The PIT Corpus Of German Multi-Party Dialogues

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

University of Ulm
[⋆]Inst. of Information Technology
[†]Inst. of Neural Information Processing
[‡]Inst. of Medical Psychology
Ulm/Donau, Germany
{firstname.lastname}@uni-ulm.de

Abstract

The PIT corpus is a German multi-media corpus of multi-party dialogues recorded in a Wizard-of-Oz environment at the University of Ulm. The scenario involves two human dialogue partners interacting with a multi-modal dialogue system in the domain of restaurant selection. In this paper we present the characteristics of the data which was recorded in three sessions resulting in a total of 75 dialogues and about 14 hours of audio and video data. The corpus is available at http://www.uni-ulm.de/in/pit.

1. Introduction

In this paper we present the PIT corpus of German multi-party dialogues recorded in the context of the *Competence Centre Perception and Interactive Technologies*[1] (PIT) at the University of Ulm. PIT joins various institutes of the University to perform interdisciplinary research in the field of advanced human-computer interaction. Our objective is to develop components and technologies for intelligent and user-friendly human-computer interaction in multi-user environments.

Future dialogue systems will be endowed with more human-like capabilities: They should be adaptive in terms of the users' needs and preferences. They should be flexible in terms of the number of users that interact with the system. They should be endowed with perceptive skills from different sensory channels (vision, hearing, haptic, etc.). And they should be aware of the users' emotional state and possess enhanced conversational skills, just to name a few of the desired character traits of future dialogue systems. Integration of perception, emotion processing, and multimodal dialogue skills in interactive systems will not only improve the human-computer communication but also human-human communication over networked systems.

At present, research on multi-party interaction is very popular especially in the context of the meeting scenario. Various corpora have been published, e.g. the ICSI (Janin et al., 2003) and the AMI (Carletta et al., 2005) corpus. The meeting scenario requires intelligent computer systems to enhance and assist the human communication during meetings, however, our aim is to integrate the computer system as an equal dialogue partner in the communication with several humans.

As far as the authors are aware, there is no existent collection of data stressing the designated features important for our research. Thus, we built our own data corpus presented in this paper. The corpus comprises transcribed audio and video data. It emerged from Wizard-of-Oz (WOZ) recordings conducted in 2006 and 2007 in the framework of the research project 'The computer as a dialogue companion - Perception and interaction in multi-user environments'. The obtained data form the basis for our research, to develop mechanisms for sophisticated human-computer interaction, which we will not go into detail here. In this paper we present our corpus which we believe to be of great value for the research community in human computer interaction.

The paper is structured as follows. The following section introduces the scenario of the recordings. Section 3 briefly describes the recording setup. Section 4 presents the collected data in detail, including the participants, audio and video data. Section 5 concludes the paper.

2. Scenario

The PIT scenario is restaurant selection and composed of three dialogue participants: Two humans are discussing their choice of restaurant, the computer assisting them. The conversation takes place at a desk, one of the users (the main user 'U1') is sitting in front of the computer. Only this user interacts directly with the system, i.e. addresses it or expresses willingness to communicate by looking at it. The system, however, overhears the complete conversation between the participants.

The system acts as an independent dialogue partner which becomes active as soon as the users start to speak about the specified domain. At the beginning of the dialogue the users talk about a random topic. As soon as the conversation enters the restaurant domain, the system 'listens' attentively. Upon an interaction request by the main user or proactively when required by the conversational situation, the system takes the initiative and gets meaningfully involved in the communication. The system is represented by an avatar (not for all dialogues, see below) and synthesised speech output. Further, a restaurant's menu, directions in form of a city map, or bus schedules are occasionally displayed on the screen. An example dialogue excerpt is shown in Figure 1 [2]. The dialogue partners were randomly

[1] www.uni-ulm.de/in/pit

[2] As the conversation language is German, the example was

U1:	Computer, could you please give us a Mexican restaurant in the city centre of Ulm.
S:	*The query for Mexican food resulted in two hits. My suggestions are restaurant Flamenco and restaurant Peppers.*
U2:	Ask him where they are.
U1:	Where are the Mexican restaurants?
S:	*Restaurant Flamenco offers Mexican and Spanish food for moderate prices, possesses a terrace and is located in the city centre next to the Danube. Restaurant Peppers offers ...*
U1:	In which street is the restaurant?
S:	*Flamenco is located at number eight Inselgasse.*
U1:	Great! How are the prices in this restaurant?
S:	[**Menu pops up on screen.**] **...**

Figure 1: Excerpt from one of the recorded dialogues (Strauß et al., 2007).

assigned different scenarios in form of descriptions of roles and situations that should be adopted in the conversation. The combination of assigned roles included amongst others employer and employee, business colleagues, friends, or a couple in a loving relationship. This procedure not only eased the beginning of the dialogue, it brought in variety in terms of different social contexts. Most times, the participants followed the role play quite well. In a few cases, however, the scenario was only used to stimulate the conversation and not referred back to later during the dialogue.

3. Data Collection Setup

The recordings of the data were performed in the Wizard-of-Oz setup described in (Strauß et al., 2006). The wizard simulates the envisioned final system's behaviour as closely as possible, replacing the system components that are not yet functional (such as e.g. the speech recogniser) using the tool described in (Scherer and Strauß, 2008). The setup of the system is shown in Figure 2. The human dialogue partners U1 and U2 interact with the system S which is operated by the wizard situated in a different room. U1 is the system's main interaction partner. The dialogues are recorded by three microphones and three cameras as follows. U1 and U2 each wear a lapel microphone (M1 and M2). The signals are sent via wireless transmission to the wizard's computer where they are recorded and played back to the wizard. Room microphone M3 records the entire scene including the system output. Camera C1 is placed as to record from the system's point of view, focusing on the face of U1. C2 is installed behind U1 in order to grasp U1's point of view, i.e. the screen and the other dialogue partner U2. C3 records the entire scene. Refer to (Strauß et al., 2006) for details on the technical equipment.

directly translated into English.

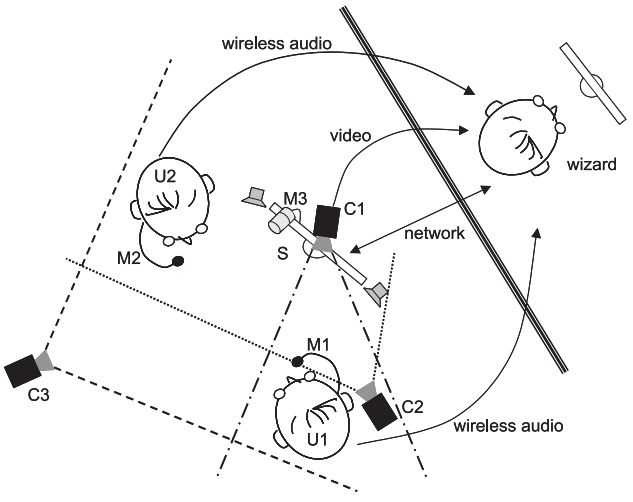

Figure 2: Data collection setup (Strauß et al., 2006)

4. Wizard Policies

Prior to recording the conversations it was necessary to assess several policies the wizard has to follow to assert uniform system behaviour throughout all dialogues in order to receive unbiased dialogue data. In general, there were two different kinds of targeted data: First, there was general or normal dialogue material and on the other hand emotional data. Both types of interaction followed the same principles: The conversation between the two dialogue partners started without wizard interaction in the beginning when the users acted out the provided scenarios. The differences emerge once the system has joined in the conversation, as pointed out in the following two paragraphs.

Standard Policy. During a standard recording the system, or wizard respectively, did not interrupt the conversations of the users. Speech recognition and language understanding was simulated to be perfect. Correct and helpful answers were given as frequently and as promptly as possible. The system's first and further interactions were triggered by the following situations:

- Reactively upon user U1 addressing the system directly
- Proactively on it's own behalf in order to make a significant contribution to the conversation (e.g. to report a problem in the task solving process)
- After a pause in the dialogue exceeding a certain threshold (if a meaningful contribution can be made)

In the case where U2 addressed the system directly, the utterance was recognised, however, no direct response was given. Yet, two different reactions were observed. Either, user U1 instantly took the turn and posed the same or similar request to the system, or U2's request was followed by a pause which again would mostly lead to an interaction of the wizard. After the users found a restaurant that pleased everyone the wizard generally closed the dialogue. In some cases, however, the participants decided to search for another locality, such as a bar for a cocktail after dinner etc. Overall, the recordings adhering to the standard policy resemble a perfectly working dialogue system.

Emotion Policy. In order to induce emotional behaviour of the users a specific emotion policy was used in several dialogues. During these recordings the wizard occasionally interrupted the users in order to appear rude. Furthermore, repetitive mistakes were made and wrong answers given.

Common mistakes included recognition and understanding errors that would sound very similar to the users' actual wishes, e.g. if the user was looking for a "not expensive restaurant" the wizard would return a selection of expensive restaurants. Additionally, the wizard would sometimes just pause to bore the users.
The emotions obtained using this strategy include anger, boredom and surprise. Emotions such as happiness and surprise were also induced by the standard policy, after receiving correct answers and useful informations. Surprise is often induced by the first interaction of the system, for example after the face of the avatar is shown to the users for the first time. In general, it is to say that the emotions expressed by the users are more moderate than artificial emotions played by actors, as in (Burkhardt et al., 2005). However, we consider these moderate emotions as more realistic and common in human computer interaction and therefore very useful for affective computing tasks.

5. Collected Data

The corpus consists of 75 dialogues from three recording sessions, refer to Table 1 for number of dialogues and durations. Session I was performed in 2006. At that stage, the system output consisted of only acoustic output (speech synthesis) and the display of the restaurant's menu in HTML format when required. For the second block of recordings (2007) the system was enhanced. The response time of the wizard was improved. An avatar was integrated to represent the system visually. Furthermore, street maps were included to be shown on the screen. For the third recording session (2007), the system was enhanced to also present bus schedules on the screen. Half (18) of the dialogues were recorded with the avatar on the screen, half without the avatar. Synthesised and visual output in form of menus and maps was the same for all recordings in this session. The shortest recorded dialogue was 2:43 minutes long (session III), the longest lasted 33:39 minutes (session II).

Session	I	II	III
Number of dialogues	19	20	36
Duration of session	3:47 h	4:18 h	5:40 h
Average dialogue duration	12 min	13 min	10 min

Table 1: Statistical information of the three recording sessions.

5.1. Participants

Participants (n=150) were students and employees of the University, who gave written consent to participate in this study. They were between 19 and 51 years of age (on average 24.4 years); 53 of them were female (4 at session I (10.5%), 18 at session II (45.0%), 31 at session III (43.1%)). Except for six participants, the native language of all participants was German.
In order to evaluate the dialogue system, several questionnaires had to be completed by the participants after the interaction with the system. The AttrakDiff (Hassenzahl et al., 2003) questionnaire was used to measure the attractiveness as well as the pragmatic and hedonic quality of the system. To evaluate the direct interaction between the human dialogue partner U1 and the computer system, a short version (n=16 items) of the SASSI (Hone and Graham, 2000) questionnaire was selected. Furthermore, data on participants' technical self assessment was collected.
The results show that the evaluation of the system significantly improved from recording session I to III in terms of attractiveness, usability and acceptance of the system. Further analysis of the data has to be done in order to reveal the impact of several changes (avatar, response time) on the evaluation of the system.

5.2. Audio Data

The audio data were recorded using three microphones: One lapel microphone for each participant and a room microphone to record the entire scene including the system output. The audio data were recorded at 16 kilohertz with 16 bit resolution. External sound cards were used to improve the recording quality.
The dialogues all follow a certain pattern marked by three phases: Each dialogue starts with a domain independent chat between the participants. The next phase of the dialogue is introduced at the point when the conversation switches over to the specified domain and the users start discussing their preferences and aversions in different aspects of the restaurant domain. The third part is characterised by the involvement of the dialogue system in the conversation to achieve the concerted task. The dialogues typically end when the users find a suitable restaurant and thank the system. Some recordings contain various iterations of the restaurant search, i.e. after finding one, instead of ending the dialogue, the users started to look for another restaurant (remaining in the third phase).

5.3. Video Data

Social interaction between humans is not only limited to verbal communication. Also visual communication plays a significant role. In a dialogue scenario, non-verbal communication is particularly characterised by analysing the gaze of a dialogue partner. Directed gaze signalises attention while averted gaze signalises inattentiveness. Therefore it is very interesting to extract and evaluate pose behaviour of individual subjects during the conversation.
The dialogues were video recorded from three different angles. Figure 3 shows the scene from the viewpoint of cameras C3 (long shot) and C1 (face of U1).
The goal is now to annotate each video frame (using the data from C1) with a specific class label to discriminate between video frames where the person (U1) attends to the system and video frames where the person attends to the human dialog partner (U2). This problem is closely linked to the field of automatic image annotation (Cusano et al., 2004), (Jeon and Manmatha, 2004), where a system automatically assigns metadata in the form of keywords to an image.
Here, we train an adaboost classifier (Viola and Jones, 2004) with a small subset of manually annotated image frames in order to automatically extract pose information

(direct gaze or averted gaze) from the video data which was acquired during the dialog session. We chose adaboost, because this approach is known to be very fast and efficient in detecting faces in images. We trained two different classifiers, one that finds frontal faces against background and one that finds averted faces against background. Finally, both classifiers are then applied to each of the remaining previously unlabeled video frames to determine their pose label. This information now enables us to statistically evaluate the amount of time user U1 spent focusing on the system as opposed to on the other user.

The video data can again be structured in three interaction phases considering the gaze direction of the main user U1. These phases differ from the dialogue phases described above. The first interaction phase is characterised by the conversation between the human dialogue partners before the first system interaction. During this time, there is almost no gaze directed towards the computer screen. The first interaction of the system initiates phase two. During this phase, U1's gaze switches between the computer and user U2, depending on speaker and addressee. The third phase is characterised by an object (other than the avatar) displayed on the screen: Generally, while a restaurant's menu, a street map, or bus schedule is shown on the screen, most of U1's gaze points towards the system. When the object is hidden, the dialogue returns to phase two.

5.4. Annotation

The data was transcribed at the utterance level and annotated with dialogue acts. Table 2 presents the basic tagset of dialogue acts we used and which proved suitable for our domain and dialogue manager requirements.

Tagset
suggest, request, inform,
accept, reject,
acknowledge, check
stall, greet, other

Table 2: Dialogue act tagset used on PIT Corpus.

6. Conclusion

In this paper we presented the PIT Corpus, a multimodal collection of 75 multi-party dialogues recorded in a Wizard-of-Oz setting. Each dialogue involved two human dialogue partners and a computer system, recorded with various microphones and video cameras. The corpus can be found on our website (http://www.uni-ulm.de/in/pit). We hope it to be of great benefit for the HCI research community.

Figure 3: Video recordings from the viewpoint of cameras C3 (left) and C1 (right) (Strauß et al., 2007).

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

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of german emotional speech. In *Proceedings of Interspeech 2005*.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*.

C. Cusano, G. Ciocca, and R. Schettini. 2004. Image annotation using SVM. In *Internet Imaging IV*, volume SPIE.

M. Hassenzahl, M. Burmester, and F. Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *In J. Ziegler & G. Szwillus (Hrsg.), Mensch & Computer 2003. Interaktion in Bewegung* , pages 187–196.

K. S. Hone and R. Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Nat. Lang. Eng.*, 6(3-4):287–303.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367.

J. Jeon and R. Manmatha. 2004. Using maximum entropy for automatic image annotation. In *CIVR*, pages 24–32.

S. Scherer and P.-M. Strauß. 2008. A Flexible Wizard-of-Oz Environment for Rapid Prototyping. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

P.-M. Strauß , H. Hoffmann, W. Minker, H. Neumann, G. Palm, S. Scherer, F. Schwenker, H. Traue, W. Walter, and U. Weidenbacher. 2006. Wizard-of-Oz Data Collection for Perception and Interaction in Multi-User Environments. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genova, Italy.

P.-M. Strauß , H. Hoffmann, and S. Scherer. 2007. Evaluation and User Acceptance of a Dialogue System Using Wizard-of-Oz Recordings. In *3rd IET International Conference on Intelligent Environments*, Ulm, Germany.

P. Viola and M. Jones. 2004. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.