ANALYSIS OF PARADISE MODELS FOR INDIVIDUAL USERS OF A SPOKEN DIALOG SYSTEM

Klaus-Peter Engelbrecht, Sebastian Möller, Robert Schleicher, Ina Wechsung

Quality and Usability Lab, Deutsche Telekom Laboratories TU Berlin
D-10587 Berlin, Germany

klaus-peter.engelbrecht@telekom.de

Abstract: In this paper, we analyze linear regression models predicting user satisfaction with spoken dialog systems. Correlations between interaction parameters and the judgments are analyzed user-wise, with the outcome that there is some congruity as to the parameters which are correlated. However, some users are less well predictable than others. A relation between the strength of correlations and user characteristics, impacting judgment behaviour and scale usage, is made. Prediction models of subgroups with specific characteristics are calculated. The results differ clearly depending on user characteristics. For some groups, transformation of the reply scale improves the prediction.

1 Introduction

The quality of spoken dialog services is a complex issue, involving multiple factors including the different modules of the system, user characteristics and resulting behaviour, and the context of use. In the past, there has been a continuous search for simple metrics which allow the comparison of systems or system variants by summing up all quality components to a single value. In PARADISE (PARAdigm for DIalog System Evaluation,[1]), it was suggested that user satisfaction, as it can be assessed with a questionnaire, is a good indicator for the system's overall quality. A linear regression equation can then be trained to predict the satisfaction score from interaction parameters (such as the dialog length). By this, a metric is derived from the data which is independent of direct user judgment.

It has been shown that mean values, corresponding e.g. to individual system configurations, can be predicted sufficiently with such a model [2]. Still, we fail to predict ratings of individual dialogs accurately. Being able to predict individual dialogs, however, opens many other applications for PARADISE such as monitoring dialogs online. Therefore, we would like to find the reasons behind this discrepancy.

The good predictability of mean ratings indicates that there is a close relation between parameters and judgments. However, such relations seem to differ between users. Therefore, we analyze relations between parameters and judgments for individual users and compare the results. For this, we use experimental data in which each user had to judge 14 tasks conducted with the same system. As we have few data points for each single user, we use correlation analysis to estimate the impact of parameters on judgment. Additionally, we check with linear regression for possible interaction effects between the parameters.

In the following section, we briefly describe the experimental set-up used for data collection. In Section 3, we perform an analysis of the relations between interaction parameters and user judgments for individual users. Factors impacting the results and implications of our findings for the application of PARADISE-style models are summarized in Section 4. Finally, Section 5 derives from the results some open questions for future research.

2 Experiment

The experiment was conducted with the INSPIRE smart home system [3], which allows control of a TV, an electronic program guide, lamps, an answering machine, blinds, and a fan. 32 users participated in the experiment, of which 16 were "old" (60-80y) and 16 were "young" (20-30y). Each user conducted 2 similar scenarios, one with dynamic help prompts and one with fixed help prompts by entry to a dialog node. Age group, scenario, and help condition were balanced in the experimental design. Each scenario consisted of 7 tasks, covering all devices integrated in the environment. The tasks were of different complexity, involving 2 to 7 concepts to convey to the system. Some tasks were composed of subtasks, e.g. "switch on two lamps and dim them". After each task, the overall impression of the interaction was rated on a continuous scale with 5 labels and extended ends [4]. The interactions were logged. From the log files interaction parameters have been calculated, following the definitions in ITU-T Suppl. 24 to P-Series [5]. By this, we obtained 14 vectors per user, containing a rating and corresponding interaction parameters. For two users, data is missing, so there are 30 users (15 young and 15 old) included in the further analysis.

3 Analysis of Individual User Models

3.1 Correlations for individual users

The aim of our modelling efforts is to establish a relation between measurable dialog characteristics (interaction parameters) and the user judgments. Such a model would allow to estimate the quality perception of the user from logged interactions. The precondition for such a model is a correlation between the parameters and the rating of the dialog. Therefore, we analyzed correlations between judgments and the following interaction parameters:

- Number of dialog turns *(#Turns)*, Task Duration (*td*), average system turn duration (*std*), *#Help-Requests*, *#Repetition-Requests*, *#Barge-in* (user interrupts system prompt), task success (*TS*), Query Density (*qd*), average number of informational concepts per utterance (*#AVPS*), number of incorrectly parsed utterances (*#PA:IC*), Concept Error Rate (*CER*)

While there are many more parameters in our database, we decided to consider only these. All others we could calculate from our data were either redundant with these or too unspecific in their meaning for the prediction of the users' judgments. For example, if the user turn duration (*utd*) is correlated with the judgment, this could be due to the naturalness of the interaction, or the age of the user (which is correlated with *utd*), or the insecurity of the user in replying, or just chance. An exemption is the inclusion of *#Turns* and the *td* in the analysis. We chose to examine both parameters, as their difference is well defined: *#Turns* is closer related to the elegance of the dialog, while *td* reflects the actual time resources spent for the dialog.

Table 1 shows for each test participant which of these parameters are significantly correlated with the user's judgments (p<0.05). It can be seen that there are differences between the participants, however, some parameters show a correlation for many of the users, namely *td* (20), *#Turns* (19), *std* (15), *CER* (14). For *#PA:IC*, there are 6 occurrences despite 12 participants were not even confronted with an incorrectly parsed utterance. This means that for every third participant confronted with an incorrectly parsed utterance, this had an influence on the judgment. All other parameters are seldom or never correlated with the user judgment.

In addition, the *number* of correlated parameters differs strongly between the users. E.g., for user 27, almost all parameters are correlated, while there are 6 users for whom no significant correlations are found.

User	Correlation	Regression	Age	Age group	Digit span	Tech. Affi.
1	*td, std, CER, #Turns*	*CER*, *td*, *#Turns*	27	young	19	1.14
3	*td, std, CER, #Turns*	*#Turns*	63	old	miss.	0.00
4	- none -	- none -	67	old	miss.	0.86
5	*CER*	*CER*	64	old	miss.	0.29
6	- none -	- none -	28	young	14	0.43
7	*td, CER, #Turns*	*td*, *std*	64	old	13	1.00
8	*td, std, CER, #Turns*	*td*, *std*	63	old	13	1.14
9	*td, std, #PA:IC, #Turns, utd*	*#PA:IC*, *#Turns*, #AVP	62	old	9	0.43
10	*td, std*	std	65	old	8	-0.14
11	*td, std, CER, #Turns*	std	72	old	8	0.43
12	*CER, #Turns*	*CER*, *#Turns*	22	young	18	1.00
13	*td, , CER qd, #AVP, #PA:IC, #Turns, utd*	*#Turns*	27	young	22	0.71
14	*std*	std	27	young	16	1.14
15	- none -	- none -	72	old	miss.	0.43
16	*td, std, qd, #Turns, utd*	*#Turns*	25	young	21	1.14
17	*td, CER, #Turns*	*td*	29	young	15	0.86
19	*td, std, #PA:IC, CER, #Turns*	*#PA:IC*, *CER*, #Barge-in	28	young	25	1.43
20	*td, std, CER*	std	75	old	13	1.00
21	- none -	- none -	70	old	14	-0.14
22	*#Turns*	*#Turns*	73	old	14	1.57
23	- none -	- none -	26	young	18	0.86
24	*td, #Barge-in, #PA:IC, CER, , #Turns*	*CER*	27	young	19	1.43
25	*td, std, qd, , #Turns*	*td*	26	young	21	1.29
26	*td, std, #PA:IC, , #Turns*	*td*	28	young	23	1.71
27	*td, std, #help_requests, qd, #AVP, #PA:IC, CER, #Turns, utd*	*#Turns*, #Barge-in	25	young	21	0.86
28	*td, std, qd, #AVP, CER, #Turns, utd*	*#Turns*, *CER*	27	young	20	0.86
29	*td, #Barge-in, qd, #AVP, #Turns*	*#Turns*	29	young	21	0.86
30	- none -	- none -	85	old	miss.	0.43
31	*td, #Turns*	*#Turns*	72	old	15	-0.14
32	*td, std, CER*	*std*	64	old	miss.	0.57

Table 1 - Significantly correlated parameters, predictors in regression analysis, age, age group, digit-span result and technical affinity value for each participant.

To check for the possibility of suppressor effects, we performed a regression analysis for each user with ratings as dependent and interaction parameters as independent variables, using a stepwise inclusion algorithm.

Here, the results are more diverse, which can be attributed to different reasons. Firstly, it could be due to the redundancy between parameters like *#Turns*, *td*, and *CER*. The redundancy is also well reflected by the reduced number of parameters in the regression models in comparison to the ones that are correlated with the judgment. Secondly, in our correlation analysis we did not analyze the relative importance, i.e., the strength of each parameter's correlation with the judgment. In stepwise regression analysis, the most important parameter is selected first, and other parameters are just added if they explain a significant part of the variance in the judgment which is not covered by the parameters selected beforehand. Thus, if despite the same correlated parameters for two users, the relative strength of the correlations differs for each, the parameter(s) included in the individual regressions differ as well.

However, the differences in the regression equations could also be explained by characteristic features of the users. E.g., they might weight the quality aspect differently regarding their importance for the overall quality judgment. Also, users might encounter different specific problems; for example, a concept insertion or substitution (*#PA:IC*) is seldom, but has a strong impact on the dialog. However, it has to be noted that the relationship between variables can be somewhat arbitrary if such few cases (14 tasks) are considered. We cannot be sure if the parameter selected first for the regression model is really the most important one for that user, as the difference to the next-best parameter might be minimal.

As a result of the regression analyses, it seems that the low correlations for some users are not attributable to suppressor effects of other variables: For the same users no significant predictor of their judgements could be found with linear regression either. In fact, overall the accuracy with which ratings of individual users could be modelled by interaction parameters differed considerably between users. R^2_{adj} of the regression models ranges between 0.258 and 1.0 (mean = 0.667, *std.* = 0.21; that some users can be predicted perfectly from the parameters is of course unrealistic for a model aiming at general validity).

Two reasons can be cited for the result that the predictability of judgments differs between the users, while the parameters correlated for each user are similar. Either there are different types of users: the ones who are affected by the aspects measured with these parameters (e.g. efficiency) and the ones who are not. Or, our results include measurement errors due to the specifics of psychometric measurement. The latter reason is specifically interesting for the quality measurement of interactive systems, as usually the experiences users make during the interaction are different. This implies that we cannot calculate mean values of many users judging the same stimulus. The following section discusses the impact of user characteristics and measurement errors on the relation between interaction parameters and judgments.

4 Discussion

4.1 Predictability of judgments and user characteristics

In the following we analyze the influence of different factors on the strength of the correlation between the interaction parameters and the judgments. The focus is first on participant age and technical affinity, which are the factors related to user characteristics. In the next section, we will examine the impact of measurement errors as described in the previous section on the parameter-judgment relation. In each case, we specifically look at the parameters which are most often correlated with the judgments (cf. Table 1), namely *#Turns*, *td*, *std* and *CER*. For each parameter, we calculated a new variable containing the correlation of this parameter with the judgment for each test participant. Such parameters are identifiable by the suffix "_cor".

Age group. Of the four users for which no model could be built, three belonged to the "old" age group. We therefore thought it would be reasonable to have a closer look at effects of age on the predictability of judgments from interaction parameters. We found that the relation between *#Turns* and judgment is significantly correlated with the age of the participants (continuous variable; r=0.38, p=0.04, N=30). No effect was found for other interaction parameters.

Technical affinity. A reasonable hypothesis would be that users with higher technical affinity are more interested in technical capabilities and efficiency of an interface. In our test, technical affinity was measured by averaging the replies to seven questions about attitude towards technology and computers. This variable is correlated with *#Turns_cor* (r=0.361, p=0.05, N=30), which means that the higher the technical affinity of the user, the closer her/his judgment is dependent on *#Turns*. For the other parameters, again no effect was found.

In our database, technical affinity is moderately correlated with age (r=-0.52, p<0.01, N=30), that is, the cause of any effect found cannot be clearly attributed to either factor. We also analyzed the effect of technical affinity on *#Turns_cor* per age group (StDev(tech_aff, young)=0.33; (StDev(tech_aff, old)=0.52). Here, no significant correlations were found, which emphasises the peril of confusing effects of technical affinity and age in our data.

We had the possibility to test the dependency between predictability of judgments and age-group in another study [6], where we found support for such a dependency. In this study, a spoken dialog service for telecommunication tariff information was analyzed. The system differed with respect to the command style and system initiative, and featured an automatic classification of user groups resulting in adoption of the system persona. Here, the dependency between *#Turns* and the judgments was correlated with the age group (adult/senior) with $|r|$=0.62, p<0.01, N=17, where older users showed less correlation between *#Turns* and the judgment on overall impression. Unfortunately, neither technical affinity nor digit-span was tested in this experiment.

4.2 Predictability of judgments and measurement errors

Although questionnaires have several advantages (e.g., they are relatively effortless and easy to apply) there are many sources of unreliability and inaccuracy decreasing the quality of these measurements [7]. For example ratings can be affected by the format of the answer scale, the order of the items or most interesting for this study by individual differences between the users (see e.g.[8][9]). Krosnick describes the cognitive processes respondents carry out when answering a question as follows: (a) Interpreting the question, (b) searching the memory for relevant information, (c) integrating the information into a judgements and (d) translating the judgement to a response. In view of these four steps, especially individual differences regarding the memory as one of the relevant cognitive abilities are likely to have a strong influence on user ratings. Furthermore different response styles are explainable with this four-steps-model: Since each of the processes described above requires much effort, only few people might be motivated to invest this effort to optimize their answers. Influenced by the intrinsic motivation and variables like the length of the questionnaire or the frequency of ratings during an interaction, the required steps are executed in a superficial manner or even skipped. Participants will then apply individual decision heuristics, choose answers like "I don't know", or the middle and the anchor point of a rating scale [8].

Cognitive abilities of users. This was tested in the experiment with the digit-span test taken from the German adoption of the Wechsler Intelligence test [10], in which the participant listens to a sequence of numbers, which s/he has to repeat forward or backward. The length of the sequence increases, and the score for the participant is calculated according to the length of the sequence s/he still could repeat. Of course, this tests just a small part of the cognitive abilities, however, we can assume that the result is related specifically to the working memory capabilities of the participant.

In the database, the digit-span result is correlated with technical affinity (r=0.57, p<0.01, N=24) as well as age (r=-0.773, p<0.01, N=24). Accordingly, we found an almost significant correlation between digit-span and *#Turns_cor* (r=0.36, p=0.081, N=24). Like in the analysis of technical affinity, we separated the young from the old users and calculated the correlation for both groups. For young users, we found correlations with digit-span for *#Turns_cor*, *td_cor* and *std_cor* (all p<0.05), while for old users, we did not observe a significant correlation.

This could be interpreted in the way that the effect saturates at a point not reached by the older participants. In any case the presence of many correlations in the group of young users indicates that age is not the main contributor to the influence of digit-span on the predictability of the judgments. This assumption is in line with the possible causes of error described above: It is a plausible hypothesis that unpredictable users are due to memory deficiencies not able to give valid and reliable ratings.

Standardization of judgments. We then compared correlations of the interaction parameters with the standardized judgments (stand_task_rate) and the unprocessed judgments (task_rate). By standardized judgments, we mean that for each participant, we calculated the mean and standard deviation (StDev) of his/her ratings and then calculated for each single rating

$$stand\_task\_rate = \frac{task\_rate - mean}{StDev}$$

Table 2 shows the correlations between the most frequently correlated interaction parameters and user judgments, allowing to compare the predictability of raw scores with the scores normalized to uniform mean and standard deviation. The overall strength of the relationship does not seem to differ considerably. For *CER*, the raw scores show a higher correlation than the standardized judgments, while for *std* it is the other way around. For the relation with the length of the interaction, standardization of judgments does not seem to matter at all.

		Concept Error Rate	**# Turns for Task**	**Task Duration**	**System Turn Duration**
task_rate	*r*	-0.506(**)	-0.536(**)	-0.557(**)	-0.399(**)
	p	0.000	0.000	0.000	0.000
	N	410	410	410	408
stand_task_rate	*r*	-,459(**)	-,547(**)	-,557(**)	-,445(**)
	p	0.000	0.000	0.000	0.000
	N	410	410	410	408

Table 2 - Correlations between interaction parameters and user judgments, once raw scores (task_rate) and once scores normalized to uniform mean and STD (stand_task_rate).

4.3 Application to PARADISE models

We proceeded by calculating prediction models for subgroups of users in our database. To build groups from technical-affinity and digit-span, these variables were dichotomized at the median leading to two equally-sized groups. Models were calculated to predict the raw judgments as well as the standardized judgments. We analyzed the relation between the parameters and judgments with R^2 on training data (ALL) and the predictive power of models to unseen cases with mean R^2 in cross validation (L1O). We also examined if the predictive reliability improves if only the predictors are used which we saw most often in our user-wise correlation analysis. Table 3 shows the R^2's of all models and the parameters included in the ALL model.

We intend her to mainly point at the differences between the groups. Roughly, we can say that the younger a user, the higher her/his technical affinity and digit span result. Looking at any of these factors, we see that young (or higher affinity, span>16) users' judgments can be predicted with a higher accuracy (R^2). The clearest effect is found for digit-span, followed by technical affinity.

Then we compared prediction models for task_rate and stand_task_rate, using the full data set. R^2's do not differ remarkably here, however, when we calculate models predicting standardized ratings of sub-groups, we observe a clear increase in R^2's for all groups except the one with lower digit-span. The result generalizes to the L1O procedure, either with all variables or with just the most promising ones from the correlation analysis.

Finally, a good choice of parameters seems to be profitable for the L1O procedure. For some models, the R^2 slightly increases when only the most promising parameters are permitted for inclusion in the model. Only in one case, the R^2 decreases (stand_task_rate predicted for users with higher technical affinity).

Configuration	ALL	L1O all parameters	L1O good predictors	Parameters
		Task_rate		
old	0.37	0.29	0.29	*td, CER, std*
young	0.40	0.30	0.32	*td, CER*
lower afinityf	*0.30*	*0.29*	*0.29*	*#Turns, std*
higher affinity	*0.37*	*0.33*	*0.35*	*td, CER*
digit-span<=16	0.34	0.20	0.20	*td, CER, std, #Barge-in*
digit-span>16	0.47	0.30	0.36	*#Turns, #PA:IC, CER*
		Stand_task_rate		
lower affinity	*0.39*	*0.31*	*0.32*	*Td, CER, #PA:IC*
higher affinity	*0.42*	*0.38*	*0.36*	*#Turns, #PA:IC, CER, std*
digit-span<=16	0.31	0.27	0.27	Td
digit-span>16	0.53	0.45	0.47	#Turns, CER, std
		All users		
task_rate	0.38	0.34	0.34	*Td, CER*
stand_task_rate	0.37	0.35	0.35	*Td, CER, std*

Table 3 - R^2's for models calculated for different sub-groups of the database and with different methods. ALL shows results on training data, while L1O shows results on user-wise cross-validation. The "good predictors" are *#Turns, td, std, CER.* The Parameters column shows the parameters included in the ALL model for each row.

5 Conclusion

In this paper, we analyzed how models for the prediction of user judgments on interactions with spoken dialog services can be improved. We started from the observation that average judgments, e.g. for a system configuration, can be predicted better than judgments of individual users. We therefore examined the relations between interaction parameters and judgments for single users and found that firstly, the parameters correlated with the judgments are relatively consistent across users, and secondly, for some users prediction works better than for others.

By analyzing the relation between the predictability of a user's judgments and characteristics of the user, we could show that individual preferences as well as scale usage play a role for the predictability. All factors (digit-span, technical affinity, age, scale standardization) show an effect on the prediction quality, however, effects are strongest for digit-span and technical affinity. This result could be confirmed with user-wise cross validation.

We also showed that standardization of judgments to cope with individual differences of scale usage improves the result for the well predictable user groups. This is mostly not possible in practise, but it shows that some part of the variance in scale usage is not due to different experiences but to different usage of the scale.

The findings pose a number of questions for the task of predicting judgments as well as the measurement of the true user ratings. Among these: To what degree are the difficulties in predicting user judgments due to individual judgment behaviour which we could consider as an experimental artefact? If users with lower memory capabilities do not judge as consistently as those with higher capabilities, is it a valid conclusion that we should measure with the latter type of users and assume that the results are valid also for the former users? Could we find a model which is constructed in accordance with knowledge about the target users, which predicts the construct we target with our questions better than the actual questionnaire?

With our ongoing work on such models, we hope to make fruitful contributions also to these kind of questions, finally not only trying to substitute evaluation methods by faster ("discount" [11]) procedures, but also improving the method itself.

References

[1] Walker, M., Litman, D., Kamm, C., and Abella, A. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, Madrid, 271–280.

[2] Engelbrecht, K.-P. & Möller, S. (2007). Pragmatic Usage of Linear Regression Models for the Prediction of User Judgments. *Proc. of SIGdial Workshop*, Antwerp, 291-294.

[3] Möller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, A., Schuchardt, D., Fakotakis, N., Ganchev, T., Potamitis, I. (2004). INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control. *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004)*, Lisbon, Vol. 5, 1603-1606.

[4] Möller, S. (2005). *Quality of Telephone-based Spoken Dialog Systems*, Springer, New York.

[5] ITU-T Suppl. 24 to P-Series Rec. (2005). Parameters Describing the Interaction with Spoken Dialogue Systems, International Telecommunication Union, Geneva.

[6] Möller, S., Engelbrecht, K.-P., Pucher, M., Fröhlich, P., Huo, L., Heute, U., Oberle, F. (2007). TIDE: A Testbed for Interactive Spoken Dialogue System Evaluation, in: *Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007)*, Moskow.

[7] Annett, J. (2002). Subjective Rating Scales: Science or Art? *Ergonomics*, 45,. 966-987.

[8] Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.

[9] Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., and Clark, F. (1991). Rating Scales: Numeric values may change the meaning of the task. *Public Opinion Quarterly*, 55, 570-582.

[10] Von Aster, M., Neubauer, A., and Horn, R. (Ed.) (2006). *WIE. Wechsler Intelligenztest für Erwachsene*. Übersetzung und Adaption.des WAS-III von David Wechsler, The Psychological Corporation, USA.

[11] Nielsen, J. (1993). *Usability Engineering*, Morgan Kaufmann, Amsterdam.