A Speech and Gesture Spatial Corpus in Assisted Living

Dimitra Anastasiou
SFB/TR8 Spatial Cognition
Computer Science/Languages Science, University of Bremen, Germany
Enrique-Schmidt Str. 5, 28359 University of Bremen
anastasiou@uni-bremen.de

Abstract

Ambient Assisted Living (AAL) is the name for a European technology and innovation funding programme. AAL research field is about intelligent assistant systems for a healthier and safer life in the preferred living environments through the use of Information and Communication Technologies (ICT). We focus specifically on speech and gesture interaction which can enhance the quality of lifestyle of people living in assistive environments, be they seniors or people with physical or cognitive disabilities. In this paper we describe our user study conducted in a lab at the University of Bremen in order to collect empirical speech and gesture data and later create and analyse a multimodal corpus. The user study is about a human user sitting in a wheelchair and performing certain inherently spatial tasks.

Keywords: Ambient Assisted Living, human-robot interaction, intelligent wheelchair

1. Introduction

Speech and gesture interaction in AAL is very important, among others, for home automation, that means to control smart devices and objects. Physical and/or cognitive disabilities can prevent people from fulfilling everyday tasks, such as reaching a high cupboard, opening a door, or changing TV channels. User-friendly adaptable technology in combination with speech and gesture recognition could help in these and many other cases to simplify daily lives and reduce the dependency on other persons. In the case of wheelchair users, speech and gesture interaction is an intuitive way to navigate in their environment. Apart from the obvious advantages of speech, gesture input recognition in AAL is necessary, since gesture is one of the communication modalities of people with muteness or speech disorders. Apart from the advantages of gesture in AAL and home automation, gestural interaction, in general, makes human-human, but also human-computer and human-robot interaction more natural and effective (see discussion in Dickinson et al., 2007).

As far as the relationship between speech and gesture is concerned, McNeill (1992) pointed out that speech and gesture must cooperate to express a person's meaning and Goldin-Meadow (2003) stated that speech-associated gestures often convey information that complements the information conveyed in the talk they accompany and, in this sense, are meaningful. In addition, in relation to spatial concepts, pointing to a destination with a deictic gesture is most often less time-intensive and more efficient than spoken language description (Anastasiou, 2011b).

We follow the gesture typologies of McNeill (1992) to design (and analyse the results of) a study in which both gesture and speech are considered as modes of communication. We focus particularly on gesticulation, the use of "unconventionalised" hand-and-arm movements that are almost always accompanied by speech. Gesticulation is subcategorised to iconic, metaphoric, rhythmic, cohesive, and deictic gestures. The description of those gestures is outside the scope of this paper.

In this paper we describe the Bremen Ambient Assisted Living Lab (BAALL) and an intelligent wheelchair-robot called *Rolland*. The goal of our study is to collect empirical AAL spatial data which can help enhance human-robot interaction (HRI). In order to improve *Rolland*'s current dialogue system, we should rely on a corpus that contains dialogue interactions that would most probably occur when users drive a wheelchair that undertakes their spoken commands and recognises gesture input. In section 2 we present some related work, while subsection 2.1 introduces BAALL and Rolland. In section 3 we discuss our study and its results (3.1), and in section 4 we present some future prospects.

2. Related Work

The three 'axes' of our research are:

i) Speech and gesture interaction
ii) Spatial domain
iii) Assistive environments.

As these axes are miscellaneous, there is lack of literature combining all of them. In the next paragraphs we describe related work combining two of the aforementioned axes (i and ii as well as i and iii).

Interaction between speech and gesture has been researched deeply the last years particularly in the spatial domain. Soma and Wachsmuth (2001) consider gestures as an inherently space-related modality and Kopp (2005) points out that gestures have sufficient specificity to be communicative of spatial information. Fricke (2007) analysed speech-accompanying gestures in way-to-destination answer in German.

The question that arises here is whether people from different *locales*, i.e. combinations of language and culture, express the same meaning by means of the same gestures. Anastasiou (2011a) points out that gestures have to be *locale*-dependent and gesture recognition software should be customised according to the *locale*. From the viewpoint of speech, an issue to be addressed is in which natural language spatial concepts are more researched. Tenbrink (2011) points out: “A range of controversies in the literature (…) may be reconciled by realising the diversity of spatial concepts (…). This is true for the well-researched English language, (…), but also for other cultures and languages, which have only partly been explored so far with respect to their spatiotemporal conceptualisations”.

As far as research in assistive environments is concerned, there is related work focusing on applications based on human-computer interaction (HCI) by means of either speech or gesture, but rarely both. One of the initiatives, which combine both modalities, is from Goetze et al. (2010) who designed a multi-media reminding and calendar system as part of a personal activity and household assistant for acoustic sound pick-up, processing, enhancement, and analysis. Furthermore, Neßelrath et al. (2011) designed a gesture-based system for context-sensitive interaction with a smart kitchen.

Last but not least, sensors are very often deployed in assistive environments (see Becker et al., 2009). Gesture control technology was developed for rehabilitation purposes by GestureTek[1], while the *Nintendo Wii* and *Microsoft Kinect* systems are used in assisted living communities, among others, to place the seniors into virtual sport and thus exercise indoors.

2.1 Related Work

The Bremen Ambient Assisted Living Lab[2] (BAALL) (see Krieg-Brückner et al., 2010) at the German Research Center for Artificial Intelligence (DFKI) in Bremen is an apartment suitable for the elderly and people with disabilities. It is 60m^2 and has all standard living areas, i.e. kitchen, bathroom, bedroom, and living room. BAALL has intelligent household appliances and furniture, e.g. separate kitchen cabinets can be moved up and down.

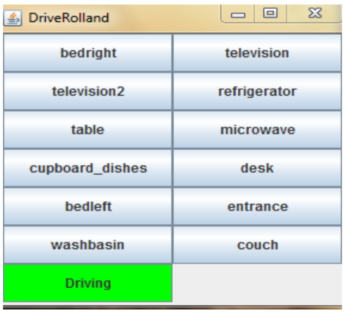
In BAALL the autonomous wheelchair *Rolland* offers mobility assistance; *Rolland* has a speech input control interface and navigation can be performed through a dialogue system; control with phone or PC tablet is also possible. The current limitations of human-robot/*Rolland* interaction are:

i) Minimal language support
ii) Lexicogrammatical limitations
iii) Absence of gesture recognition.

We explain how we mitigate those limitations with our following study.

3. User Study

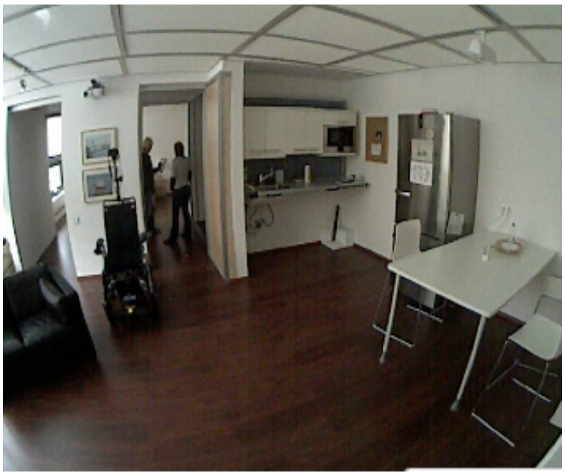
Our user study includes a real-life everyday scenario of a wheelchair user to navigate in her environment (go to the bathroom, eat sth. in the kitchen, take a book and read it on the sofa)[3]. The design and actual preparation of the study had various challenges. Those challenges included programming new destinations for Rolland and developing software for controlling it remotely[4] (Screenshot 1), acquiring and positioning appropriate technical equipment (cameras, camera stand, recording software), and test subjects recruitment.



Screenshot 1. User Interface for *Rolland'*s remote control

During the preparation of written and spoken instructions[5], there were several issues to be addressed. One was how whether we should explicitly state to the participants to perform gestures. The risk here is that they would be biased towards the demonstrated gestures. Our decision was to show an example-gesture which would not come up during the study rather than stating it explicitly.

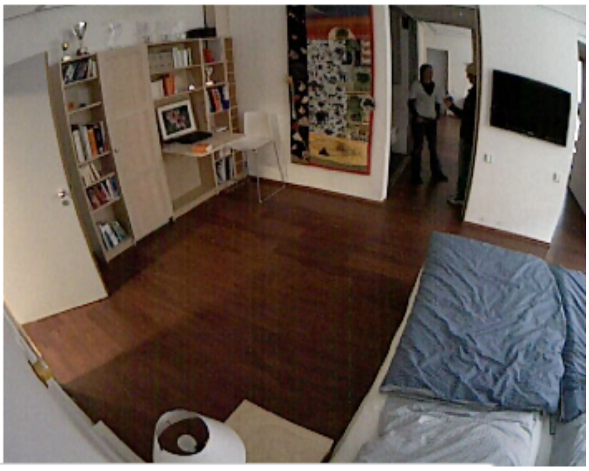
The experiment lasts about 20 minutes and is Wizard-of-Oz (WoZ) controlled, i.e. a researcher is in her office looking via camera software what is going on in BAALL (two views in Pictures 1 and 2).



Picture 1. Living room and kitchen

[1] http://www.gesturetek.com/, 11.10.11
[2] www.baall.net, 11.10.11
[3] Rolland answers neutrally: “Yes, I’m driving you there”, “OK”, etc.

[4] Acknowledgments for these two tasks are due to Bernd Gersdorf for his valuable support.
[5] These instructions are available at http://ai.cs.uni-sb.de/~stahl/d-anastasiou/DiaSpace/Resources/



Picture 2. Bedroom and desk

Another researcher is inside BAALL giving instructions to the participants and following them during task execution. The participants were requested to perform the tasks, i.e. go to specific rooms and do daily activities by means of natural speech and gesture interaction.

In the end of each session of the experiment, we followed a retrospective protocol (Dorst and Dijkhuis, 1995); participants were asked to go through the tasks that they just performed and speak loud what they were thinking. They were also asked to recommend future improvements of the human-wheelchair communication.

The participants used a Bluetooth microphone headset and the experimenter in the office could see and hear what they said or how they gestured through live audio and video streaming; the experimenter had a direct, noise-free access to the speech input. It is worth noting that in many WoZ studies with dialogue systems, some level of noise is added to the input available to the experimenter to mimic the challenges faced by the final system when operating in a real environment. Although a dialogue system is implemented in *Rolland*, it was not used in this study, as our focus was to gather empirical speech and gesture data and not evaluate the dialogue system's performance. Moreover, the grammar used in the speech recognition system is limited and not all users' commands would have been covered in this grammar.

As far as the situation parameters in terms of mutual 'visibility' between cameras and participants is concerned, two IP cameras are available in BAALL (one in the living room and one in the bedroom) which cover most of the BAALL. A floor plan in BAALL is available online[6]. An SLR camera was placed on the back of *Rolland*, so users can gesture "in front" of it.

For our actual study we recruited 20 German students (mean age 26). We considered it important that the participants should not have lived many years abroad, because their gesture might have been influenced by the foreign *locale*. People coming from different *locales* will be recruited at a second stage of the experiment[7]. Then a comparative analysis between cross-lingual spatial spoken commands and gestures will be conducted. In further user studies we plan to have elderly people as participants, as these are the main target audience of AAL. However, the goal of the study presented in this paper is to collect speech and gesture data and not make a comparative analysis based on the age of the participants.

3.1 Results

Here we refer to some results of our aforementioned user study. A relatively unexpected fact is that the example gesture was considered by the majority of the subjects as a robot-activating call and not as a modality to control *Rolland*.

In total we collected 317 spoken commands. In our written instructions we did not include the words representing the rooms, i.e. 'living room, 'bathroom', but the activity instead, e.g. 'I want to wash my hands'. Surprisingly, 20% gave as a command the activity rather than the destination point. 30% were supposedly "testing" *Rolland* by giving as a command both the destination point and the activity interchangeably. The remaining 50% gave commands to the destination point. Some subjects repeated their commands, as they found that *Rolland* reacted slowly and might have not heard the users properly. Some commands, such as "stop here", "go closer", "turn here", "turn on the light", though not specified in the instructions, were intuitively uttered. Moreover, context-sensitive commands were given, such as "come here" instead of "come to the sofa". We saw a diversity of the spoken commands based on the research discipline of the subjects; students of computer science and computational linguistics tried more 'expected' commands for a machine than students from linguistics.

As far as gesture frequency is concerned, 35% of the participants (7 sessions) performed a gesture. In 2 out of 7 sessions participants employed at least one gesture during a session. Most of gestures were deictic, as expected and happened mostly in cases when something went wrong, e.g. *Rolland* "parked" too far from the participant.

It is worth noting that nobody of the participants realised that the experiment was WoZ, believing indeed that *Rolland* moved based on their own commands.

As far as the retrospective protocol is concerned, most of the participants said that *Rolland* "parked" too far and a person with disabilities would not have been able to reach it. Another participant (female) said that she expected *Rolland* to have a female voice (see discussion in Crowell et al. 2009).

4. Discussion and Future Prospects

AAL is a purposeful research field with various research subfields combined; although AAL often focuses on ICT of medical care, dialogue systems and gesture interaction to control personal assistants, like an intelligent wheelchair, is another important research field. Our observational user study collects multimodal data of what wheelchair users say and gesture in real-time using

[6] http://ai.cs.uni-sb.de/~stahl/d-anastasiou/DiaSpace/Resources/

[7] Instructions and *Rolland*'s backchannel feedback will be respectively in the language in question.

intuitive and natural dialogue. Our analysis and usability evaluation will be conducted on recorded videos with a future lexical enrichment of the grammar and gesture annotation. We will use the tool ANVIL (Kipp et al., 2007) which allows annotation on multiple tracks and within gesture stroke movements.

Regarding sensing technology, in further user studies *Kinect* sensors will be placed in BAALL and 3D body data will be collected and analysed. Moreover, a wide-angle lens is required for the SLR camera to cover the whole gesture space of the wheelchair users.

Based on the results of our user study we will compile a multimodal corpus, then design an ontology, and draw conclusions on diversity regarding intuitive spatial gesture-speech interaction in an assistive environment.

5. Acknowledgments

We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition - Subproject I5-DiaSpace.

6. References

Anastasiou, D. (2011a). Speech Recognition, Machine Translation and Gesture Localisation. In *Proceedings of TRALOGY: Translation Careers and Technologies: Convergence Points for the Future*, Paris, France, 2011.

Anastasiou, D. (2011b). Gestures in assisted living environments. In *Proceedings of the Ninth International Gesture Workshop*, Athens, Greece.

Becker, E., Le, Z., Park, K., Lin, Y., and Makedon, F. (2009). Event-based experiments in an assistive environment using wireless sensor networks and voice recognition. In *Proceedings of the Second International Conference on PErvasive Technologies Related to Assistive Environments* (PETRA), Crete, Greece.

Crowell, C.R., Scheutz, M., Schermerhorn, P., and Villano, M. (2009). Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), St. Louis, MO, USA, pp. 3735--3741.

Dickinson, A., Arnott, J., and Prior, S. (2007). Methods for human-computer interaction research with older people. *Behaviour & Information Technology*, 26(4), pp. 343--352.

Dorst, K., Dijkhuis, J. (1995). Comparing paradigms for describing design activity. *Design Studies,* 16, pp. 261--274.

Fricke, E. (2007). *Origo, Geste und Raum: Localdeixis im Deutschen*. PhD Thesis. Berlin: de Gruyter.

Goetze, S., Moritz, N., Appell, J.E., Meis, M., Bartsch, C., and Bitzer, J. (2010). Acoustic user interfaces for ambient-assisted living technologies. *Inform Health Soc Care*, 35(3-4), pp. 125--43.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Kipp, M., Neff, M., and Albrecht, I. (2007). An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation Journal*, 41, pp. 325--339.

Krieg-Brückner, B., Röfer, T., Shi, H., and Gersdorf, B. (2010). Mobility assistance in the Bremen Ambient Assisted Living Lab. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 23(2), pp. 121--130.

McNeill, D. (1992). *Hand and Mind – What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, London.

Neßelrath, R., Lu, C., Schulz, C. H., Frey, J., and Alexandersson, J. (2011). A gesture based system for context-sensitive interaction with smart homes. Wichert, R., Eberhardt, B. (Eds.), *Proceedings of the Fourth AAL-Kongress*, pp. 209--222.

Sowa, T., Wachsmuth, I. (2001). Interpretation of shape-related iconic gestures in virtual environments. In Wachsmuth, I., Sowa, T. (Eds.), *Proceedings of the Fourth International Gesture Workshop, Gesture and Sign Language in Human-Computer Interaction*, 2298/2002, pp. 21--33.

Tenbrink, T. (2011). Reference frames of space and time in language. *Journal of Pragmatics,* 43(3), pp. 704--722.