Evaluation of the PIT Corpus Or What a Difference a Face Makes?

Petra-Maria Strauß[⋆], Stefan Scherer[†], Georg Layher[†], Holger Hoffmann[‡]

[⋆] Institute of Information Technology
[†] Institute of Neural Information Processing
[‡] Medical Psychology Section, University Clinic for Psychosomatic Medicine and Psychotherapy
University of Ulm, 89069 Ulm, Germany
firstname.lastname@uni-ulm.de

Abstract

This paper presents the evaluation of the PIT Corpus of multi-party dialogues recorded in a Wizard-of-Oz environment. An evaluation has been performed with two different foci: First, a usability evaluation was used to take a look at the overall ratings of the system. A shortened version of the SASSI questionnaire, namely the SASSISV, and the well established AttrakDiff questionnaire assessing the hedonistic and pragmatic dimension of computer systems have been analysed. In a second evaluation, the user's gaze direction was analysed in order to assess the difference in the user's (gazing) behaviour if interacting with the computer versus the other dialogue partner. Recordings have been performed in different setups of the system, e.g. with and without avatar. Thus, the presented evaluation further focuses on the difference in the interaction caused by deploying an avatar. The quantitative analysis of the gazing behaviour has resulted in several encouraging significant differences. As a possible interpretation it could be argued that users are more attentive towards systems with an avatar - the difference a face makes.

1. Introduction

Spoken language dialogue systems are increasingly being deployed in more and more different domains and applications. They are thus confronted with the challenge to meet new demands such as being aware of their users and the context in which they are interacting. The present work focuses on such a novel sort of dialogue system that acts as an independent dialogue partner in the interaction with two users. The users are having a conversation about any topic. As soon as they come to speak of the system's specified domain, the system starts listening closely and takes the initiative to interact as soon as it can contribute to the interaction. The system has been simulated as part of an extensive Wizard-of-Oz environment (Strauss et al., 2006) that has been used to record the PIT Corpus of multi-party dialogues (Strauss et al., 2008) which builds the basis of the evaluation presented in this paper. Figure 1 shows a screenshot of the system in use as it appears to the users.

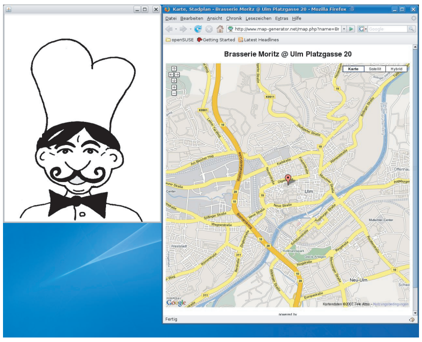

Figure 1: Screenshot of the system showing the avatar and a city map.

2. Experimental Setup

2.1. Data Recordings

The dialogues used for the evaluation have been collected in a Wizard-of-Oz environment that simulated a proactive spoken dialogue system interacting with two users in the example domain of restaurant selection. 76 dialogues were recorded over three recording sessions. The participants who took part in the recordings were not aware of the fact that the system was only a simulation. For a thorough description of the setup refer to (Strauss et al., 2006). The simulated system, i.e. the wizard interaction tool (Scherer and Strauss, 2008), has been enhanced between the different recording sessions in terms of interaction speed and features. The participants filled out a questionnaire prior and subsequent to the recordings to submit their subjective estimation of the interaction. Technical self assessment was performed which shows that all participants have a similar attitude towards technology (Strauss et al., 2008). In each interaction, one of the users was the system's main interaction partner leading the interaction with the system. Only these users' ratings were used for the evaluation. In Figure 2 a scene recorded with three different cameras is shown. The user displayed in the left most picture is the main user.

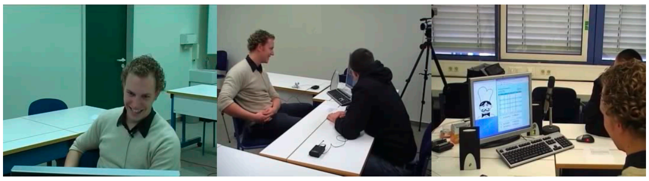

Figure 2: Typical scene taken from recording Session III with three different camera angles.

The data was recorded using different setups. An animated avatar[1] was deployed from Session II whereas in Session III half of the recordings were performed with, half without the avatar using only voice output. A set of Session III dialogues was further recorded using an emotion-eliciting strategy the wizard adopted in the interaction including simulated understanding or database errors of the system. Table 1 gives a short overview of the corpus; for a more detailed description please refer to (Strauss et al., 2008).

2.2. Procedure

The questionnaires used for the evaluation of the subjective ratings contained SASSISV, a shortened version of SASSI (*Subjective Assessment of Speech System Interfaces*) questionnaire (Hone and Graham, 2000) which consists of only 16 of 35 items. Due to the fact that it highly correlates (p<.001) in all scales with the original version it is taken to be valid for the present evaluation. In order to validate the results, a second questionnaire, AttrakDiff (Hassenzahl et al., 2003), was deployed additionally. While all results were at all times consistently found between both methods, only the SASSISV results are displayed in this paper.

For the gaze direction evaluation presented below, the dialogues have been annotated manually in terms of the gaze direction of the main interaction partner. The first comparison has been performed using 20 dialogues: 8 Session I (without avatar) and 12 Session II (with avatar). For the evaluation of Session III, 16 randomly chosen videos (8 with avatar and 8 without) were used. All dialogues were hand-annotated using a software written in Matlab. Manual annotation was chosen over an automatic approach due to the quality of data. Thorough inspection of the dataset has shown that the naturalness of the behavior of the users during the interaction - as it was not restricted at all - includes a manifold of different gaze direction, head pose and body pose configurations (for instance subjects tend to squint on the screen while orienting their head towards the secondary person). Current standard approaches (i.e. Viola Jones (Viola and Jones, 2004)) mostly neglect the actual gaze direction and simply use the head orientation as a rough estimate of the line of sight and therefore are not capable of achieving sufficiently accurate features for an analysis of human computer interactions. We consider the recorded data due to its naturalness and non sterile laboratory conditions a very valuable dataset that can be utilised in order to benchmark future approaches capable of dealing with such a large variety of behaviour. For such benchmarks ground truth annotations are necessary, which only humans can currently provide.

3. Evaluation

Evaluation has been performed with two different foci. Usability evaluation has been performed to acquire an overall assessment of the subjective ratings of the system. Gaze direction analysis has further been performed in order to assess the difference in the user's gazing behaviour if interacting with the computer versus the other dialogue partner.

3.1. Usability Evaluation

The evaluation of Session III dialogues of the PIT corpus is presented in terms of the system's usability. All four different setups are compared using SASSISV, see Figure 3.

A first look is taken at the dialogues recorded without emotion eliciting strategy in order to be able to assess the general usability of the system as well as to see what a difference a face (i.e. the avatar) makes in this context. Thus, considering only IIIa and IIIb it can clearly be seen that the dialogues with avatar achieve better ratings throughout all scales. This finding is consistent also if including the dialogues recorded with emotion eliciting strategy (IIIc and IIId) into consideration: The setups with avatar in general score higher than without avatar.

Considering the ratings over all four setups an interesting result can be observed: Surprisingly, the perfectly working system as it was simulated for IIIa dialogues did not always achieve the highest scores. The IIIc setup scored best regarding *Likeability* (IIIc mean: 5.68 vs. IIIa mean: 5.54), *Cognitive Demand* (IIIc mean: 5.55 vs. IIIa mean: 5.18) and *Annoyance* (IIIc mean: 1.5 vs. IIIa mean: 2.25). In the remaining scales (SRA, H, S) IIIa scored best. The Mann-Whitney U Test (Mann and Whitney, 1947) finds no significant differences between IIIa and IIIc, however a highly significant difference between IIIb and IIId for *System Response Accuracy* (p=.002) which is an expected result due to the fact that the system's behaviour was directed by the wizard to be less accurate. While the ratings go apart between all pairs of setups with and without avatar, it can be observed that the difference in the ratings is more prominent for the emotion-elicited dialogues.

The observed results show that users do not necessarily rate a perfectly working system best. Instead of feeling disturbed the users might feel more involved and stimulated by a system which commits a few mistakes once in a while and thus rate this setup better. Due to the small amount of dialogues the results, however, have to be treated with caution.

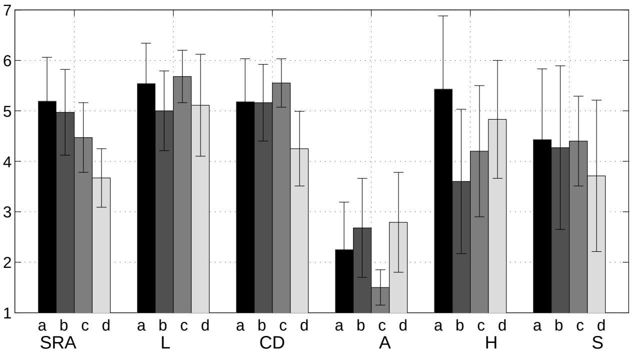

Figure 3: Usability evaluation of Session III dialogues over the different setups using SASSISV. The sets of dialogues are labeled a to d as described in Table 1. The scales are System Response Accuracy (SRA), Likeability (L), Cognitive Demand (CD), Annoyance (A), Habitability (H), and Speed (S).

[1] The avatar (see Figure 1) moves its mouth when speaking and occasionally blinks with one eye.

Session	I	II	III				Total
			IIIa	IIIb	IIIc	IIId	
Number of dialogues	19	20	37				76
			14	11	5	7	
Duration of session	3:47 h	4:18 h	5:40 h				13:45 h
Min dialogue duration	3:15 m	4:18 m	2:43 m				2:43 m
Max dialogue duration	26:11 m	33:39 m	18:24 m				33:39 m
Mean dialogue duration	12 m	13 m	9:44 m				11 m
Avatar	-	+	+	-	+	-	51.3%
Emotion-eliciting strategy	-	-	-	-	+	+	15.8%

Table 1: Statistical information of the PIT corpus.

3.2. Gaze Direction Analysis

The gaze direction of the main user has been analysed in order to assess user acceptance of the dialogue system. The difference in the behaviour of the user towards the computer versus towards the other human are investigated. It is further differentiated between addressing behaviour, i.e. gaze of the user while addressing the other dialogue partner, and listening behaviour, i.e. behaviour while listening to the other dialogue partner speak. Results of recording sessions I and II are given in the following. Between Session I and II the system has been enhanced and usability has been improved. Thus, the results have to be treated with caution as comparability of these sessions is limited. Evaluation of Session III (which includes both dialogues with and without avatar for direct comparison) is still ongoing, first results are presented further below. The dialogues are divided into three phases for this evaluation. Phase 1 denotes the part of the dialogue before the system's first interaction. Phase 2 describes the interaction between all three dialogue partners. Phase 3 denotes the phase in the dialogue when an object other than the avatar is displayed on the screen (e.g. restaurant menu or city map).

Table 2 presents the general gazing behaviour of the main user throughout the entire dialogue and according to the different phases. It lists the mean values of the percentages of the main user's gaze pointing to either dialogue partner, i.e. the other dialogue (U2) or the system (S), regardless of who is speaking or who is addressed. The values differ only slightly between the two setups. During Phase 1, the gaze towards U2 is predominant, as expected, as the system is not yet interacting. During Phase 2, the gaze goes more towards the system than to U2, however consistent for both setups. During Phase 3 a further object is displayed on the screen which attracts most of the U1's gaze.

The fact that during Phase 2 the gazing behaviour towards both dialogue partners yields very similar values for both setups does not necessarily denote that the system is treated equally by the user. There is also no obvious explanation why the system is looked at more than the other user. Thus, further investigation has been performed analysing Phase 2 in terms of the current speaker. First, a look is taken at the addressing behaviour. For this, only the points are considered when the main user U1 is addressing the respective dialogue partner. The results are shown in Table 3. It can be observed that the system is looked at slightly more when speaking to it when an avatar is deployed. The difference is however not very large and not statistically significant (p=0.23).

	no avatar (I)		avatar (II)	
U1 looking at	S	U2	S	U2
Phase 1	10.0%	70.4%	9.2%	58.1%
Phase 2	51.1%	39.1%	50.8%	37.7%
Phase 3	67.4%	27.5%	77.8%	18.8%
Dialogue	44.0%	43.9%	45.0%	40.5%

Table 2: Gazing behaviour of main user (U1) towards the second user (U2) and the system (S) according to different phases of dialogue.

	no avatar (I)		avatar (II)	
U1 addressing and looking at	S	U2	S	U2
Phase 2	72.3%	72.7%	77.4%	74.9%

Table 3: Addressing behaviour of main user (U1).

Table 4 presents the results of the listening behaviour, i.e. the other dialogue partner is speaking and user U1 is looking at the speaker. The system attracts the user's gaze significantly more when represented by the avatar (p=0.023) than if represented only by voice. The difference towards the values regarding the other user U2 is very prominent: The other dialogue partner (U2) attracts the main user's gaze to a large extent more than the system. However, deploying an avatar shows improvement in this aspect.

Evaluation of Session III is currently still in progress. First results are presented in the following. The left hand side

	no avatar (I)		avatar (II)	
U1 looking at speaker	S	U2	S	U2
Phase 2	37.6%	83.8%	55.2%	78.8%

Table 4: Listening behaviour of main user (U1).

blue plots denote the relative focus duration of the primary user towards any of the targets if an avatar is present or about to be shown on the screen while the right hand side red plots indicate a system configuration without an avatar. Figure 4 shows the gaze directions over the whole dialogue towards either of the three foci: Other dialogue partner (U2), system (S), or else. The stars denote significant differences (one star p<.05 and two star for p<.01), while rejecting the hypothesis that the data with and the data without an avatar stem from the same distribution, by applying a standard t-test. The most prominent result is the highly significant difference (**) between the amount of time the user is focusing on the second dialogue partner vs. the time spent focusing on the screen of the system, while there is an avatar present. Additionally, the difference of the relative focus time towards the system if an avatar is present vs. no avatar is also significant (*). Therefore, it could be argued that the system receives more attention of the user if a face is present as opposed to a sterile user interface.

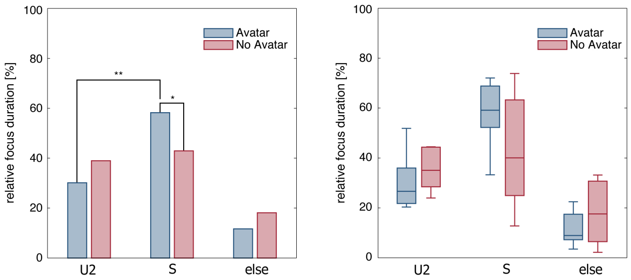

Figure 4: Gazing behaviour of main user towards either the second user (U2), the system (S) or elsewhere (else) over complete Session III dialogues is displayed. Significantly different results are marked with brackets on the left side figure (* ... p < .05 and ** ... p < .01). The mean values are displayed. On the right side figure box plots of the observed relative focus durations are shown.

In Figure 5 the plot is separated into the three different interaction phases mentioned above. Again blue plots indicate the presence, red plots the absence of an avatar. As expected, the user herein is focusing on the system significantly more time if an avatar is present. The other two phases again do not result in significant differences, which was also anticipated, since visually displayed information should provoke attention even without an avatar.

Finally, Figure 6 contains the same information as Figure 4 on the relative focus duration towards either of the three foci during the second interaction phase. For further comparison with the results received from the earlier recordings (Sessions I and II) please refer to Table 2 and the corresponding table of Session III Table 5. For the interaction with avatar no prominent difference can be observed to the former results. In the case of not deploying an avatar, the attention towards the second user rises while the attention towards the system sinks. Further evaluation shall be performed in order to investigate the cause of this phenomenon.

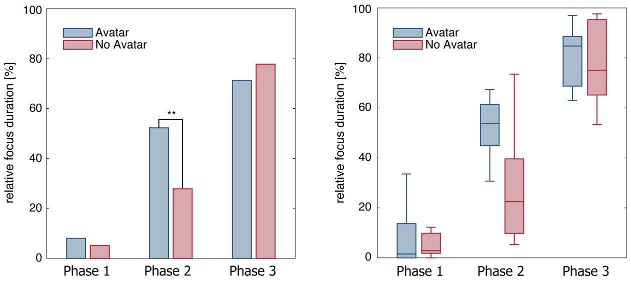

Figure 5: The relative focus duration of main user according to different dialogue phases of Session III dialogues towards the system (S) is displayed. Phase 1 corresponds to the time before the system interactions for the first time. Phase 2 represents the interaction between the system and the users with no visual information displayed and Phase 3 corresponds to the time when additional visual information is displayed. Significantly different results between the observations with or without an avatar are marked with brackets on the left side figure (** ... p < .01). The mean values are displayed. On the right side figure box plots of the observed relative focus durations are shown.

	no avatar (III)		avatar (III)	
U1 looking at	S	U2	S	U2
Phase 1	5.1%	56.1%	8.0%	50.5%
Phase 2	27.8%	49.5%	52.2%	34.6%
Phase 3	77.7%	16.4%	81.3%	15.2%
Dialogue	38.1%	41.7%	53.2%	33.1%

Table 5: Gazing behaviour of main user (U1) towards the second user (U2) and the system (S) according to different phases of dialogue in recording Session III.

4. Conclusions

The presented evaluation of the recorded dialogues of the PIT corpus shows very positive results. Overall, we believe that the results are very promising and show that a friendly face does have a significant impact on the usability as well as attentional behavior towards the system. It can clearly be observed that over all different evaluations, the avatar contributes significantly to the attractiveness and good usability ratings of the system and thus towards the system being accepted as an independent dialogue partner. This is confirmed also by the results of the gazing direction analysis: The user is focusing on the system more often than during the interaction without an avatar and is paying less attention towards the secondary user. Therefore, it might be argued that a simple avatar as ours is already shifting the attentional state of the primary user to a significant amount.

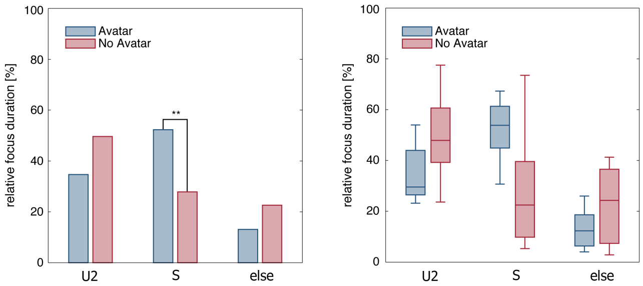

Figure 6: Gazing behaviour of main user towards either the second user (U2), the system (S) or elsewhere (else) over Phase 2 (interaction without visual information) of Session III dialogues. Significantly different results between the observations with or without an avatar are marked with brackets on the left side figure (** ... p < .01). The mean values are displayed. On the right side figure box plots of the observed relative focus durations are shown.

Acknowledgment

The presented work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

5. References

M. Hassenzahl, M. Burmester, and F. Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *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). *Natural Language Engineering*, 6(3-4):287–303.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

S. Scherer and P.-M. Strauss. 2008. A flexible wizard of oz environment for rapid prototyping. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

P.-M. Strauss, H. Hoffmann, H. Neumann, W. Minker, G. Palm, S. Scherer, F. Schwenker, H. Traue, and U. Weidenbacher. 2006. Wizard-of-oz data collection for perception and interaction in multi-user environments. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

P.-M. Strauss, H. Hoffmann, W. Minker, H. Neumann, G. Palm, S. Scherer, H. Traue, and U. Weidenbacher. 2008. The pit corpus of german multi-party dialogues. In *Proceedings of LREC 2008*.

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