Towards Addressee Recognition in Smart Robotic Environments

An Evidence Based Approach

Viktor Richter
Bielefeld University (CITEC)
Inspiration 1, 33619 Bielefeld, Germany
vrichter@techfak.uni-bielefeld.de

Franz Kummert
Bielefeld University (CITEC)
Inspiration 1, 33619 Bielefeld, Germany
franz@techfak.uni-bielefeld.de

ABSTRACT

This work explores what people do to inform their surroundings about who is the main addressee of their communicative acts in smart environments. A corpus of naive users, solving daily tasks in a smart home, which is additionally inhabited by a robot, is investigated. Evidence drawn from the corpus is used to create a first model for addressee recognition in smart environments. Finally, the performance of the model is evaluated using the corpus, and possible future improvements and challenges are discussed. The main contribution of this work is a detailed analysis of human addressing cues in smart environments, and the resulting, evidence based addressing model.

CCS Concepts

•Human-centered computing → **Ambient intelligence; User models; User centered design;** *Natural language interfaces;* **•Software and its engineering** → *Requirements analysis;*

Keywords

addressing; smart environments; multi-modal; natural interaction; social robot

1. INTRODUCTION

With technology ever becoming smaller and cheaper during the last decades, social robots and smart environments more and more start to enter our daily life. Many of these smart environments, especially in the context of smart homes and ambient assisted living, are specifically designed to support the daily routines of inhabitants, and allow easy, intuitive interaction. All possible devices, that can be found in a typical home, are being equipped with sensors and actuators, and then combined into a huge network of smart objects. Although automation is the mainly used way to assist inhabitants in smart homes, people repeatedly demand to be able to override the systems behaviour [10]. Rich user interfaces on computers and smart phones were created to allow inhabitants to control their homes, but the interfaces tend to become complicated. Additionally, it happens to feel tedious, if a frequently used function requires the user to get and unlock the smart phone, start the control application and then find the desired function to finally be able to trigger it.

This is one of the reasons why modern smart homes additionally contain social robots and virtual interactive agents, which - among other things - can interact with inhabitants and bridge the gap between the smart devices and the user. Moreover, other ways of controlling continuously get further refined. Smart appliances can be directly controlled via touch, speech, by using various gestures or rarely even through brain machine interfaces[6]. However, controlling lots of devices without switches or a Graphical User Interface (GUI) requires lots of modalities and metaphors, leading to ambiguity and the additional problem of addressee distinction. Which light should be switched if someone claps, and if a user makes a waving gesture does he or she want the tv to turn on or the alarm to stop? The same problem arises in verbal interaction. When should the robot react to speech, when a virtual avatar or even the apartment as a whole?

There are many possible answers to these questions which usually are situation, appliance, and user specific, but a general answer is hard to find. To get closer to a general model of addressee recognition it needs to be investigated how people naturally communicate in their daily life and how they show their surroundings who is meant by their communicative acts.

The rest of this paper is organised as follows. Section 2 examines related research for common approaches towards addressee recognition. A corpus of unconstrained, task-directed interactions of naive users in a robotic smart home is introduced in Section 3. In the following Section the corpus is further examined for evidence on human addressing cues. The resulting findings are used to create and evaluate a first addressee recognition model in Section 5. Finally, future enhancements and possible problems of the proposed model are discussed in Section 6, and a summary is given in Section 7.

2. RELATED WORK

To get an overview on the current state of the art in addressee recognition, we look into current work in human-robot interaction (HRI), interaction with *classical* smart environments and finally interaction in smart environments inhabited by humans and robots.

In HRI it is widely accepted that social robots should be able to interact via speech. Spexard et al. emphasise that "The user should be able to communicate with the system by, e.g., natural speech, ideally without reading an instruction manual in advance." [26] and [9] found that 71% of the participants of their study wanted a robot companion to communicate in a *human-like* manner. This results in various social robots utilising speech as the main modality in their interaction with human partners [4][19][7]. However, because

addressee recognition is often not the focus of the corresponding research, it often gets replaced by simple heuristics like reacting to everything that can be understood.

Gross et al. developed a robot as a companion for people with mild cognitive impairments which can be controlled using "speech or the external tablet" [11], but does not go into detail on how the robot decides whether it is addressed by human speech or not. Many works implicitly assumes that the robot is in a dyadic interaction with only one interaction partner. The receptionist robot in [13] and the interactive learning robot in [18] both are tailored towards dyadic interaction and do not consider other addressees in the scene. Even *in the wild*, interactions are often simplified to one-on-one situations, where the robot chooses an interaction partner and sticks to it until the end of the interaction, meanwhile attributing all recognised speech to the current interaction partner [17]. In multi-party interactions a robot needs to distinguish between multiple humans and handle situations where they talk to each other. One common approach in such cases is to track the current speaker and assume the corresponding addressee to be the speakers current visual focus of attention (VFOA). To detect the speaker [25] use close-talk microphones for every participant, and [2] a microphone array and sound-source localisation. Both works use video processing techniques to find the speakers head orientation and infer the corresponding VFOA. A different approach is implemented in a multi-party interaction scenario by Richter et al. [23] where not all participants can be seen by the robot at once. In this system, the robot *Meka* considers itself addressee, when it shares mutual gaze with a speaking person. Whether a person is speaking is detected visually and the participants have various, multi-modal ways of attracting the robots attention.

Research in smart homes and smart environments often focuses on unobtrusive adaptation and automation, trying to reduce the need for explicit control of the available smart appliances [12]. However, Dragone et al. argue that, when users feel to be passive recipients of smart services, it decreases their perceived value, trust and understanding of the system [10]. Because there always will be the need to override a otherwise optimal system behaviour, active control metaphors in such environments additionally need to be intuitive and easy to remember. The addressee of a touch gesture is rarely ambiguous, e.g. touching a lamp can switch its state. This is an already common metaphor but it require the person to move around to reach the corresponding interfaces. Rich GUIs [3] can, in combination with a smart-phone, allow users to control an apartment without having to move around, but the interfaces tend to become cumbersome and complicated when many functions and options are available. [15] use a distinct gesture for every functionality in their smart home. This way, the addressee of the gesture is encoded in the gesture itself, but 8 different gestures are needed only to switch the light and open or close the curtains. This certainly is hard to scale to modern smart homes with lots of various interactive appliances, and hard to remember for a user, especially when the functionality is not frequently used. [16] use only a small set of gestures but specify the addressee of the gesture via a smart-phone screen, which may be somewhat redundant in case of a simple switch. A different, multi-modal approach is implemented in [8], where a pointing device is used to specify the addressee before the person can use gestures or speech to control the specified device. For multi-party, verbal interaction Potamitis et al. [22] require their participants to explicitly state the addressee of their speech by starting commands with the addressed agent's name.

In mixed scenarios, where human participants can interact with the smart environment and one or more social robots, it is even harder to find information about addressee recognition. In [20] a robotic smart house is presented, where the different agents can be addressed by the resident via explicit verbal naming or by pointing at it (similar to [22] and [8]). The same group later proposes to use a social robot or avatar as a mediator between the human and the smart environments functionalities in [21], but does not explain how they plan to distinguish whether the robot was addressed or not. Unfortunately this is the case in various works on robot inhabited smart environments. [28] uses a ubiquitous home robot as a human-machine interface in the smart environment which can be ordered or talked to by residents. Baeg et al. present their smart environment for service robots which provide humans with services [24]. Neither of them considers the problem of addressee distinction in such a context. Bernotat et al. conducted a *wizard of oz* study [1] to investigate the addressing behaviour of naive users in a smart robotic environment while solving simple daily tasks. They empirically show which modalities and interfaces people prefer to use for different daily tasks. Addressee estimation - not being the focus of the study - is done by human *wizards* and annotators and is not further explored.

Although addressee recognition does not seem to get much attention in the presented work, most of the presented systems have some kind of way to explicitly exclude, simplify, or just ignore ambiguities in the addressing behaviour of human interaction partners. The systems which explicitly distinct between different addressees, utilise VFOA or mutual gaze in multi-party HRI and explicit pointing or naming in smart environments. While these heuristics seem plausible at first, they lack the empirical basis on how naive users convey the addressee of their commands, when they are free to do it without restrictions.

3. INTERACTION CORPUS

To be able to recognise human addressing behaviour, we first need to analyse what people do to specify the addressee of a deliberate, communicative act in a smart environment. The corpus by Holthaus et al. [14] encompasses multi-modal interaction and addressing behaviour of naive users in a robot-inhabited, smart home. It was created during a study in the Cognitive Service Robotics Apartment (CSRA) and provides a good basis for our analysis. Bernotat et al. [1] conducted the corresponding interaction study with naive users to investigate how and whom they address in such a setting to solve everyday tasks. The corresponding corpus builds the basis for our investigations on how people inform about the addressee of their communicative acts in smart environments. We therefore first give a short summary of the corpus.

The aim of the *wizard of oz* study (explained in detail in [1]) was to find out how and what people address in a smart robotic apartment to solve different mundane tasks. More precisely, the questions were (I) who is addressed (robot, light, apartment) (II) which modality is used (speech, gesture, touch) to solve a task. Seven everyday tasks, consisting of 1. turn on light in hallway 2. turn off light in hallway (from kitchen) 3. listen to music 4. check mail 5. check phone-calls 6. check time 7. set brightness of a lamp (without using speech) were chosen to be solved in the apartment. The participants had nearly no restrictions in their solution of the task. Only the ordinary light switches on the walls could not be used. When the *wizard* decided that the participant made an attempt to solve the task, the corresponding functionality was triggered in the apartment. The *wizard* additionally chose if the robot or the apartment were more probably addressed and should therefore react. Additionally, the trials were split into a *verbal* and a *nonverbal* condition. In tasks, where information needed to be presented to the user (i.e. tasks 4, 5, 6), the first group got a verbal response from the apartment or robot. For the second group only visual cues

were used, consisting of attention guiding gestures by the robot and apartment and information printed on various screens.

The corresponding corpus (further explained in [14]) contains video and audio recordings of the study trials, which were carried out in the CSRA. Additionally, the corpus contains recordings of all system events during the trials and *ELAN*[5] annotations. Among other things, the system event recordings contain the wizards decisions, the corresponding actions of the smart environment (including screens, lamps, robot, speech) and sensory information about the apartment state (power consumption, doors, windows, cupboards, etc.). Some of this data provides the basis for the *ELAN* annotations. These contain further, hand-annotated information about the course of the study and emotional expressions of the participants (e.g. smiling). Information about the participants goal-directed actions like the used modality, language features, focus of attention and final addressee is also available.

According to the outcome of the study, in such an unconstrained setting people often prefer verbal interaction, but also use gestures and touch interaction. When a task has a corresponding physical interface, this will be directly addressed in most cases. In case of information requests or tasks where there is no obvious physical addressable entity, people tend to more often address the robot than in the other cases. However, in such cases the addressee often becomes *unspecific*.

4. REQUIREMENTS ANALYSIS

The used corpus can give us information about who is addressed in each task (and via which modality), but does not expose the way in which people indicate the addressee of their actions. To acquire such information the available data needs to be further investigated. We concentrate on the annotations at the moment of task solution. This is when the *wizard* decided that the participant actively tried to solve a given task and which agent is addressed thereby. It therefore can be seen as a measure for human-level performance in command and addressee recognition. We further look into the following information: 1. *agent* (apartment, robot or floor-lamp): shows which appliance was addressed according to the *wizard* in each task 2. *modality* (speech, gesture, touch): the modality used by the participant to solve the task 3. *addressee* (one of various appliances, the robot, the apartment, the participant or undecided): who was addressed by the participant according to the annotator 4. *VFOA* (same options like addressee): the participant's focus of attention just before solving the task. All mentioned significance tests are performed using the Chi-Square test with a significance level of 0.05.

4.1 Visual Focus of Attention

It is apparent from the related work (Sec. 2) that the VFOA is a commonly used feature in addressee recognition in HRI. This is often used with the assumption that the VFOA and addressee usually match [27]. To first revise this assumption for goal directed actions in a smart environment, we examine how often the VFOA and addressee are equal during the study. The resulting plot can be seen in Figure 1.

Figure 1 shows, that the addressed entity is the same as the focused entity over all tasks in 89% of the interactions. The distribution of addressed entities are highly dependent on the given task. In tasks where a physical entity can be addressed (the lights in tasks 1, 2 and 7) this is the most common addressee in the task solution. On the other hand, when a task does not have a distinct addressable interface, like in the information requests (tasks 4, 5, 6) and when listening to music (task 3), the robot or an unspecific entity is addressed. Moreover, the difference in the congruence

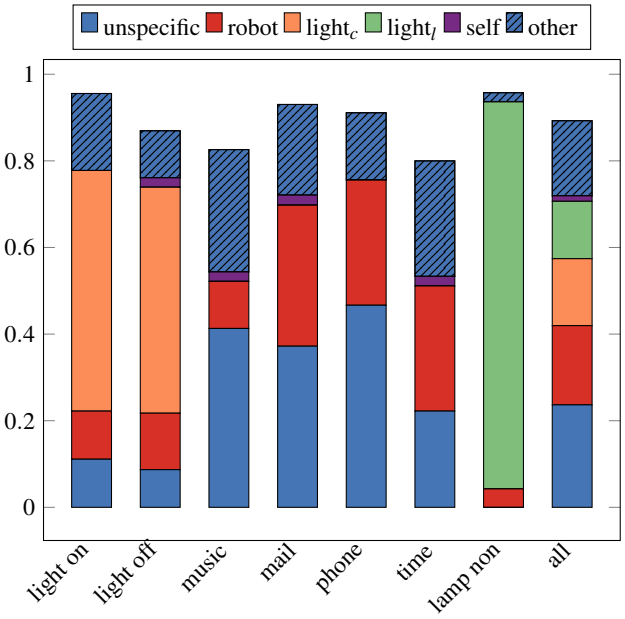

Figure 1: Plots how often (relative) the focus of attention of participants matched their addressee during the task solution. Each bar is subdivided into different VFOA. The mean over all tasks is shown in the rightmost bar (*all*). The addressee *other* is a combination of rarely addressed entities (various screens, switches, etc.). ©*Viktor Richter*

of addressee and VFOA between tasks with a distinct responsible entity (the light tasks) and tasks without such an entity (listen to music and information requests) is not significant. The set of rarely addressed entities, which were combined into the addressee *other*, contains switches, screens and the apartment as a whole. Another rare addressee during the trials was the participant (in 1% of the interactions, *self* in Figure 1). This is a somewhat special case in which participants talk to themselves, or are looking at their task description while at the same time trying to solve it. An addressee which is common, especially in tasks without a distinct responsible entity is *unspecific* (24% over all tasks). According to Bernotat et al. this was when "[...] it was obvious that participants addressed an interface within the apartment, but it was unclear which one."[1]. When the addressee was *unspecific* according to the annotators, the *wizards* decided that the apartment (and not the robot) should react in 94% of the times.

The inspection of the correlations between addressee and VFOA shows that the latter is a highly informative cue for addressee recognition, but does not suffice in all situations. Additionally, it is still a challenging task to automatically classify the VFOA of an inhabitant in a smart home in a unobtrusive and reliable manner. As far as can be said from the used corpus, the correlation between addressee and VFOA is independent from the type of the task - whether information request or control of an appliance does not significantly change the matching probability. Finally, the cases where addressee and VFOA are at the participant or *unspecific*, and there is no other information about an addressee usually result in the apartment being the right entity which should react.

4.2 Other Addressing Cues

The addressee cannot always be determined solely on the basis of VFOA. In fact, 11% (34 of 317) of the interactions had a mismatch between these two classifications. Additionally, in au-

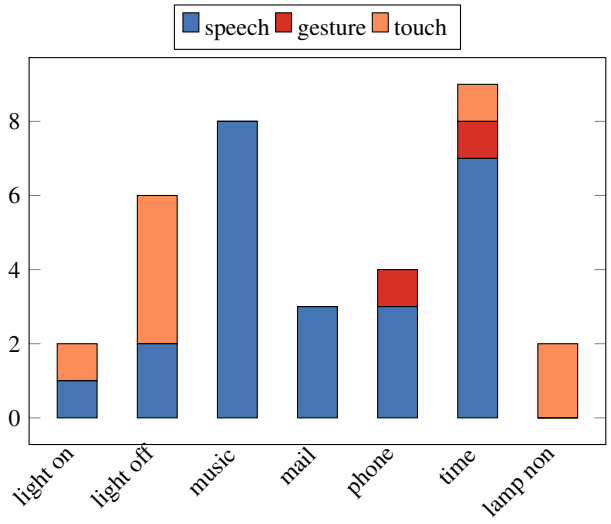

Figure 2: Plots how often (absolute) the VFOA was not equal to the addressee during the task solution. Each bar is subdivided into the used modalities. ©*Viktor Richter*

tonomous systems the information about the VFOA of a user will likely be less reliable than in the hand-annotated corpus. To be able to model these kind of interactions we first need to further analyse these cases. The plot in Figure 2 shows mismatching interactions split by task and modality. A significant change in the mismatch proportions between the modalities cannot be found. The mismatch between addressee and VFOA was significantly higher in the nonverbal condition ($p = 0.03$) over all modalities. Within the modalities, a significant difference could only be found in speech ($p = 0.02$). In the following, the cases where addressee and VFOA do not match are examined in detail.

The 8 mismatches in the *touch* modality consist of: Three cases where participants touched screens or switches (addressee) while looking if the state of the lamp changes (VFOA). Four times where users switched off the light in the corridor by closing the door between the corridor and the kitchen. One time where (according to the annotator) the user did not solve the task but the *wizard* decided differently (wizard error). The 2 mismatches in the *gesture* modality consist of: Once waving in front of a screen while looking into the room and once waving at the robot while looking at screens. Both cases happened in the nonverbal condition, where information requests always were answered by text messages on screens. In all of the 24 mismatches in the *speech* modality either the addressee or the VFOA is *unspecific*, *self* or *apartment*. Correspondingly, the *wizards* attributed all but two of the tasks to the apartment. In the other two cases, the participant was looking at the task description while interacting only with the robot during the whole trial.

Direct verbal specification of the addressee was not used in case of the apartment. The robot was addressed 4 times using the term *robot* and 7 times using the pronoun *you*. In all of these cases the robot was additionally the participants VFOA (no mismatches). The light in the corridor was verbally named 32 times while being VFOA and 6 times with *unspecific* attention. However, it is hard to say if the participants addressed the light as an entity or wanted another entity to switch the light.

The inspection of interactions, where addressee and VFOA did not match, allows a more sophisticated interpretation of the participants attention. Touch and gesture interactions show that the users attention may differ from the addressee to monitor the progress of the task in question. Verbal interactions show that it is usually possible to assume the apartment to be addressee, when no other addressee is probable enough. Additionally, verbal interaction can establish a context which is more meaningful than the visual attention of the participants.

4.3 Discussion

Some insights on the addressing behaviour of people in smart, robot inhabited environments could be drawn from the analysis of the multi-modal interaction corpus. First of all, the common assumption, that the VFOA of a person equals the addressee does not hold in all cases, but is a highly informative cue. It is valid in interactions with appliances and robots (Section 4.1). When unspecific attention is considered as attention towards the whole environment as an entity even the apartment is addressable using VFOA in most cases. This assumption is supported by the following observations: (1) The *wizards* decisions usually comply with the interpretation that *unspecific* attention addresses the apartment. (2) In contrast to the apartment, which is hardly detected, the annotators frequently find that the addressee is *unspecific*. (3) The participants contradict the annotations by stating that they most frequently addressed the apartment [1]. However, automatic detection of VFOA is more prone to classification errors than the human annotation, which was used in this corpus. Additionally, freely moving people can yield even more challenges to an automatic VFOA classification system which results in situations where this cue is not available.

Because mismatches between VFOA and addressee happen in all modalities, and this information may not be available in all situations more cues are needed to get a reliable model. The analysis of other addressing cues (Section 4.2) allows the following observations: (1) The VFOA may differ from the addressee when people use touch or gestures, and the outcome of the task should be directly visible. In these cases the user may observe the controlled appliance while using a different appliance for control (e.g. switching lamps via screens). (2) The interaction history may be more essential than the current attention of a person. This especially can happen in verbal interactions. (3) The content of a communicative act can explicitly specify the addressee, thus overriding all other cues.

5. INITIAL MODEL & EVALUATION

A first addressee recognition model can be created directly from the insights which were drawn from the investigation of the corpus. A simple Bayesian Network that takes the findings from section 4 into account can be seen in Figure 3. The model uses information about the users attention, whether an addressee was specified explicitly, the used modality, and the conversational context of the interaction to calculate the most probable addressee.

The newly created model can be evaluated using the data from the corpus. To this end, the rarely used addressees from the corpus are combined into one addressee set of outliers. The annotations on attention and modality are used as is. If some appliance was addressed directly (*addressed by content*) during the task solution, this was additionally annotated on top of the corpus. For an approximation on the conversational context the addressee of the previously solved task is used. During the first task the conversational context is always *unspecific*, because the interaction just started.

Two configurations of the Bayesian Network are trained. The first (*full*) uses the complete model. For the second network (*context-free*) the *conversational context*-node is removed from the model. This way the *context-free* model can be realised without using long-term information. The trained models are evaluated in two conditions. In the first condition all input information is observed and the addressee needs to be inferred. In the second condition no information

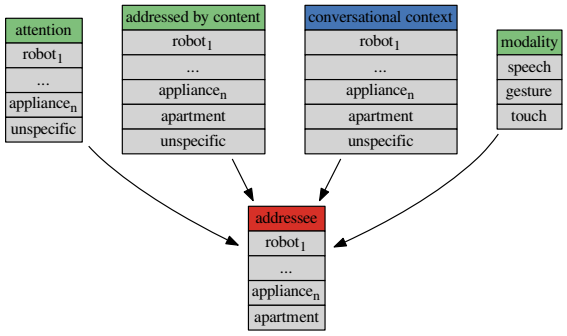

Figure 3: A simple bayesian model of addressee recognition using attention, content addressee, conversational context, and modality. The boxes are nodes of a directed Bayesian Network with the nodes name on the left and its possible states on the right side. The arrows depict the dependencies between the nodes. The green and blue nodes are observed to infer the probability distribution of addressees (red node). The blue node (*conversational context*) was not used in all evaluations. ©*Viktor Richter*

	baseline	*full*	*context-free*
obs. VFOA	0.8991 ± 0.03	0.8991 ± 0.03	$\mathbf{0.9148 \pm 0.03}$
unobs. VFOA	0.2808 ± 0.05	$\mathbf{0.6348 \pm 0.05}$	0.4984 ± 0.05

Table 1: The mean classification performance and confidence intervals of the *baseline* model and the trained networks *full* and *context-free*. The results are shown for inference with fully observed input data (obs. VFOA), and inference without information about the VFOA (unobs. VFOA). Confidence intervals are calculated using the Adjusted Wald Method. ©*Viktor Richter*

about the user's attention is available. The baseline classification performance, when the VFOA is always considered addressee, is rather high (89.91%). This is a little bit higher than the proportion mentioned in Section 4.1 (89.27%), because of the reduced set of addressable entities. When no information about the user's VFOA is available, we choose the overall most probable addressee for the baseline. This results in 28.08% correct classifications and is better than the random choice (20%).

A leave-one-out cross-validation is performed to calculate the performance of the models and evaluate them on the corpus data.

The corresponding results can be seen in Table 1. In the condition with observed VFOA no significant improvements of the classification can be achieved. With the human annotated VFOA, the complete network (*full*) can not beat the baseline performance. The classification using the network that ignores the long-term context (*context-free*) achieves the best performance (91.48%) in this condition, but is still within the confidence interval. In the condition with information about the user's VFOA missing, both Bayesian Models greatly surpass the baseline performance. In such a situation, the fully trained network performs best (63.48%).

The evaluation of the proposed model shows that, although the human annotations on VFOA are hard to beat as a cue for the addressee of an action, the classification quality can be further enhanced using a statistical model. Moreover, in situations with missing information about the VFOA the trained Bayesian Model can use other cues to infer the addressee and achieve a much higher performance than the baseline.

6. OUTLOOK & CHALLENGES

The proposed model for addressee recognition was created on the basis of evidence from an interaction corpus of unconstrained human interaction with a robotic smart home. It uses observations of interactions of naive users to choose an appropriate addressee for a communicative human action. Therefore, it is a good step towards intuitive interaction between humans and embodied and non-embodied agents in smart environments. Knowledge about the situation and conversational context could be used to further enhance the model in future iterations. Additionally, the model which currently generalises over different users, could be adapted over time to better reflect the habits of different users.

For future iterations of the model, an evaluation in an online situation should be conducted to evaluate how much the model generalises to other interactions in smart environments. Some challenges need be considered before such an online model can be evaluated. The data that was used in this evaluation comes from human annotations. Automatic detection of the context is not always possible. VFOA can be detected using head orientations or gaze directions, but this information is likely to be much more noisy than the human-annotated data from the corpus. Often, depending on the activity of the user, the VFOA may even be not observable. As seen in the evaluation (Section 5), it is likely that such noise will greatly decrease the performance of the *baseline* model, while the Bayesian Models, to some extent, should be able to deal with it. To find out if speech contains an explicit addressee, a component for natural language understanding is needed. Using the last addressee as interaction context is a heuristic too. This could be further enhanced by explicitly modelling the temporal dynamics of interactions using a Hidden Markov Model (HMM) or similar approaches. Additionally, further information about the user could be used to improve the model.

The model is created from single user interactions. To extend the model to multi-user interactions in smart environments, additional persons could be integrated as possible addressees. This way each user could have an own addressee recognition which accounts for the other users. Furthermore, it is possible that people act differently in a multi-user scenario. Therefore, it currently can not be said how good (or bad) this model would perform in such a case. This is open for further evaluations.

7. CONCLUSIONS

The goal of this paper was to create an evidence based addressee recognition model for smart environments inhabited by humans and robots. First, interactive robots and smart environments were inspected for their approaches at addressee recognition. Then, a recent multi-modal corpus of unconstrained human interactions with a robot-inhabited smart home was analysed. The evidence, found in the interaction corpus was used to create and evaluate a first model for addressee recognition. It could be shown that the proposed model can yield better results than a simple approach based only on VFOA, especially in case of missing or noisy input data. Finally, possible future enhancements and challenges of a deployment in a real smart environment were discussed.

8. ACKNOWLEDGMENTS

Thanks to Holthaus et al. [14] for creating and making available the corpus used in this paper.

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).

9. REFERENCES

[1] J. Bernotat, B. Schiffhauer, F. Eyssel, P. Holthaus, C. Leichsenring, V. Richter, M. Pohling, B. Carlmeyer, N. Köster, S. Meyer, R. Zorn, K. F. Engelmann, F. Lier, S. Schulz, R. Bröhl, E. Seibel, P. Hellwig, P. Cimiano, F. Kummert, D. Schlangen, P. Wagner, T. Hermann, S. Wachsmuth, B. Wrede, and S. Wrede. Welcome to the future - How naïve users intuitively address an intelligent robotics apartment. In *International conference on Social Robotics*, 2016.

[2] D. Bohus and E. Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. *Proceedings of the SIGDIAL Conference*, 2011.

[3] L. Borodulkin, H. Ruser, and H.-R. Trankler. 3D virtual "smart home" user interface. In *International Symposium on Virtual and Intelligent Measurement Systems*. IEEE, 2002.

[4] A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *International Conference on Robotics and Automation*, volume 4. IEEE, 2002.

[5] H. Brugman and A. Russel. Annotating multi-media / multi-modal resources with ELAN. *International Conference on Language Resources and Language Evaluation*, 2009.

[6] R. Carabalona, F. Grossi, A. Tessadri, P. Castiglioni, A. Caracciolo, and I. de Munari. Light on! Real world evaluation of a P300-based brain-computer interface (BCI) for environment control in a smart home. *Ergonomics*, 55(5), may 2012.

[7] B. Carlmeyer, D. Schlangen, and B. Wrede. Towards Closed Feedback Loops in HRI. In *Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*, New York, New York, USA, 2014. ACM Press.

[8] S. Carrino, A. P'eclat, E. Mugellini, O. Abou Khaled, and R. Ingold. Humans and smart environments: A Novel Multimodal Interaction Approach. In *International conference on multimodal interfaces*, New York, New York, USA, 2011. ACM Press.

[9] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry. What is a robot companion - Friend, assistant or butler? In *International Conference on Intelligent Robots and Systems*, 2005.

[10] M. Dragone, J. Saunders, and K. Dautenhahn. On the Integration of Adaptive and Interactive Robotic Smart Spaces. *Paladyn, Journal of Behavioral Robotics*, 6(1), jan 2015.

[11] H.-M. Gross, C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley, T. Langner, M. Merten, C. Huijnen, H. van den Heuvel, and A. van Berlo. Further progress towards a home robot companion for people with mild cognitive impairment. In *International Conference on Systems, Man, and Cybernetics*. IEEE, oct 2012.

[12] K. Guo, Y. Li, Y. Lu, X. Sun, S. Wang, and R. Cao. An Activity Recognition-Assistance Algorithm Based on Hybrid Semantic Model in Smart Home. *International Journal of Distributed Sensor Networks*, 12(8), aug 2016.

[13] P. Holthaus. *Approaching Human-Like Spatial Awareness in Social Robotics - An Investigation of Spatial Interaction Strategies with a Receptionist Robot*. PhD thesis, Bielefeld University, 2014.

[14] P. Holthaus, C. Leichsenring, J. Bernotat, V. Richter, M. Pohling, B. Carlmeyer, N. Köster, S. Meyer zu Borgsen, R. Zorn, B. Schiffhauer, K. F. Engelmann, F. Lier, S. Schulz, P. Cimiano, F. Eyssel, F. Kummert, T. Herrmann, D. Schlangen, U. Rückert, S. Wachsmuth, B. Wrede, and S. Wrede. 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*. ELRA, 2016.

[15] D. Kim and D. Kim. An Intelligent Smart Home Control Using Body Gestures. In *International Conference on Hybrid Information Technology*, volume 2. IEEE, nov 2006.

[16] C. Kühnel, T. Westermann, F. Hemmert, S. Kratz, A. Müller, and S. Möller. I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies*, 69(11), oct 2011.

[17] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. a. Fink, and G. Sagerer. Providing the basis for human-robot-interaction. In *International conference on Multimodal interfaces*, New York, New York, USA, 2003. ACM Press.

[18] I. Lutkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter, and R. Haschke. The curious robot - Structuring interactive robot learning. In *International Conference on Robotics and Automation*. IEEE, may 2009.

[19] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations. In *International conference on Human robot interaction*, volume 2, New York, New York, USA, 2009. ACM Press.

[20] K.-H. Park, Z. Bien, J.-J. Lee, B. K. Kim, J.-T. Lim, J.-O. Kim, H. Lee, D. H. Stefanov, D.-J. Kim, J.-W. Jung, J.-H. Do, K.-H. Seo, C. H. Kim, W.-G. Song, and W.-J. Lee. Robotic smart house to assist people with movement disabilities. *Autonomous Robots*, 22(2), jan 2007.

[21] K.-h. Park, H.-e. Lee, Y. Kim, and Z. Z. Bien. A Steward Robot for Human-Friendly Human-Machine Interaction in a Smart House Environment. *IEEE Transactions on Automation Science and Engineering*, 5(1), jan 2008.

[22] I. Potamitis, K. Georgila, N. Fakotakis, and G. Kokkinakis. An integrated system for smart-home control of appliances based on remote speech interaction. *European Conference on Speech Communication and Technology*, 2003.

[23] V. Richter, B. Carlmeyer, F. Lier, S. Meyer zu Borgsen, F. Kummert, S. Wachsmuth, and B. Wrede. 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. In *International Conference on Human-agent Interaction*, Singapore, 2016.

[24] Seung-Ho Baeg, Jae-Han Park, Jaehan Koh, Kyung-Wook Park, and Moon-Hong Baeg. Building a smart home environment for service robots based on RFID and sensor networks. In *International Conference on Control, Automation and Systems*. IEEE, 2007.

[25] G. Skantze, M. Johansson, and J. Beskow. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In *International Conference on Multimodal Interaction*, 2014.

[26] T. Spexard, M. Hanheide, and G. Sagerer. Human-Oriented Interaction With an Anthropomorphic Robot. *IEEE Transactions on Robotics*, 23(5), oct 2007.

[27] R. Stiefelhagen, Jie Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4), jul 2002.

[28] T. Yamazaki. Beyond the Smart Home. In *International Conference on Hybrid Information Technology*, volume 2. IEEE, nov 2006.