Evaluating Telephone-Based Interactive Systems

Sebastian Möller

Deutsche Telekom Laboratories
Technische Universität Berlin, Germany
sebastian.moeller@telekom.de

Abstract

In order to evaluate the quality of telephone-based interactive systems, two approaches are commonly followed. Firstly, system and user behavior are logged, transcribed and annotated, in order to quantify the performance of the system components and the flow of the interaction between user and system in a parametric way. Secondly, the entire system is evaluated from a user's point of view, with the help of questionnaires and quantitative rating scales. For both approaches, recommendations have been issued, defining interaction parameters and the practical set-up of experiments with human test subjects. In addition, prediction algorithms have been proposed to map interaction parameters to subjective user judgments, thus providing quality estimations without relying on user judgments. The present contribution describes what has been reached for each of the approaches, but also the limitations of each methodology. On the basis of experimental data collected with two exemplary systems, shortcomings are identified and future research directions are outlined.

1. Introduction

The spoken interaction between a human user and an application based on speech technology (e.g. for railway information or telephone banking) gains ground in modern telephone networks. The underlying applications contain speech recognition and understanding components, a dialogue manager, a response generation component, and a component for output of pre-recorded or synthesized speech. The interaction between these components is rather complex, and it is difficult to decide how well individual components have to perform to guarantee an acceptable overall quality for the user.

The quality of the interaction with a telephone-based speech service can be addressed from two separate points-of-view. System developers, on the one hand, are interested in quantifying the contribution of their system or system module to the entire interaction between user and system; they focus on the *performance* of the system (modules), i.e. on whether the system is able to provide the function it has been designed for. For example, the performance of a speech recognizer may be described by quantifying its ability to transcribe the speech signal into its corresponding textual form, e.g. in terms of word or sentence accuracies.

On the other hand, the *quality* of the interaction with a telephone-based speech service largely depends on the perceptions of the user. In this context, quality has been defined as the "result of appraisal of the perceived composition of the service with respect to its desired composition" [1]. This definition, which is based on earlier work by Jekosch [2], shows that perceived quality results from a comparison between what the user expects or desires, and the characteristics he/she perceives while using the service. Quality is thus highly dependent on the situation in which the perception and judgment take place. This fact has to be taken into account when measuring quality in subjective quality evaluation experiments, namely by creating a more-or-less natural test situation and a realistic test user motivation.

Both points-of-view on quality – the one of the system designer and the one of the user – have to be considered in the design and set-up phase of interactive telephone services. System designers need to quantify the performance of system modules and the interaction between user and system, e.g. by calculating so-called *interaction parameters*. These parameters may be measured instrumentally (e.g. the duration of a dialogue, the number of system and user turns, etc.), or they rely on a transcription and annotation by a human expert. In both cases, they do not take the perception and judgment process of the user into account. As a consequence, interaction parameters are no direct indicators of quality, nor can they be considered to be "objective", in the sense that they could be obtained independently of the human user. In fact, they are "subjective" in a double sense, namely that their definition and partly also their measurement depends on the human evaluation expert, and their value strongly depends on the behavior of the human test subject.

When the focus is on *quality*, perceiving and judging human subjects are indispensable in the measurement process. A relatively broad view on quality and its composing dimensions may be obtained from questionnaires which are given to test subjects after interacting with the service under consideration. The interactions are often incited by providing exemplary tasks to the test subjects. In this way, direct measurements of quality may be obtained; the measurement results may however be biased, because the test situation will influence the user's motivation and expectation – thus the reference the perceived characteristics of the service are compared to.

In case that the evaluation takes place under controlled laboratory conditions, both interaction parameters and (test) user judgments can be collected in parallel. From this data, a direct relationship between system characteristics (quantified in the interaction parameters) and user perceptions (quantified in the user judgments) can be established. The relationship may be expressed as a simple correlation, or as a more complex model which aims at predicting quality judgments on the basis of interaction parameters. With the help of such a *prediction model*, quality may be estimated independently of user judgments; the estimation is however not independent of the user, because the interaction parameters have to be calculated from his/her interactions with the system.

It is the aim of this overview paper to present current methods for the evaluation of telephone-based spoken dialogue services, following each of the described ways – interaction parameters, user judgments and prediction models (Section 2). The validity and reliability of these methods will be investigated with the help of experimental data collected with two exemplary systems, see Section 3. The analysis presented in Section 4 reveals that user judgments and interaction parameters provide complementary information to the evaluator. As a result, it is difficult to predict user judgments on the basis of interaction parameters. Reasons for the observed limitations are discussed, and research necessary to overcome the limitations is outlined in Section 5.

2. Assessment and evaluation methods

The following paragraphs provide an overview of the three approaches for assessing and evaluating telephone-based spoken dialogue systems. Two of the approaches seem to be well consolidated so that recommendations on their use have been issued by expert groups or standardization bodies, namely the Expert Advisory Group on Language Engineering Standards, EAGLES [3], and the Telecommunication Standardization Sector of the International Telecommunication Union, ITU-T [1][9]. Still, the validity and the reliability of the information obtained with each approach is a topic for further study. Initial results in this respect are presented in Section 4 and in a separate paper [4].

2.1. Interaction parameters

Interaction parameters may be extracted from log files of real or test user interactions with the system under consideration. In case that the fully integrated system is not yet available, it is possible to collect parameters in a so-called "Wizard-of-Oz" paradigm, where a human experimenter replaces missing parts of the system under test. Parameters which relate to the surface form of the utterances exchanged between user and system, like the duration of the interaction or the number of turns, can usually be extracted instrumentally from the signal. Transcription and annotation by a human expert is necessary when not only the surface form is addressed, but also the contents and meaning of system or user utterances.

Based on a literature survey, parameters were identified which have been used in different assessment and evaluation experiments during the past 15 years [5]. These parameters address different levels of the interaction between user and system. They can broadly be classified as follows:

- *Dialogue- and communication-related parameters*: These parameters provide a rough quantitative description of the flow of the interaction between user and system, as well as of the flow of information. Typical parameters address the duration (dialogue duration, system/user turn duration, system/user response delay), the number of words or utterances exchanged between user and system, the number of system/user questions, and parameters describing how quickly new information can be absorbed by the system (query density, concept efficiency).
- *Meta-communication-related parameters*: They provide a quantitative description of what "happens" in the dialogue. Most parameters count the number of "dialogue problems", e.g. help requests, time-out prompts, rejections from the speech recognizer, error messages from the system, barge-ins from the user, or cancel attempts. Other parameters quantify the recovery from these problems, namely the system/user correction rate, and the implicit recovery parameter.
- *Cooperativity-related parameter*: One parameter is directly related to the cooperativity of system utterances. This so-called "contextual appropriateness" parameter counts whether a system utterance "fits" into the immediate dialogue context, defined by whether it violates one or more of Grice's maxims of communicative behavior, see [3], [6] and [7].
- *Task-related parameters*: These parameters require that a defined task can be solved with the system under consideration, and that this task can be made explicit by the experimenter or by the user. In that case, the task success may be quantified in a binary way (success or failure), potentially amended by the expected reason for the failure. Alternatively, if the task can be described with the help of attribute-value pairs (so-called "slots"), the correctness of the solution reached by the end of the dialogue may be calculated by the κ coefficient, as it is described in [8].
- Speech-input-related parameters: They are related to the performance of the speech recognizer (word/sentence accuracy or error rate, number of errors per sentence, etc.) or to the one of the speech understanding module. The latter may be quantified in terms of correct attribute-value pairs extracted from each user utterance (concept accuracy), or indirectly by the correctness of system answers – an approach used in the DARPA ATIS program.

Exact definitions of each parameter can be found in [5]. By the time of writing, the parameters are proposed for a new Supplement to ITU-T P-Series Recommendations [9]. A common definition of these parameters is highly desirable, as it will facilitate cross-system comparison and provide common data for setting up prediction models.

2.2. User judgments

Following the definition of quality, its measurement has to rely on judgments from real or test users. In order to collect such judgments in a common and quantifiable form, questionnaires have been developed. Questionnaires which are distributed to test subjects before interacting with the system help to obtain unbiased information on the users' background and expectations. Other questionnaires are distributed after some interaction experience, to reflect the current impression of using the system.

A questionnaire which can be distributed to test participants directly after an interaction with the system was developed by Hone and Graham [10]. This so-called "SASSI" questionnaire (Subjective Assessment of Speech System Interfaces) has been designed on the basis of subjective experiments with eight different systems, all showing speech input capability, and some also speech output capability. The questionnaire contains 44 declarative statements of opinion (e.g. "the system is easy to use") with which respondents rate their agreement on 7-point Likert scales. A factor analysis revealed six underlying perceptive dimensions ("system response accuracy", "likeability", "cognitive demand", "annoyance", "habitability" and "speed") to which the statements can be attributed.

A different – but partly overlapping – list of questions is given in ITU-T Rec. P.851 [1]. This recommendation mainly addresses telephone-based spoken dialogue services, but has been successfully used for other spoken dialogue systems as well [4]. It distinguishes between three types of questionnaires: (1) Questionnaires collecting information on the user's background, and distributed at the beginning of an evaluation experiment; (2) questionnaires with questions related to individual interactions with the system under test; and (3) questionnaires related to the user's overall impression of the system, to be answered at the end of an experiment. For each type of questionnaire, an open list of topics is proposed; the topics then have to be translated into precise questions or statements according to the purpose of the evaluation and the system/service under test. Exemplary questions and statements are provided which are rated on 5-point Likert scales or on continuous rating scales. In addition, general guidelines are given for the experimental set-up, the test scenarios, as well as the selection of test participants.

2.3. Prediction models

An approach to predict user judgments on the basis of interaction parameters has been proposed by Walker et al. [8]. Their PARADISE (PARAdigm for DIalogue System Evaluation) model combines a set of interaction parameters κ (task success) and c_i into an estimation of user satisfaction, *US*, in the following way:

$$US = \alpha \cdot N(\kappa) - \sum_{i=1}^{n} w_i \cdot N(c_i) \qquad (1)$$

In this equation, *US* is calculated as the arithmetic mean of eight or nine user judgments on Likert scales, and N(.) is the Z-score normalization function, normalizing each input parameter to a mean of zero and unity variance.

Input interaction parameters and target user judgments first have to be determined in a controlled laboratory experiment. Then, a multivariate linear regression analysis is carried out, determining the weighting coefficients α and w_i of the linear prediction function. This function once being determined, it can be used to predict mean user judgments for unseen dialogues without directly asking the users.

3. Experimental data

The three approaches have been analyzed on the basis of experimental data obtained with two different systems. The system characteristics are outlined in Section 3.1, and Section 3.2 describes the test design and run. Some interesting results from the data analysis are summarized in Section 4; further details on the experimental set-up and the data analysis can be found in [4] and [5].

3.1. Experimental systems

The first system is a telephone-based spoken dialogue system for obtaining restaurant information, called BoRIS (Bochumer Restaurant-Informations-System). The system is implemented as a finite-state machine and optionally uses an explicit confirmation strategy. Because the speech recognition accuracy was foreseen to be too low when the experiment was carried out, the speech recognition module was replaced by a transcribing wizard producing a close-to-perfect transcription of the user speech. On this transcription, errors have been generated in a controlled way, leading to an adjustable recognition performance of between 60 and 100%. Speech output is implemented either via pre-recorded messages from two non-professional speakers (1 male, 1 female), or via a text-to-speech (TTS) system. The individual system options have been combined in order to generate 10 system configurations differing in speech recognition performance, speech output, and confirmation strategy.

The second is a smart-home system developed in the EC-funded IST project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces). It provides speech control over a number of domestic devices (TV, video recorder, electronic program guide, lights, blinds, fan, answering machine) through a unique interface with speech input and speech/audio-visual output. In addition, the user will directly experience the device feedback (TV switching, blinds operation, etc.) when operating the system in the home environment. As with the BoRIS system, the speech recognizer was replaced by a transcribing wizard, but the transcription was left unmodified, leading to a recognition performance of close to 100%. The system is embodied via a speaking avatar visualized on a screen, via an immaterial assistant providing feedback through a set of loudspeakers in the whole home environment, or via "intelligent devices" where a specific voice is connected to each individual device.

3.2. Test design

Two experiments have been carried out with these systems. In exp. 1, test subjects interacted five times with the BoRIS system over a simulated phone connection in an office environment, following five different scenarios (tasks for obtaining restaurant information). In exp. 2, test subjects were asked to solve three scenarios in a simulated home environment, including the operation of several domestic devices each (12-14 tasks per scenario). The experiments were carried out with different user groups, at distant points in time and in different room environments.

Following each interaction with the system, the subjects had to rate 26 (37 in exp. 2) different statements or questions following the procedure of ITU-T Rec. P.851 [1]. During each interaction, text and audio log files were produced by the system. On the basis of the logged information, 52 interaction parameters (53 for exp. 2) have been extracted, using an annotation tool which has been developed in [11]. 40 subjects (11 f, 29 m) interacted with the BoRIS system (18-53 years, mean 29 years), and 24 subjects (10 f, 14 m, 19-29 years, mean 23.7 years) interacted with the INSPIRE system. Details on the experimental procedure, the questionnaires, the parameters and the annotation procedure are given in [4] and [5].

4. Data analysis and discussion

The experimental results have been analyzed with respect to their validity – how well do they measure what they aim at measuring – and partly also their reliability – how well do they measure what they actually measure. Section 4.1 extracts the dimensions underlying the interaction parameters and user judgments, providing an indication of the object of measurement, and thus of the reliability of the respective approach. A limited reliability analysis of the questionnaire used in exp. 2 is presented in Section 4.2, and Section 4.3 addresses the correlation between interaction parameters and user judgments observed in this experiment. Finally, the validity and reliability of prediction models is investigated in Section 4.4.

4.1. Underlying dimensions

The interaction parameters extracted from exp. 1 have been analyzed using a factor analysis and a hierarchical cluster analysis, see [5]. Both analyses reveal clusters related to speech input performance, duration and communication efficiency, meta-communication, the system's answer capability, task success, and cognitive demand. Apparently, the parameters address quality on the level of the dialogic interaction between user and system. Global quality aspects like usability, user satisfaction and acceptability are not directly covered by the interaction parameters.

A principal component analysis of the user judgments collected in exp. 2 reveals 8 factors underlying the subjective user judgments [4]. These factors have been labeled "acceptability", "cognitive demand", "task efficiency", "system errors", "ease of use", "cooperativity" and "speed of the interaction"; only one factor could not been interpreted in an unambiguous way. Although these dimensions are not fully congruent with the dimensions extracted from the SASSI questionnaire (see Section 2.2), there are several similarities: SASSI's "likeability" dimension, which includes both the overall opinion about the system and the affect/feeling of the user, seems to be related to the "acceptability" dimension of exp. 2; SASSI's "system response accuracy" dimension contains several statements which are also found in the "task efficiency" dimension of exp. 2; and the "cognitive demand" and the "speed" dimensions are common to both analyses. In turn, SASSI's "annoyance" and "habitability" dimensions were not extracted as separate factors from the exp. 2 data.

4.2. Reliability analysis

The reliability of the questionnaire used in exp. 2 has been determined by calculating Cronbach's α for the statements/questions loading higher +/- 0.4 on a particular factor. The results show that $\alpha \geq 0.7$ for all but the cooperativity factor; the first four factors show even $\alpha \geq 0.8$, which is generally regarded as sufficient for widely used scales.

4.3. Correlation analysis

For exp. 2, correlations between interaction parameters and user judgments have been calculated. The obtained values were disappointingly low. Although the speech recognizer was simulated and thus showed a nearly perfect performance, highest correlations were found for the recognition-related parameters with the perceived control over the dialogue (0.60), the pleasantness (0.51), the usability (0.47) and the smoothness of the dialogue (0.42). Correlations were particularly low in cases of "obvious" relationships: Between the understanding accuracy parameter and the perceived understanding of the system only 0.41, between the measured and perceived length of the interaction only 0.09, and between annotated and perceived task success only 0.15. Apparently, the user judgments are guided by characteristics which are not covered by the interaction parameters so far.

4.4. Prediction model analysis

The discrepancy between user judgments and interaction parameters is also reflected by the prediction accuracy of PARADISE-style regression models. Such models have been calculated from the data of both experiments, predicting either an overall quality judgment (target A) or the arithmetic mean of all user judgments (target B), see [12]. The entire set of interaction parameters has been used as an input to the regression analysis, amended by an expert annotation of task success (*TSw* or κ), or by the user's judgment of task success (*TSr*).

Table 1: Model performance on training data.

System	Task success	Output target	Model	
			R^2_{corr}	n
BoRIS	*TSw*	A	0.165	3
BoRIS	*TSw*	B	0.309	6
BoRIS	κ	A	0.165	3
BoRIS	κ	B	0.279	5
BoRIS	*TSr*	A	0.356	2
BoRIS	*TSr*	B	0.597	11
INSPIRE	*TSw*	A	0.247	2
INSPIRE	*TSw*	B	0.420	4
INSPIRE	*TSr*	A	0.436	3
INSPIRE	*TSr*	B	0.509	4

The prediction results for the training data are presented in Table 1. In this and the following tables, the model performance is expressed by the amount of variance in the user judgments R^2_{corr} which is covered by the respective model, using the number of input parameters n which have been selected by the stepwise inclusion algorithm of the regression analysis. In most cases, only between 30 and 50% of the variance is covered by the respective model. This indicates that about half of the information which is contained in the user judgments can not be modeled on the basis of interaction parameters – at least not with the linear PARADISE approach.

Table 2: Prediction accuracy for different user groups.

Training			Test	
Subj. no.	R^2	n	Subj. no.	R^2
7-24	0.421	3	1-6	0.203
1-6 & 13-24	0.529	3	7-12	0.116
1-12 & 19-24	0.425	3	13-18	0.124
1-18	0.536	5	19-24	0.025

Table 3: Prediction accuracy for cross-system models.

Target	Training			Test	
	System	R^2	n	System	R^2
A	BoRIS	0.181	3	INSPIRE	0.108
A	INSPIRE	0.283	2	BoRIS	0.001
B	BoRIS	0.331	6	INSPIRE	0.011
B	INSPIRE	0.452	4	BoRIS	--

Prediction results are even worse when the models are tested on an independent set of data. This has been tested with the exp. 2 data (target B, *TSw* for describing task success), using the data from 18 subjects for training and the remaining six subjects for testing, see Table 2. In all cases, R^2 decreases significantly from the baseline of $R^2_{corr} = 0.420$. The same problem occurs when a model is trained on one system and tested on another one, see Table 3. Apparently, the models are specific to the user group and the system they have been developed for.

The latter finding seems to be in contrast to observations described by the PARADISE authors in [13], where the prediction accuracy suffered from changes in the user group, but not from changes between systems. Two potential reasons may be given so far. Firstly, the BoRIS and INSPIRE systems differed to a high degree, higher that it can be expected for the three systems described in [13] which are presumably based on the same system components, and were all operated over the telephone channel. Secondly, a much larger number of interaction parameters have been collected in the present study; the selected values may therefore be more specific for the systems considered in the training.

5. Conclusions and future work

Three types of approaches have been described which aim at quantifying the quality of telephone-based spoken dialogue services. Real measurements of quality can only be obtained from real or test users interacting with the system, for example using questionnaires. The flow of the interaction between user and system as well as the performance of system components can be quantified in a parametric way, in terms of interaction parameters. These parameters are useful in the system development and set-up phase, but they are not directly related to quality. In fact, correlations to subjective user judgments are disappointingly low, and simple prediction models like PARADISE fail in validly predicting user judgments. In addition, quality predictions seem to be limited to a particular user group and system.

Two major reasons may be responsible for these limitations. Firstly, system characteristics which are relevant for quality from a user's point-of-view are not yet covered in the parametric description, e.g. the speech output quality, the personality of the system, the cognitive demand required from the user, or service and task efficiency aspects. Secondly, there is no reason why a simple linear model should be sufficient for predicting quality. On the contrary, optimum values may exist for some of the parameters, and this can only be taken into account by non-linear algorithms.

In order to find more informative interaction parameters and define better prediction algorithms, it is necessary to collect data with more diverse systems and user groups. The mentioned ITU-T Recommendation [1] and Supplement [9] have been set up to support a standardized evaluation protocol, and to ensure comparability of the collected data. ITU-T Study Group 12 plans further activities in this direction, as it is described under http://www.itu.int/ITU-T/studygroups/com12/q12roadmap/index.html.

6. Acknowledgements

Part of the present work has been performed at the Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum. The case study was supported by the EC-funded IST project INSPIRE (IST-2001-32746, see http://www.knowledge-speech.gr/inspire-project/index.html). Partners of INSPIRE were: Knowledge SA, WCL/University of Patras, TNO Human Factors, Philips Electronics Nederland B.V, ABS Jena, IKA/Ruhr-Universität Bochum and EPFL. The author would like to thank his former colleagues at IKA for their help during the experiments, and Deutsche Telekom Laboratories for their support in publishing the results.

7. References

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

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

[3] Fraser, N., "Assessment of Interactive Systems", in: *Handbook of Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, Berlin, 564-615, 1997.

[4] Möller, S., Smeele, P., Boland, H. and Krebber, J., "Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study", accepted for *Computer Speech and Language*, 2005.

[5] Möller, S., *Quality of Telephone-Based Spoken Dialogue Systems*, Springer, New York, 2005.

[6] Grice, H.P., "Logic and Conversation", in: *Syntax and Semantics*, Vol. 3: Speech Acts (P. Cole and J.L. Morgan, eds.), Academic Press, New York, 41-58, 1975.

[7] Bernsen, N.O., Dybkjær, H. and Dybkjær, L., *Designing Interactive Speech Systems: From First Ideas to User Testing*, Springer, Berlin, 1998.

[8] Walker, M.A., Litman, D.J., Kamm, C.A. and Abella, A., "PARADISE: A Framework for Evaluating Spoken Dialogue Agents", in: *Proc. ACL/EACL 35th Meeting*, Madrid, 271-280, 1997.

[9] ITU-T Contribution COM 12-17, *Proposal for a New Supplement to P-Series Rec. "Parameters Describing the Interaction with Spoken Dialogue Systems"*, Federal Republic of Germany (Author: S. Möller), ITU-T SG12 Meeting, 17-21 October, Geneva, 2005.

[10] Hone, K.S. and Graham, R., "Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)", *Natural Language Engineering*, 6:287-303, 2000.

[11] Skowronek, J., *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem* (Development of Quality-of-Service-Prediction Models for the Interaction with a Spoken Dialogue System), unpublished Diploma thesis, IKA, Ruhr-Universität Bochum, 2002.

[12] Möller, S., "Towards Generic Quality Prediction Models for Spoken Dialogue Systems - A Case Study", in: *Proc. 9th European Conf. on Speech Communication and Technology (Interspeech 2005)*, Lisboa, 2489-2492, 2005.

[13] Walker, M., Kamm, C. and Litman, D., "Towards Developing General Models of Usability with PARADISE", *Natural Language Engineering*, 6:363-377, 2000.