a ground truth annotation was carried out for each dialogue act.

RESULTS

To assess the performance of different approaches to addressee recognition, task specific utterances are extracted and classified into "robot is addressee" (positive condition) and "robot is not addressee" (negative condition). Comparing the classification results of the different approaches yields the corresponding 2x2 confusion matrix, which can be used to calculate accuracy, recall and precision (cf. Figure 5).

As baseline approach we accepted all results from the speech recognition (C_0). This approach does not reject tasks, thus its recall is 100% and the accuracy equals the prevalence of the dataset for the robot being addressed (84%, for an interpretation of this rather high amount of tasks addressed towards the robot see section: DISCUSSION).

We compare the results of the baseline test with the following approaches:

- (C_l) *lip movement* Accept tasks only when movement of the lips was detected.
- (C_g) *mutual gaze* Accept tasks only when mutual gaze with the robot was detected.

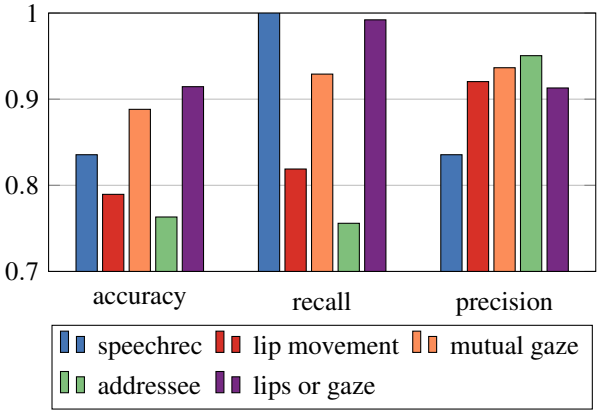

Figure 5. Accuracy, recall and precision of different addressee recognition approaches.

- (C_a) *addressee* Accept tasks only when the robot was recognized as addressee. This requires both mutual gaze and lip movement, and is the condition that was actually used throughout the trials.
- (C_x) *lips or gaze* Accept tasks when either lip movement or mutual gaze (inclusive disjunction) were detected.

As depicted in Figure 5 it is evident that *mutual gaze* can preserve a high recall (93%) compared to the baseline. The detection of lip movements does not perform as well (82%) as the baseline, and the conjunction of lip movement and mutual gaze detection C_a achieves only 76% recall. The accuracy of C_l and C_a is lower than the baseline's accuracy too. In contrast, the accuracy of C_g (89%) and C_x (91%) is higher than the baseline's accuracy. All non-baseline conditions show a precision of $> 90\%$, with a maximum of 95% for C_a in contrast to the baseline precision of 84%.

Many interactions between the participants in our scenario were not recognized as tasks by the robot. The resulting recognitions of short statements were out of context and therefore could be automatically rejected by the dialogue system (see section: Data Recording and Annotation). This results in the relatively unbalanced prevalence of the dataset, with 84% of the tasks actually addressed at the robot. The ratio between statements directed at the robot and statements exchanged between the participants is rather specific to our scenario. We therefore calculated the diagnostic odds ratio (DOR) for all conditions. This measure is an indicator of test quality, like accuracy, but decoupled from the prevalence of the test set. A DOR of x can be interpreted as: *The odds of being correctly classified as addressee are x times higher than the odds of being falsely classified as addressee.* [11].

Considering the DOR, all conditions perform better than the baseline C_0 with C_x indicating the best performance (cf. Figure 6). In contrast to the accuracy results, the conjunction of mutual gaze and lip movement detection has a higher DOR than lip movement detection only. This shows that the addressee recognition is, in general, more reliable than C_l.

We looked into the ten cases in which the robot was addressed but failed to look at the respective speaker. While in 3 of

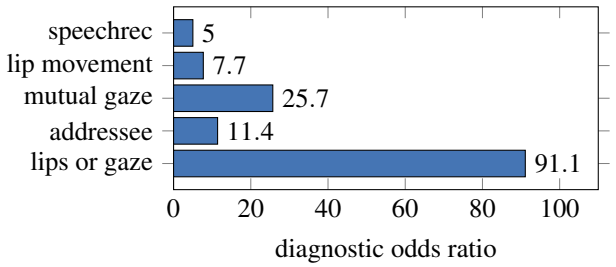

Figure 6. Diagnostic odds ratio of different addressee recognition approaches.

these cases none of the approaches' (C_l, C_g, C_a, C_x) cues were identified, mutual gaze with the chosen attendee was involved in 6 of the other cases. Mouth movements were misclassified twice, once because an attendee had to laugh.

DISCUSSION

There are multiple reasons for the relatively high amount of tasks directed towards the robot (84%) in contrast to tasks recognized from interpersonal speech (16%). First, participants used a much more variable wording and lower voice in interpersonal speech, decreasing the probability of the robot recognising a task. Additionally, in one trial, the participants resorted to showing the task descriptions to each other instead of stating the tasks. Finally, the addressee recognition – needing the highest certainty for acceptance – often rejected tasks, forcing the participants to repeat their assignment.

Nonetheless, the results show that detection of mutual gaze at the end of an utterance improves the accuracy and precision when recognizing whether or not a robot was verbally addressed. This confirms our first hypothesis.

Detection of lip movements does not perform that well in our scenario. When used as a single feature for addressee recognition in our scenario it has a much lower accuracy than the baseline, which means that our second hypothesis could not be confirmed. Additionally, it results in low accuracy and recall in the *addressee* condition where we require both mutual gaze and lip movements.

However, this does not eliminate lip movements as a feature per se. When mutual gaze and lip movement detection is used in conjunction, the system shows a high precision of 95%, which makes it more suitable for scenarios where the robot is rarely addressed. The result produced by *lips or gaze* reach the highest accuracy and DOR of all tested approaches, showing that a simple logical combination of mutual gaze and lip movement information increases the performance of addressee recognition, and thus confirms our third hypothesis.

Furthermore, we observed that in multiple cases the robot was addressed but did not detect mutual gaze or lip movements. One explanation for this is that either of these features could not be recognized at the relevant moment due to movements of the robots head and the resulting motion blur. Another explanation is that the robot sometimes did not look at the speaker when the task was stated and executed. While in such cases the focused person did not speak, there are multiple reasons for other persons to establish mutual gaze, which allows the robot to recognize itself as addressee although it is not looking at the speaker. For instance: in multi party interaction people do not only look at the speaker but also at the addressee [18]. This is especially valid while a turn is transferred from one agent to another, where the attention is typically directed towards the agent most probable to take the turn [32]. However, the exact reasons for this apparent supplementing of the two features are subject to further investigation.

Considering that it was impossible to see all participants of the interaction at once in our scenario ,it is evident that the generic attention management, combining bottom-up and top-down saliency features, provides a considerably good basis for addressee recognition using only visual features. This results in the observation that already a simple addressee recognition can increase the performance of an agent in a multi party interaction. Nevertheless another consequence is that the attention module has a direct impact on the quality of addressee recognition.

LESSONS LEARNED & FUTURE WORK

The first observation is that the attention module sometimes triggers unintended behaviour of the robot. In only a few cases the participants had difficulties to acquire the robot's attention while it looked at another person. This is based on the fact that face detection results have the highest priority. Although it is possible to override this attention cue, e.g., verbally, this becomes difficult in the case of multiple persons speaking simultaneously (habituation).

A second observation is that people do not always wait until the robot looks at them. Occasionally they are already talking while the robot is still turning around. Often, in this case it is not possible to align visual features, such as gaze, and speech recognition results. In addition, an addressee recognition based solely on visual features is very challenging during head movements due to motion blur. Based on these observations other modalities or arbitration mechanisms should be considered.

We believe that the proportion of recognized dialogue acts may be different in long term multi party interaction. In our scenario most of the recognized dialogue acts were actually addressed towards the robot. We expect that in long term multi party interactions the dialogue acts not addressed to the robot will increase. In such cases addressee recognition becomes even more important. Therefore we additionally should consider scenarios with more interpersonal communication for evaluation.

However, with the recorded dataset we are able to tackle some of these issues. We will now be able to train and test different classifiers for lip movement detection to improve the accuracy of the classification of this visual feature. Furthermore, we will investigate different approaches for data fusion. On the one hand, a more sophisticated model for late feature fusion could be used. On the other hand, it is possible to explore various techniques for early fusion based on the raw data.

In addition, our observations show that the integration of other modalities is required. For instance, information from the

attention management could be used in addressee recognition and vice versa. Apart from these low level features, we want to investigate the inclusion of high level features. One example are the speech recognition results. The verbal addressing of the robot by either using its name or the word "robot" should be exploited in order to improve the results of the addressee recognition.

We are also interested in the evaluation of the influence of such an attentive system on the subjective ratings of the robot by the participants. Therefore, we will carry out further experiments to measure different key concepts in HRI such as anthropomorphism and likeability of the robot.

CONCLUSIONS

In this work we investigated the impact of attention management and addressee recognition on a robot's capability to distinguish utterances directed at it from communication between humans. A multi party interaction study was carried out and the recordings annotated with ground truth information. Based on the evaluated results, we can show that attention management facilitates addressee recognition, especially in situations where it is not possible for the robot to see all participants of the interaction at the same time. It can further be verified that mutual gaze at the end of an utterance, is a meaningful signal for turn yielding. Verification of a speaker through the observation of lip movements decreases false positive addressee recognitions. Furthermore, already simple logical combinations of gaze and lip movement classifications yield good performance when it comes to finding out who is being addressed. However, more work is required to create a fusion model that performs well in all situations or can be tuned for a specified precision or accuracy in a continuous way. Extra effort is needed to enhance the interoperation between attention management and addressee recognition in order to be able to cope with some of the observed corner cases.

ACKNOWLEDGMENT

This work was supported by the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and by the German Federal Ministry of Education and Research (BMBF) via the KogniHome project (project number: 16SV7054K).

REFERENCES

1. Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*. 29–32. http://nbn-resolving.de/urn:nbn:de:0070-pub-25145558
2. Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Ieeernational Conference on Multimodal Interfaces, Workshop on Machine Learning for Multimodal Interaction*. 1. DOI: http://dx.doi.org/10.1145/1891903.1891910
3. Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Special Interest Group on Discourse and Dialogue*. http://dl.acm.org/citation.cfm?id=2132903
4. Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and Autonomous Systems* 42, 3 (2003), 167–175.
5. Cynthia Breazeal and Brian Scassellati. 1999. A Context-dependent Attention System for a Social Robot. In *International Joint Conference on Artificial Intelligence*. 1146–1151. http://dl.acm.org/citation.cfm?id=1624312.1624382
6. Allison Bruce, Illah Nourbakhsh, and Reid Simmons. 2002. The role of expressiveness and attention in human-robot interaction. In *International Conference on Robotics and Automation*, Vol. 4. 4138–4142. DOI: http://dx.doi.org/10.1109/ROBOT.2002.1014396
7. Birte Carlmeyer, David Schlangen, and Britta Wrede. 2014. Towards Closed Feedback Loops in HRI: Integrating InproTK and PaMini. In *Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction (MMRWHRI '14)*. 1–6. DOI: http://dx.doi.org/10.1145/2666499.2666500
8. Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362, 1480 (2007), 679–704. DOI: http://dx.doi.org/10.1098/rstb.2006.2004
9. Boris De Ruyter, Privender Saini, Panos Markopoulos, and Albert Van Breemen. 2005. Assessing the Effects of Building Social Intelligence in a Robotic Interface for the Home. *Interacting with Computers* 17, 5 (2005), 522–541. DOI:http://dx.doi.org/10.1016/j.intcom.2005.03.003
10. Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems* 42, 3 (2003), 143–166. DOI:http://dx.doi.org/10.1016/S0921-8890(02)00372-X
11. Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, and Patrick M.M. Bossuyt. 2003. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 56, 11 (2003), 1129–1135. DOI: http://dx.doi.org/10.1016/S0895-4356(03)00177-X
12. Marcel Heerink, Ben Kröse, Vanessa Evers, BJ Wielinga, and others. 2008. The influence of social presence on acceptance of a companion robot by older people. *Journal of Physical Agents* 2, 2 (2008), 33–40. DOI: http://dx.doi.org/10.14198/JoPha.2008.2.2.05
13. Patrick Holthaus. 2014. *Approaching Human-Like Spatial Awareness in Social Robotics - An Investigation of Spatial Interaction Strategies with a Receptionist Robot*. Ph.D. Dissertation. Bielefeld University.
14. Patrick Holthaus, Christian Leichsenring, Jasmin Bernotat, Viktor Richter, Marian Pohling, Birte Carlmeyer, Norman Köster, Sebastian Meyer zu Borgsen, René Zorn, Birte Schiffhauer, Kai Frederic Engelmann, Florian Lier, Simon Schulz, Philipp Cimiano, Friederike Eyssel, Thomas Hermann, Franz Kummert, David

Schlangen, Sven Wachsmuth, Petra Wagner, Britta Wrede, and Sebastian Wrede. 2016. How to Address Smart Homes with a Social Robot? A Multi-modal Corpus of User Interactions with an Intelligent Environment. In *International Conference on Language Resources and Evaluation* (23-28).

15. Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. 2011. How Can I Help? *International Journal of Social Robotics* 3, 4 (11 2011), 383–393. DOI: http://dx.doi.org/10.1007/s12369-011-0108-9
16. Dinesh Babu Jayagopi and Jean-Marc Odobez. 2013. Given that, should i respond? Contextual addressee estimation in multi-party human-robot interactions. In *Human-Robot Interaction*. 147–148. DOI: http://dx.doi.org/10.1109/HRI.2013.6483544
17. Martin Johansson and Gabriel Skantze. 2015. Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. In *Special Interest Group on Discourse and Dialogue*. 305–314.
18. Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2014. Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction. In *Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. 21–26. DOI:http://dx.doi.org/10.1145/2666242.2666249
19. Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874. DOI: http://dx.doi.org/10.1109/CVPR.2014.241
20. Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The CMU SPHINX-4 speech recognition system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1. Citeseer, 2–5.
21. Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot a Fink, and Gerhard Sagerer. 2003a. Providing the basis for human-robot-interaction. In *International Conference on Multimodal Interfaces*. 28. DOI: http://dx.doi.org/10.1145/958432.958441
22. Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. 2003b. Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot. In *International Conference on Multimodal Interfaces*. DOI: http://dx.doi.org/10.1145/958432.958441
23. Liyuan Li, Qianli Xu, and Yeow Kee Tan. 2012. Attention-based addressee selection for service and social robots to interact with multiple persons. In *Proceedings of the Workshop at SIGGRAPH WASA*, Vol. 1. 131. DOI: http://dx.doi.org/10.1145/2425296.2425319
24. Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations. In *Human Robot Interaction*, Vol. 2. 61. DOI: http://dx.doi.org/10.1145/1514095.1514109
25. Julia Peltason and Britta Wrede. 2010. Pamini: A Framework for Assembling Mixed-Initiative Human-Robot Interaction from Generic Interaction Patterns. In *Special Interest Group on Discourse and Dialogue (SIGDIAL '10)*. 229–232.
26. Gill A Pratt and Matthew M Williamson. 1995. Series elastic actuators. In *Human Robot Interaction and Cooperative Robots*, Vol. 1. IEEE, 399–406.
27. Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *International Conference on Robotics and Automation Workshop on Open Source Software*, Vol. 3. 5.
28. Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hornstein, Jose Santos-Victor, and Rolf Pfeifer. 2008. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *International Conference on Robotics and Automation*. 962–967. DOI: http://dx.doi.org/10.1109/ROBOT.2008.4543329
29. Lars Schillingmann and Yukie Nagai. 2015. Yet another gaze detector: An embodied calibration free system for the iCub robot. In *International Conference on Humanoid Robots*. 8–13. DOI: http://dx.doi.org/10.1109/HUMANOIDS.2015.7363515
30. Marc Schröder and Jürgen Trouvain. 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology* 6, 4 (2003), 365–377. DOI: http://dx.doi.org/10.1023/A:1025708916924
31. Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In *International Conference on Multimodal Interaction*. 67–74. DOI: http://dx.doi.org/10.1145/2818346.2820749
32. David Traum. 2004. Issues in Multiparty Dialogues. In *Workshop on Agent Communication Languages*. 201–211. DOI:http://dx.doi.org/10.1007/978-3-540-24608-4_12
33. Michael L Walters, Kerstin Dautenhahn, Sarah N Woods, Kheng Lee Koay, R Te Boekhorst, and David Lee. 2006. Exploratory studies on social spaces between humans and a mechanical-looking robot. *Connection Science* 18, 4 (2006), 429–439. DOI: http://dx.doi.org/10.1080/09540090600879513
34. Johannes Wienke and Sebastian Wrede. 2011. A Middleware for Collaborative Research in Experimental Robotics. In *IEEE/SICE International Symposium on System Integration (SII)*. 1183–1190. DOI: http://dx.doi.org/10.1109/SII.2011.6147617
35. Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Language Resources and Evaluation Conference*, Vol. 2006. 5th.