A Closer Look at Quality Judgments of Spoken Dialog Systems

Klaus-Peter Engelbrecht, Felix Hartard, Florian Gödde, Sebastian Möller

Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany

{Klaus-Peter.Engelbrecht,Florian.Goedde,Sebastian.Moeller}@telekom.de,
Felix.Hartard@Berlin.de

Abstract

User judgments of Spoken Dialog Systems provide evaluators of such systems with a valid measure of their overall quality. Models for the automatic prediction of user judgments have been built, following the introduction of PARADISE [1]. Main applications are the comparison of systems, the analysis of parameters affecting quality, and the adoption of dialog management strategies. However, a common model which applies to different systems and users has not been found so far. With the aim of getting a closer insight into the quality-relevant characteristics of spoken interactions, an experiment was conducted where 25 users judged the same 5 dialogs. User judgments were collected after each dialog turn. The paper presents an analysis of the obtained results and some conclusions for future work.

Index Terms: spoken dialog systems, PARADISE, evaluation

1. Introduction

The evaluation of Spoken Dialog Systems (SDSs) is a complex issue, involving the assessment of many different components and their interrelations. Therefore there is a desire for simple metrics comprising all aspects of the system's quality in one contrastable measure. Such metrics could also be used for dialog strategy adoption. In 1997, Walker et al. [1] introduced the PARADISE framework, which follows the basic assumption that user satisfaction is an adequate measure of the system's overall quality. This view is supported for example by Hone and Graham [2] and the definition of "Quality" in Jekosch [3], stating that it is the result of appraisal of the perceived composition of a service with respect to its desired composition.

According to Walker et al., user satisfaction is achieved by maximizing task success and minimizing costs in terms of efficiency and dialog quality. This assumption breaks down user satisfaction into interaction parameters which are measurable independently of the observer. By training a linear regression equation with the parameters as an input and satisfaction as the target variable, a prediction model can be built. However, several studies have shown that correlations between interaction parameters and user judgments are surprisingly low, e.g. [4]. Therefore, prediction models build according to PARADISE usually predict about 50% of user satisfaction only [1, 5]. Moreover, experience shows that models derived from dialogs with different systems usually include differing parameters and coefficients. On the other hand, it has been shown that mean values, e.g. for different experimental conditions, can be predicted quite accurately with a model trained on all configurations together [6].

Two reasons can be cited for these findings. Firstly, being able to predict mean values but not individual cases might be due to differences between the users' rating behaviors. This problem has been discussed e.g. in [7]. Indeed, by taking several judgments from each user, it was shown that the parameters which are correlated with the quality judgments were largely shared among the users, however, the strength of the relationship differed between the users. When dividing users according to their technical affinity and cognitive skills, for users with higher technical affinity and higher cognitive skills prediction models performed better [8].

Secondly, the system characteristics might influence the parameters relevant for the judgment. For example, if the speech recognition performs poorly, the number of no-matches might be strongly correlated with the judgment. However, if speech recognition is not a problem, user judgments might more reflect the system voice or something else [9]. Similar assumptions have been made by Nielsen, who recommends iterative testing, because new problems might occur once the most severe design problems have been removed [10].

While the cited literature supports our assumptions that judgment prediction models depend on the system features and user characteristics, the exact nature of this dependence is so far unknown. We therefore designed an experiment in which user judgments could be tracked meticulously and compared among the users. To do this, we confronted our users with as similar dialogs as possible, using the Wizard-of-Oz (WoZ) method. Interactions followed predefined scripts, which included problematic situations at particular dialog turns. In addition, we asked the users for a judgment after each turn. In the next section, we describe how the experiment was designed and conducted. In section 3, we present findings we could derive from the data. These are discussed in section 4, before conclusions for future work are drawn in section 5.

2. Experiment

In this section, we describe the experiment by first introducing the experimental design and the system, and then explaining how the experiment was conducted.

2.1. Collection of quality issues

A set of tasks had to be defined covering as many quality-related issues as possible. Furthermore, we were interested in user judgments when confronted with combinations of problems. Also, we wanted to compare judgments of different users, i.e. all users would be confronted with the same dialogs.

As a first step, we collected as many interaction problems as possible in a brainstorming session. Problems were partly observed in our former experiments, or known from the SDS design literature, as summarized in [11]. The resulting list of issues was amended later whenever a new issue came to our mind until the design of the dialogs was completed. Afterwards, we evaluated if each problem could be forced to appear in a Wizard-of-Oz-driven interaction, and how this could be done. Unfortunately, for some very interesting issues no solution could be found. E.g., it is impossible to force a user to barge into system prompts. Even a very long prompt would not guarantee barge-in for all users, which however is a precondition for the comparability of the dialogs and ratings.

After collecting the problems and their possible forced realization in a dialog, concrete dialog scripts had to be designed which all users would have to complete. We tried to arrange the issues in a way allowing analysis of the effects of each issue alone and in combination with other issues.

2.2. Selection of the system

In order to keep up a plausible interaction scenario, we decided to design a consistent dialog strategy (i.e. system) for all tasks. In former evaluations, we had worked with research prototypes featuring mixed-initiative dialog strategy, natural language understanding and template-based prompt generation. The complexity of these systems sometimes led to dialog situations which were difficult to anticipate even for the system designer. Furthermore, some of the problems arising from this complexity were very specific to the system. Therefore, an adequate and generic parametric description of the resulting dialogs would be difficult to find, while it would currently apply to only few systems.

We therefore decided to design a new version of our prototype BoRIS (Bochumer Restaurant Informations-System, [11]), resembling in its dialog strategy and competence a typical commercial dialog system. BoRIS helps users to find a restaurant in the German city Bochum via the telephone. For the new version, we built a WoZ interface, allowing the wizard to simply replay the system prompts foreseen at each dialog turn. Remember that dialogs were predefined, including all interaction problems. The wizard also had at hand some prompts for "emergency cases", e.g. when the user did not say anything or said something for which the predefined system reply would be implausible. This way, the illusion of a real system could be held up more easily.

The system mock-up realized a simple system-initiative dialog, however, flexible enough to generate different types of problems. The issues which could be imposed directly on the dialogs include understanding errors of different severity (partial understanding, no-match, concept-substitution). These were combined with different types of confirmation. Also, prompt wording problems were "generated", e.g. where the reply options were not clear (*S: I could not find a matching restaurant. You can change day, time, location, cuisine or pricing.*), or where the vocabulary was not shared with the user (*U: Sushi – S: OK, Japanese*). As an issue of controllability, the system was designed to hang up after providing the information, without the option to repeat the information. On the efficiency level, the amount of concepts which can be specified by the user in one turn was varied, and some prompts were deliberately wordy. Finally, task failure could be ensured by not asking for a criterion (the "reason" being that just one restaurant was left in the search space).

To complete a dialog, a user has to provide the desired day and time of the meal, plus location, cuisine and pricing of the restaurant in this order. Day and time can be uttered at once, but have to be explicitly confirmed by the user. If the user disconfirms, the system passes to asking day and time separately, using implicit confirmation. All following information is provided one-by-one by the user, and implicitly confirmed by the system. However, the last information (pricing) is not confirmed, and instead information is provided directly. If no restaurant could be found, the system allows the user to change constraints until a solution could be found. A typical dialog could look like this:

S: Welcome to the telephone-based restaurant information system for Bochum. When would you like to eat?
U: Tomorrow night.
S: So you want to eat tonight?
U: No.
S: On which day would you like to eat?
....
U: Medium pricing
S: Unfortunately, I could not find a restaurant matching your query. Would you like to change time, location, cuisine or pricing?
U: Yes, cuisine.
S: Which type of cuisine are you searching for?
U: Italian.
S: The possible restaurants are: {name, address}

Finally, we came up with five different dialogs containing all the situations we were able to generate in a controlled way.

2.3. Conducting the experiment

25 users recruited on the campus, but covering various demographic groups, participated in the experiment. Each user performed all five tasks, however, in differing order. After each turn, the user had to rate the quality of the dialog up to the current moment on a keyboard (number pad). To improve the scale characteristics, we added a graphical measurement scale with labels from "poor" to "excellent" on top of the respective keys with lines pointing from each label to the corresponding key to press (Fig. 1). We chose to let users rate the quality *up to the current moment in the dialog* as we were interested in the development of the users' opinion along the dialog. The question was written on top of the rating scale, and users confirmed verbally that they had understood it correctly.

Tasks were described roughly as mind settings [11], in order to set the users in a believable situational context. As we were interested in judgments for the same situation rather than varying behavior, the concepts to convey in each turn were summarized below the scenario. This also helped users to not get lost in the dialog when they had a high cognitive load on judging the actual turn. In addition, each participant performed one training dialog, with the experimenter standing besides and notifying her when she forgot the rating.

Users communicated with the system through a headset of high quality. This allowed them to keep their hands free for the number pad, which we deemed more important than the realness of the experience. The WoZ just replayed the prompts foreseen at each turn, however, not before the user had rated the previous turn. This sometimes caused a delayed response before the user noticed that she forgot to give a rating. Users were allowed to rate either before or after their own utterance.

After each dialog, the users provided a final judgment on a paper scale and stated whether they thought the task was successful. After the experiment, individual characteristics of the interactions were judged on a 43-item questionnaire designed according to [12] and covering different aspects of the system. We also collected information about the users' attitude towards SDSs and their general technical affinity. The latter had shown to impact judgments of whole dialogs in [8].

After the experiment, dialogs were annotated with commonly used labels describing the interaction in terms of understanding errors, system confirmation, system and user speech acts, prompt length in number of words *(#Words*), contextual appropriateness of prompts (annotated according to Grice's maxims, see [11]), dialog length as the current turn number (*#Turn*), and task success.

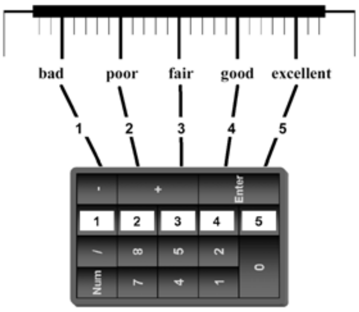

Figure 1. *Rating scale and keypad used during dialogs*

3. Results

All data were aggregated in an SPSS sheet, comprising 1027 turns with valid ratings. The distribution of judgments is: 43 "bad", 143 "poor", 229 "fair", 389 "good", 223 "excellent". We first look at the univariate relations between per-turn judgments and parameters annotated for this particular turn. Then, we present models for multivariate relationships, using the general linear model approach.

For univariate relations, Pearson correlations (r) were calculated where possible. The impact of parameters on the nominal level was tested with ANOVA, and significance was confirmed with a Kruskal-Wallis-Test because of the inhomogeneity of variances (according to Levene's test). We report F (pointing to the significance of the relation), and the effect size η^2 (which is equal to R^2 when mean values are predicted for each group). For correlations, the squared correlation (r^2) is added for better comparability with η^2.

Generally, single parameters explain only a small part of the variance in the ratings. The highest relation was found with task success ($F(1)$=18.2**; η^2=0.148), followed by user speech act ($F(3)$=41.1**; η^2=0.119) and understanding errors ($F(3)$=38.4**; η^2=0.101). For all other parameters, as well as the user feature SDS attitude, we found highly significant relations, which however covered an even smaller part of the variance in the ratings i.e. r^2 or η^2 was lower.

The relations become stronger when ratings are normalized for each user, so that they feature the same mean and standard deviation. The impact of task success raises to η^2=0.241 ($F(1)$=33.4**). For all other parameters describing the dialog, the increase in effect size gained by the normalization is smaller than 0.025. Analysis of the impact of user characteristics on the standardized ratings is not meaningful, as normalization practically means equalization of the differences between the users.

We also tested the relations of parameters to the change in ratings since the previous turn (-1 if rating decreased, 0 if no change, +1 if rating increased). Here, we observed a clearly higher impact of the errors, while all other parameters except *#Turn* showed a lower impact. For example, task success had no significant influence on the relative judgment. Differences between user groups were not assessed for relative judgments, as no plausible hypothesis could be formulated for such differences.

Finally, we checked whether the absolute judgments are impacted by the dialog history, by analyzing their relations to errors and judgments in the previous turns. The relation between current ratings and errors in the previous turn is characterized by $F(3)$=30.7**; η^2=0.083. Also, the current ratings are correlated with previous ratings (r^2=0.295 with rating_lag1; r^2=0.078 with rating_lag2), which means that they change only gradually over time. This also confirms the validity of our measurements in that ratings summarize the quality perception of the previous dialog steps.

While single dialog parameters, dialog history parameters or user characteristics explain only a small part of the variance in the ratings, the parameters in their combination cover a good part of the spread. This is partly due to interactions between the parameters, where the true impact of a parameter on the judgment is discovered only when the data are split by another parameter's values. In other cases, the parameters appear to be complementary.

We expected to find interactions of understanding errors with the other parameters, either because they might determine the severity of the error (e.g. when they occur with certain speech acts or confirmation strategies), or because they might hide the impact of less severe issues (e.g. lengthy or badly formed prompts). However, we found only a trend for an interaction with prompt length ($F(1)$=3.6; p=0.058).

On the other hand, analyzing the impact on ratings by their context in the dialog history, we could show that understanding errors affect the ratings differently depending on the understanding performance in the previous turn ($F(5)$=2.96*; η^2_{Model} =0.181**). Also, we found an interaction between the two previous ratings ($F(15)$=1.9*; η^2_{Model} =0.355**).

With respect to user characteristics, technical affinity determines how a user judges situations characterized by different speech acts ($F(3)$=4.19**; η^2_{Model} =0.135**) and errors ($F(3)$=5.6**; η^2_{Model} =0.119**). In particular, users with higher affinity judged confirmations and repetitions better Also, they did not punish understanding errors as much as the low-affinity users. An analysis of the impact of SDS attitude on judgments in case of errors revealed that users with higher attitude rated no-matches relatively well, besides their better judgments overall ($F(3)$=2.88*; η^2_{Model} =0.174**).

Models can be enriched with more parameters, leading to a reasonable level of prediction accuracy. For example, a model comprising the previous two ratings, error, current user speech act, and the user characteristics SDS attitude and technical affinity, reaches η^2=0.759**. However, still 24% of the variance in the ratings are not explained by the model, despite the inclusion of user characteristics and temporal context.

4. Discussion

In the previous section, we showed that user judgments depend on various parameters and their interrelations. Combinations of parameters describing the context in terms of time, user characteristics, and co-occurrence of events could predict the biggest part of the variance in the rating, but not all of it.

We therefore analyzed the ratings themselves. Fig. 2 shows ratings of 25 users for one task, i.e. all users depicted were confronted with exactly the same dialog, apart from minor differences in the user utterances. The impact of understanding errors ("FA", "PA") on the judgments can clearly be seen. However, the spread of the ratings is relatively high, especially when it comes to errors. This could not be explained completely by the user characteristics we measured.

Users even started off with different ratings (between 3 and 5) in the first turn. T-tests analyzing the effect of the user groups were insignificant. In addition, the users differed in the range their ratings occupied.

Moreover, analysis of relative judgments revealed that some users improved their judgments when an error occurred, while most users decreased their judgment in such cases. A post-experimental interview with the former users revealed that they liked the system reaction, which provided help and a rephrased question. The other users seem to have focussed more on the error.

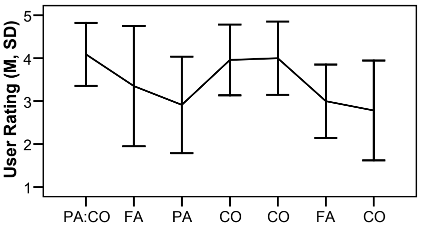

Figure 2. *Ratings by 25 users (M, SD) for one of the dialogs, x-axis labeled with errors in each turn (CO= parsing correct, PA=partially correct, FA= failed).*

These findings imply that our assumptions about the impact of user characteristics on the ratings need to be refined. Firstly, the user's reaction seems to depend on more complex characteristics than a simple combination of well-examined traits known from usability testing (such as attitude towards SDS or technical affinity), despite the undoubted impact of these parameters on the ratings. Secondly, and more importantly, the differences between users in how they judge certain events are not only quantitative, but qualitative. This concerns the prioritization of usability issues (e.g. judging the error or the recovery strategy) as well as their lenience when problems occur.

For system development and assessment, this implies that the optimal dialog strategy can differ for different users or groups of users. For the more theoretical issue of predicting the users' judgments, the results show that user characteristics are an important factor, which determines not only how well judgments can be predicted from dialog events (cf. [8]), but also how a specific user reacts to these events. Therefore, different models are necessary for different user groups.

Also, as our experiment showed that users do rate the same dialog differently even within one user group, predictions of user judgments should not aim at one correct value, but take into account and target the spread of the judgments of all potential users.

A further implication for judgment prediction models concerns the integration of time. Interaction parameters, which describe the number of occurrences of a problem, but not their relation in time, are not sufficient for an accurate prediction of user judgments. Such relations concern the simultaneity of different problems, as well as the succession of events. As a "special case", users are relatively forgiving and increase their judgments after having passed a problematic situation (cf. Fig. 2). This is a new point of view which has not been taken into account in judgment prediction models, nor in the creation of interaction parameters for the assessment of SDS components, yet.

Criticism of our results may concern their validity. The problems which occurred in the dialogs were deliberately designed, which might reduce their variability and strengthen their relation to quality judgments. Also, the situation was rather artificial, and the two tasks of talking to the system and judging it at the same time might have interfered. Finally, some quality issues, such as barge-in, could not be measured with the experimental setup.

Also, it is not absolutely clear which question the participants of such an experiment should be asked. While evaluators are often interested in absolute ratings, relative ratings might be easier to state for the participants. Changes in the rating also showed a stronger relation to understanding errors. A clearer picture of the impact of simultaneous events might be attained by asking for the quality of the current turn instead of the quality "so far".

5. Conclusion

In this paper, we reported on a new type of experiment, which allowed us to analyze user judgments about interactions with SDSs in a more fine-grained fashion than usual user tests would do. By this we were able to analyse judgments in more depth, leading to new insights about how users rate dialogs.

Despite some drawbacks of the method, we could generate very rare data describing relations between judgments and situational parameters in great detail. Gaining insight into general aspects of the judgment process, such as dependence on time or user characteristics, also enables a meaningful analysis of relations between other parameters and judgments collected in a more usual test setup.

In the next steps of our work, we will take the findings of this study into account in modelling quality perception of SDSs. We plan to use classifier models which consider the classified object as a time series, in particular Hidden Markov Models. Furthermore, we will examine the possibility of predicting the spread of judgments for a number of users, in order to take into account unexplained individual differences. Such a model could then be used for decision taking in adaptive SDSs or for the analysis of dialog corpora. E.g., all situations in which the user judgment is likely to be low could be filtered out of the data and presented to the designer for inspection.

6. References

[1] M. Walker, D. Litman, C. Kamm, A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in *Proc. of ACL/EACL*, 1997, pp. 271–280.

[2] K. S. Hone, R. Graham, "Subjective Assessment of Speech-system Interface Usability," in *Proc. of EUROSPEECH*, 2001, 2083-2086.

[3] U. Jekosch, *Voice and Speech Quality Perception. Assessment and Evaluation*, Berlin: Springer, 2005.

[4] E. Frøkjær, M. Hertzum, K. Hornbæk, "Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?" Preprint version, in *Proceedings of ACM CHI*, 2000, pp. 345-352.

[5] S. Möller, K.-P. Engelbrecht, R. Schleicher, "Predicting the Quality and Usability of Spoken Dialogue Services," *Speech Communication*, vol. 50, pp. 730-744, 2008.

[6] K.-P. Engelbrecht, S. Möller, "Pragmatic Usage of Linear Regression Models for the Prediction of User Judgments," in *Proc. 8th SIGdial*, 2007, pp. 291-294.

[7] M. A. Okun, R. M. Weir, "Toward a Judgment Model for College Satisfaction," *Educational Psychological Review*, vol. 2, no. 1, pp. 59-76, March 1990.

[8] K.-P. Engelbrecht, S. Möller, R. Schleicher, I. Wechsung, "Analysis of PARADISE Models for Individual Users of a Spoken Dialog System," in *Proc. of ESSV*, 2008, pp. 86-93.

[9] K.-P. Engelbrecht, C. Kuehnel, S. Möller, "Weighting the Coefficients in PARADISE Models to Increase Their Generalizability," in *Proc. of PIT*, 2008, pp. 289-292.

[10] J. Nielsen, *Usability Engineering*, Amsterdam: Morgan Kaufmann, 1993.

[11] S. Möller, *Quality of Telephone-based Spoken Dialog Systems*. New York: Springer, 2005.

[12] ITU-T Rec. P.851, *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*, International Telecommunication Union, Geneva, 2003.