Wizard-of-Oz Data Collection for Perception and Interaction in Multi-User Environments

Petra-Maria Strauß[⋆], Holger Hoffmann[‡], Wolfgang Minker[⋆], Heiko Neumann[†], Günther Palm[†], Stefan Scherer[†], Friedhelm Schwenker[†], Harald Traue[‡], Welf Walter[†], and Ulrich Weidenbacher[†]

University of Ulm
[⋆]Dept. of Information Technology
[†]Dept. of Neural Information Processing
[‡]Dept. of Psychosomatics and Psychotherapy
Ulm/Donau, Germany
{firstname.lastname}@uni-ulm.de

Abstract

In this paper we present the setup of an extensive Wizard-of-Oz environment used for the data collection and the development of a dialogue system. The envisioned Perception and Interaction Assistant will act as an independent dialogue partner. Passively observing the dialogue between the two human users with respect to a limited domain, the system should take the initiative and get meaningfully involved in the communication process when required by the conversational situation. The data collection described here involves audio and video data. We aim at building a rich multi-media data corpus to be used as a basis for our research which includes, inter alia, speech and gaze direction recognition, dialogue modelling and proactivity of the system. We further aspire to obtain data with emotional content to perfom research on emotion recognition, psychopysiological and usability analysis.

1. Introduction

The objective of the presented work is the development of components and technologies for intelligent and user-friendly human computer interaction in multi-user environments. Future multimodal systems are endowed with perceptive skills from different sensory channels (vision, hearing, haptic, etc.) as well as conversational skills to be able to capture and analyse spoken, gestural, as well as emotional utterings. The basis for that is the development of adaptive system components to capture the spacial, temporal, and user specific context of an interaction process. Integration of perception, emotion processing, and multimodal dialogue skills in interactive systems will not only improve the human-computer communication but also the human-human communication over networked systems.

In the framework of the research project Perception and Interaction in Multi-User Environments we have set up a Wizard-of-Oz (WOZ) environment to assist the development of a Perception and Interaction Assistant and to facilitate data collection and interaction model building. A human so-called "Wizard" simulates a dialogue system (or essential components thereof) that interacts with the human users just like the envisioned final system. Ideally, the users do not notice the simulation and behave as if they were interacting with an automatic system rather than a human.

The collected data (interaction model, acoustic and video signals, language and semantic models) will result in a rich multi-media data corpus to be used as a basis for our research.

Due to the various departments involved in the project the research performed is multi-faceted. One research direction is dialogue modelling and proactivity of the system. In this context, it will be explored how to find the right point in time for the system's pro-active involvement in the conversation. We also aim to analyse whether and to what extent dialogues between the human conversational partners differ from human-computer dialogues. We will further investigate cues to detect whether the dialogue falls within the specified domain.

Other aspects of our research include speech and gaze direction recognition. We also aspire to obtain data with emotional content to perform emotion recognition. Further, psycho-biological usability and acceptability analyses, such as the analyses of emotion patterns and user actions using the recorded video data will be carried out.

This paper is structured as follows. The following chapter 2 addresses related work in the area. Chapter 3 describes the data collection scenario. We present the Wizard-of-Oz setup in section 4. Section 5 talks about the data collection before the paper is concluded in chapter 6.

2. Related Work

Similar work focusing on the possibilities offered by enriching system with the perception of the user's state, context, and needs is currently addressed by CHIL (Stiefelhagen et al., 2004) which aims at developing environments in which computers serve humans giving them more freedom to focus on the interaction with other humans. One of the tools developed in the scope of CHIL is the connector (Danninger et al., 2005). It perceives activities, preoccupations, and social relationships of its users in order to determine their disposability and the appropriate device for communication. Another tools is the memory jog, a context- and content-aware service providing the user with helpful background information and memory assistance related to an ongoing event or other participants.

The earlier projects SMARTKOM (Wahlster et al., 2001; Beringer, 2003) and VERBMOBIL (Wahlster, 2000) dealt

1)	U1: Look at this weather, isn't it beautiful!	`keyword_spotting`
2)	U2: Absolutely. Why don't we leave it for today?	
3)	U1: Good idea. I'm getting **hungry** anyway.	
4)	U2: Do you want to **go eat** something? I feel like a huge **pizza**!	([U2][eat]suggest) `start_restaurant_application` ([U2][food:italian]suggest)

5)	U1: That **sounds good**! Let's go somewhere where we can **sit outside**. **[looks at computer]** Is there anything **in the park** maybe?	([U1][setting:outside_seating]request)
		`start_interaction`
		([U1][location:park]request)
		`check_database`

6)	S: Unfortunately not. However, I can suggest Pizzeria Napoli. It is close to the park and has a patio.	`$prompt` ([S][restaurant:Napoli]suggest)
7)	U1: That **sounds good**. What's the **address**?	([U1][restaurant:Napoli]confirm) ([U1][Napoli:address]request)
8)	S: Pizzeria Napoli is situated on Forest Avenue number fifteen. Would you like to see the menu?	`$prompt`
9)	...	...

Figure 1: Example dialogue and system reaction

with integrating speech and gesture processing for a natural interaction with the system in different scenarios.

3. Data Collection Scenario

The envisioned dialogue system will interact with two human users. The computer acts as an independent dialogue partner. Passively observing the dialogue between the human users and capturing the relevant conversational context, the system should take the initiative and get meaningfully involved in the communication process when it is required by the conversational situation.

The data collection scenario is restaurant selection. Two human users are engaged in a conversation talking about anything. The system listens passively, spotting keywords in order to notice when the conversation topic changes to the specified domain. At one point, the conversation switches over to the restaurant domain: The users decide they want to eat out and discuss their preferences about choosing the appropriate restaurant. The system notices the topic change and starts listening attentively to the conversation to store all of the relevant data in the dialogue history. One of the two human users is the system's main interaction partner. When this dialogue partner shows will to communicate with the system by turning his attention towards it, either explicitly by addressing the system directly or implicitly by looking at it, the system gets involved in the conversation. It assists the users in finding a suitable restaurant by providing information about restaurants and presenting their menus on the screen. For an example dialogue and the corresponding system reaction refer to Figure 1. Keywords that trigger system reaction are in boldface.

4. Wizard-of-Oz Setup

The dialogue system in use for the Wizard-of-Oz setup is built so as to simulate the final system's behaviour as closely as possible. Appropriate tools are necessary for a Wizard to be able to simulate or control ideally all parts of a typical dialogue system: Speech recognition, semantic analysis, dialogue management, domain knowledge base, natural language generation, and text-to-speech conversion [refer to Figure 2]. As a component becomes functional during the simulation process, easy integration into the system should be ensured. In doing so, the system evolves from pure simulation to a partial system, eventually resulting in the final system.

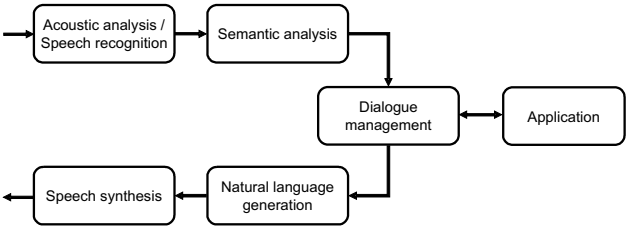

Figure 2: Spoken Language Dialogue System

Our system architecture is already partly functional. Speech recognition is currently performed by the Wizard. The final system will employ a key word spotting mechanism during passive mode. The system waits for certain words specified by the associated grammar which are then passed on to the dialogue management component. In active mode, a regular speech recogniser tries to transcribe the complete utterances. Semantic analysis extracts the meaning of what was recognised before and passes it on to the dialogue management. The domain knowledge base is in our case a database containing restaurants and their properties such as cuisine, setting, address, and phone-number.

Dialogue management is already performed automatically by an initial version of the system. Using a VoiceXML front-end, the Wizard enters commands that are understood and processed by the system implemented in Java: Data base queries are triggered and natural language generation is performed automatically. However, the Wizard can modify a system utterance before it is synthesised and played back to the users. We use a MBROLA voice in a FreeTTS text-to-speech synthesis engine.

The setup of the system is shown in Figure 3. The human dialogue partner A1 and A2 interact with the system A3 which is operated by the Wizard situated in a different room. A1 is the system's main interaction partner. The following programs are running on the computer A3: First of

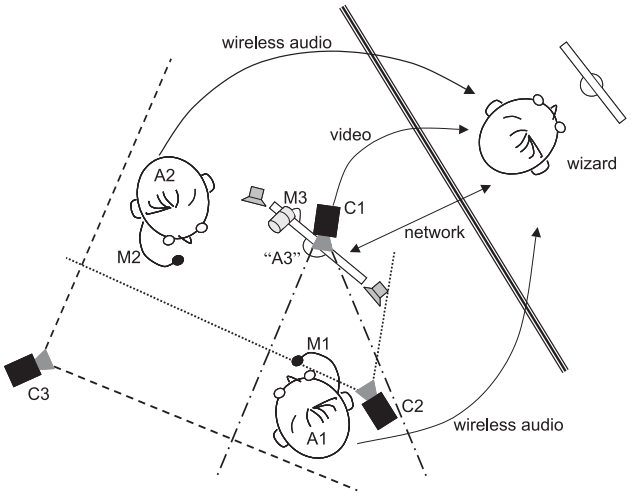

Figure 3: Data collection setup

all, the dialogue system server which produces acoustic and visual output for the dialogue partners. Acoustic output is the synthesised dialogue output of the system, visual output consists of the restaurants' menus shown on the screen. The picture of camera C1 records the main user A1 for online gaze detection analysis. It is recorded on this machine and made available to the Wizard. Finally, the input from microphone M3 is recorded.

The Wizard's computer is connected to the dialogue system server via a network connection. The Wizard operates the system hearing what the users are saying through microphones M1 and M2. Their signals are transmitted via wireless connection and recorded on this computer. A client for the camera server shows what is going on in the room (via C1) and what person A1 points his attention to.

The psycho-biological usability and acceptability analysis is not possible online. Therefore, the dialogue situation has to be recorded for later play-back for selected studies in the laboratory. For this reason, we installed two further camcorders: One to record user A1's perspective (C2), the other one to capture the entire scene (C3).

5. Data Collection

Audio Processing The speech signals are recorded by three different microphones. One for each human dialogue partner (M1, M2)[1] and one room microphone (M3)[2] to capture the entire scene including the system output. The data is used to train the speech recogniser as well as for emotion recognition, it's transcriptions for dialogue modelling. All audio data is recorded at 16 kilohertz with 16 bit resolution – the standard quality for speech recognition. External sound cards are used for improved quality and to be independent of the recording computer.

Emotion Processing Apart from information transmitted by the explicit channels in human to human communication, such as speech, there is also implicit information (Cowie et al., 2001). Understanding the other party's emotions is one of the main tasks in order to decode the implicit information. Decoding this information is not only necessary for "healthy" human to human interaction but becomes more and more interesting for human-computer communication. For that purpose, it is also necessary to record emotional data. In order to obtain such material, we record not only so-called free dialogues without providing any kind of guidelines for the users but intend to stimulate additional user emotions by giving the users specific problems that need to be solved during the dialogue process. An example of such a problem is shown in Figure 4. Another way to obtain emotional data is to direct the Wizard's reaction in a certain way. The Wizard could for instance purposely suggest a restaurant that does not match the users' preferences or that is situated far away in a completely different place.

Situation:	two colleagues A and B want to go eat something after a long day at work
Person A	- main dialogue partner of computer
	- loves Greek food
	- had a bad day (is tetchy)
	- is short in money
Person B	- is disgusted by garlic
	- likes plain German food

Figure 4: Example scenario description

Video Processing Social interaction between humans is not only limited to verbal communication but also visual communication plays a significant role. In a dialogue scenario, non-verbal communication is particularly characterised by analysing the gaze of a dialog partner. Directed gaze signalises attention while averted gaze signalises inattentiveness (see Figure 5).

In the WOZ-environment we employ images of the head from the main interaction partner of the system (A1) captured by camera C1. The goal of the video processing unit is to analyse whether the gaze of A1 is directed to the Wizard (A3) or to the dialogue partner (A2).

To estimate the gaze direction of A1, we extract two important visual cues from the face to discover the focus of attention from a persons head: (1) the orientation of the head (head pose) and (2) orientation of the eyes within their socket (eye gaze). For this purpose, we initially estimate the head pose (Gee and Cipolla, 1994; Krüger et al., 1997) of manually labelled heads and further investigate temporal redundancies of image sequences (Ke et al., 2003). Additionally, the eye gaze of A1 is estimated by analysing the eye region depending on the detected head pose. Our approach is motivated by experimental observations of (Langton et al., 2000; Sinha, 2000) where they find that the scleral contrast between the dark and bright regions in the eye is used as a key feature by humans to detect where someone is looking. Finally, both informations (head pose and eye gaze) are combined to get a robust mechanism that is able to reliably estimate the focus of attention of user A1.

Psychophysiological and Usability Analysis Analysis of psycho-biological usability and acceptability cannot be performed online. Therefore, we record the entire scene on further camcorders (C2, C3) to be able to later reconstruct the situation in the laboratory. Camera C2 records

[1] AKG 97L with AKG WMS40

[2] AKG 1000S

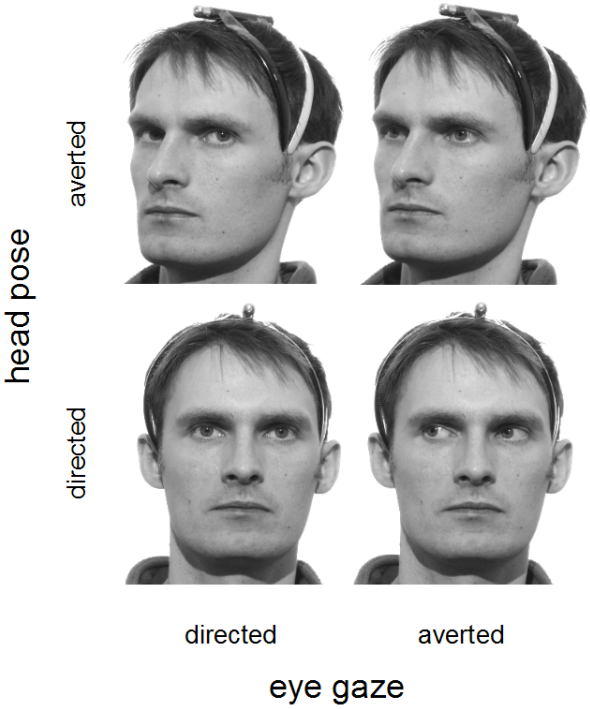

Figure 5: This figure shows four different combinations of head pose and eye gaze. In the lower left head pose and eye gaze are both directed to the viewer which signalises a high degree of attention. In the lower right and the upper left we illustrate two conditions where only one component, either head pose or gaze, is directed to the viewer which leads to a lower level of perceived attention. In the upper right both components are averted, suggesting the impression of inattentiveness.

the scene from the point of view of the system's main interaction partner A1 including both dialogue partners, i.e. the human dialogue partner A2 and the computer screen (A3).

In order to evaluate the scenario in terms of usability, acceptability (behaviour towards the system) and the emotional state ("emotion patterns") of the user, recorded video data (C2, i.e. from A1's perspective) is presented to several subjects as input. To simulate the real situation the screen is divided into two halves: On the left side user A2 is visible; the right side presents the computer screen A3. Subjects sit in front of the computer screen simulating the original scene and receive varying instructions (e.g. finding out the relevant information). Various behavioural and psychophysiological data are recorded during the session to extract typical patterns descriptive of a defined emotion. Data collected are eye-tracking parameters (pupillary dilation, saccades and fixations), electroencephalography (EEG) and multimodal physiological signals (heart rate, temperature, skin conductance level, electromyography (EMG) and respiration rate), which are synchronised by a logging system.

The aim of this subproject is to obtain emotional response patterns extracted from multiple subjects that can be applied in the evaluation of A1's reaction to the real scenario. This measure provides online data of A1's emotional patterns.

6. Conclusion

We have set up the WOZ-environment described in this paper for data collection and development of the final dialogue system. The collected data is used for research on the presented objectives. We currently started our recordings with users unfamiliar with the fact that the system is only simulated and has not fully evolved yet. For further analysis, we ask the users to fill out a questionnaire upon using the system. In the future, we also aim at including an avatar as a personification of the system and virtual dialogue partner.

7. Acknowledgements

This work has been supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg (Az:23-7532.24-13-19/1).

8. References

N. Beringer. 2003. The smartkom multimodal corpus - data collection and end-to-end evaluation. In *Colloquium of the Department of Linguistics*, University of Nijmegen.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. 2001. Emotion recognition in human-computer interaction.

M. Danninger, G. Flaherty, K. Bernardin, H. K. Ekenel, T. Köhler, R. Malkin, R. Stiefelhagen, and A. Waibel. 2005. The connector: facilitating context-aware communication. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 69–75, New York, NY, USA. ACM Press.

A. H. Gee and R. Cipolla. 1994. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647.

Wang Ke, Wang Yanlai, Yin Baocai, and Kong Dehui. 2003. Face pose estimation with a knowledge based model. *IEEE Int. Conf. Neural Networks and Signal Processing*, pages 1131–1134.

N. Krüger, M. Pötzsch, and C. von der Malsburg. 1997. Determination of face position and pose with a learned representation based on labelled graphs. *Image Vision Comput.*, 15(8):665–673.

S.R.H. Langton, R.J. Watt, and V. Bruce. 2000. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Science*, 4(2):50–59.

P. Sinha. 2000. Last but not least. here's looking at you, kid. *Perception*, 29:1005–1008.

R. Stiefelhagen, H. Steusloff, and A. Waibel. 2004. CHIL - Computers in the Human Interaction Loop. In *NIST ICASSP Meeting Recognition Workshop*, Montreal, Canada, May.

W. Wahlster, N. Reithinger, and A. Blocher. 2001. Smartkom: Multimodal communication with a life-like character. In *Proceedings of Eurospeech 2001, 7th European Conference on Speech Communication and Technology*, volume 3, pages 1547–1550.

W. Wahlster. 2000. *Verbmobil, Foundations of Speech-to-Speech Translation*. Springer, Berlin.