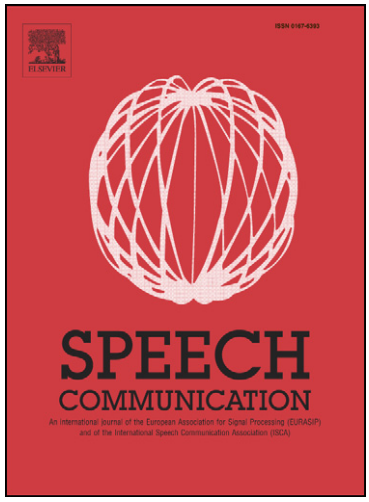
Predicting the Quality and Usability of Spoken Dialogue Services

Sebastian Möller, Klaus-Peter Engelbrecht, Robert Schleicher



Predicting the Quality and Usability of Spoken Dialogue Services

Sebastian Möller, Klaus-Peter Engelbrecht, Robert Schleicher

PII: S0167-6393(08)00035-6
DOI: 10.1016/j.specom.2008.03.001
Reference: SPECOM 1694

To appear in: *Speech Communication*

Received Date: 29 June 2007
Revised Date: 3 January 2008
Accepted Date: 14 March 2008

Please cite this article as: Möller, S., Engelbrecht, K-P., Schleicher, R., Predicting the Quality and Usability of Spoken Dialogue Services, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.03.001

Predicting the Quality and Usability of Spoken Dialogue Services

Sebastian Möller, Klaus-Peter Engelbrecht, Robert Schleicher

Quality and Usability Lab
Deutsche Telekom Laboratories
Berlin University of Technology
Ernst-Reuter-Platz 7
D-10587 Berlin
Germany
E-Mails: sebastian.moeller@telekom.de; klaus-peter.engelbrecht@telekom.de;
robert.schleicher@telekom.de

Corresponding author:

Sebastian Möller
Quality and Usability Lab
Deutsche Telekom Laboratories
Berlin University of Technology
Ernst-Reuter-Platz 7
D-10587 Berlin
Germany
Tel.: +49 30 8353 58465
Fax: +49 30 8353 58409
E-Mail: sebastian.moeller@telekom.de

Abstract

In this paper, we compare different approaches for predicting the quality and usability of spoken dialogue systems. The respective models provide estimations of user judgments on perceived quality, based on parameters which can be extracted from interaction logs. Different types of input parameters and different modeling algorithms have been compared using three spoken dialogue databases obtained with two different systems. The results show that both linear regression models and classification trees are able to cover around 50% of the variance in the training data, and neural networks even more. When applied to independent test data, in particular to data obtained with different systems and/or user groups, the prediction accuracy decreases significantly. The underlying reasons for the limited predictive power are discussed. It is shown that – although an accurate prediction of individual ratings is not yet possible with such models – they may still be used for taking decisions on component optimization, and are thus helpful tools for the system developer.

Keywords

Spoken dialogue system, quality, usability, prediction model, optimization.

1. Introduction

Spoken dialogue services (SDSs) with increasingly sophisticated speech and language processing capabilities are available on the market offering information, transactions, and device control. Examples are train and air timetable information services, telephone banking, tourist services, or smart speech-controlled home environments. The underlying systems have speech recognition and interpretation capabilties, a dialogue manager which maintains the interaction with the user and is capable of meta-communication (e.g. feedback, correction, or reference solution), a response generation component, as well as a module for the output of concatenated pre-recorded or synthesized speech. Because of their interdependence, improvement of these underlying modules is not an easy, but necessary task in order to deliver an optimal quality to the human user.

The *quality* a user perceives during the interaction with the system is a perceptual event. It results from a perception and a judgment process, during which the user compares what s/he perceives with what s/he desires or expects, considering the own background knowledge as well as the experience with this or other systems (Jekosch, 2005). Because of the complexity of the perception process, and because of the inaccessibility of the desired or expected reference the perceptual event is compared to, quality can until now only be measured by asking the user about his/her percept, e.g. using questionnaires with different types of rating scales.

Hone and Graham (2000, 2001) developed such a questionnaire for systems with speech input capability (so-called SASSI questionnaire), based on standard usability evaluation tools. The questionnaire has been extended towards systems with speech output capability, and is now recommended for evaluating telephone-based SDSs by the International Telecommunication Union, ITU-T (ITU-T Rec. P.851, 2003).

Considering the large effort necessary for carrying out subjective tests under laboratory conditions, SDS developers try to limit the need for such tests, and the collection of both user- and system-related information herein. System-related diagnostic information can be extracted from log-files collected during real or test interactions with their systems. This information is captured in so-called *interaction parameters* which are determined on the signal level (number of turns, turn duration, pause duration, speech and noise levels) as well as on the symbolic level (number of words, word accuracy, concept accuracy, appropriateness of system prompts, etc.). A large number of interaction parameters have been developed for this purpose, see e.g. Simpson and Fraser (1993), Fraser (1997), or Möller (2005). Recently, an extensive set has been recommended for telephone-based systems by the ITU-T in its Suppl. 24 to P-Series Recommendations (2005). Interaction parameters describe the *performance* and the *behavior* of the system and the user during the interaction, but do not necessarily reflect perceived quality.

Interaction logs can also be annotated to detect and identify interaction problems, and thus to determine the *usability* of a spoken dialogue system. Usability is commonly defined as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-11, 1998). Usability is degraded when interaction problems occur. Such problems may be quantified using dedicated classification schemes. For example, Bernsen et al. (1998) classified dialogue design and "user errors" according to

cooperativity guidelines which have been violated during the interaction. Oulasvirta et al. (2006) classified errors on different layers such as goal-level, task-level, command-level, and concept level errors, and judged also the consequences (stagnation, regression or partial progress) linked to the occurrence of each error. It was shown that especially the consequences correlate moderately with user judgments on quality. Constantinides and Rudnicky (1999) classified problems along the bows of a fishbone diagram, each bone indicating a specific source of errors (recognition, understanding, output, etc.) and thus providing direct feedback to the system developer. The frequency of such errors is an indication of poor system performance, although it is not necessarily linked to perceived quality.

Several approaches have been made to relate performance metrics to perceived quality. The most popular one is the PARADISE framework developed by Walker et al. at AT&T (Walker et al., 1997). The idea is to estimate subjective judgments of "user satisfaction" (calculated as the arithmetic mean of 8-9 questionnaire judgments) as a linear combination of several parameters which can be determined from interaction logs. In this way, questionnaires may partly be skipped and the need for subjective testing may significantly be reduced during the system design process. It would also become possible to estimated user judgments from log data during real-life usage, when the users are not accessible to the service operator.

Parameters used as an input to the PARADISE model relate to task success as well as to dialogue costs, the latter being composed of "dialogue efficiency costs" (e.g. the number of utterances) and "dialogue quality costs" (e.g. the recognition performance or the number of help requests). Different modeling approaches have also been presented, using regression tree models (Walker et al., 2000b; Hastie et al., 2002), or neural networks (Compagnoni, 2006).

PARADISE models have been extensively used for system optimization and also "user satisfaction" prediction. Usually, the amount of variance of the training data (R^2) which is captured by a model with 3-4 predictors is around 35-50%, see e.g. Walker et al. (1998) and Kamm et al. (1998). However, the models have only rarely been analyzed with respect to their predictive power on unseen data. An exception is the analysis by Walker et al. (2000) showing that an extrapolation to other systems may be possible, but that a change in the user group (namely from novices to experts) significantly reduces R^2. In contrast to this, Möller (2005a) found that the predictive power was significantly reduced when extrapolating from one system to another.

It is the aim of the present paper to analyze the capacity of models relating performance indices to quality judgments, both in interpolating known data and in extrapolating towards unknown data. Data from three interaction experiments were available for this purpose, addressing two types of systems. The experimental data will be briefly reviewed in Section 2. Using these data, different types of models are constructed, based on different input parameters, modeling algorithms and target variables, see Section 3. Their predictive power is analyzed on distinct training and testing sets, and their extrapolation capability towards new system versions and systems is examined in Section 4. The results are discussed in Section 5, and in Section 6 it is shown how – despite the observed inherent limitations – models can still be very useful for system design and optimization. Section 7 summarizes the main findings and gives a perspective for future work.

2. Experimental Data

Three databases were available for the analysis. They have been collected with two different types of systems, namely a telephone-based system for restaurant information (BoRIS) and a spoken dialogue interface to domestic devices (INSPIRE). The latter system has been investigated at two different stages of the development process, reflecting two system versions differing in their vocabulary, speech understanding capabilities, and system prompts. In order to avoid an excessive impact of the speech recognition component which was not yet optimized when the experiments were carried out, this component has been replaced in all experiments by a transcribing wizard. In two of the experiments, speech recognition errors have been generated by introducing substitutions, insertions and deletions according to a pre-determined confusion behavior, so that the system reflects real-world behavior during these experiments.

All experiments have been carried out in a laboratory environment at Ruhr-University Bochum, Germany, to ensure stable conditions across subjects. During each experiment, the test subjects had to carry out scenario-driven interactions with the system under test. Subjective quality judgments were collected on questionnaires designed according to ITU-T Rec. P.851 after each interaction. The interactions have been logged, and the log-files have later been annotated by experts in order to collect performance indices. As the interactions have been carried out at different points in time and for different purposes (namely system optimization starting from the current state of development), the participant group of each experiment was different.

In the following Sections 2.1 and 2.2, we briefly review the relevant experimental characteristics. Section 2.3 gives an overview of the resulting three databases.

2.1 BoRIS Restaurant Information System

With the help of the Bochumer Restaurant-Informations-System (BoRIS), a user can obtain information about restaurants in the town of Bochum and its surroundings by specifying the desired suburb, the type of food, the day and the time the restaurant should be open, as well as the price category. The system is implemented as a finite-state machine with the help of the CSLU toolkit (Sutton et al., 1998). Since 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 and realistic way, leading to an adjustable recognition performance between 60 and 100%. The error generation is based on a confusion matrix which has been determined for a commercial recognizer, and then scaled to a given target "recognition rate" (Möller, 2005). With the help of this matrix, some of the transcribed words have been replaced by others (substitutions), deleted or inserted, leading to a defined "recognition accuracy". The method has been validated in Möller (2005) and showed an acceptable agreement between observed and target recognition accuracy (median of the observed values was within 2% of the respective target recognition performance).

In addition to the simulation of recognition errors, the system could be set to an explicit confirmation strategy to avoid misunderstandings, or (in a different version) be used without confirmation at all. Speech output was implemented either via pre-recorded messages from two non-professional speakers (1 male, 1 female), or via a text-to-speech (TTS) system. Different speech output options could be used for the fixed and the variable (e.g. restaurant names and addresses) message parts. For the experiment described here, the individual system options have been combined in order to generate 10 system configurations differing in speech recognition performance, speech output, and confirmation strategy, see Table 1.

Table 1: Different configurations of the BoRIS system used in experiment 1.

Target recognition rate	*Speech output*		*Confirmation strategy*
	Fixed	*Variable*	
100	female	female	no
100	male	female	no
100	female	TTS	no
100	TTS	TTS	no
70	female	female	no
100	female	female	explicit
90	female	female	explicit
80	female	female	explicit
70	female	female	explicit
60	female	female	explicit

40 participants (11 f, 29 m) interacted five times with the BoRIS system through a simulated telephone line in an office environment. They were between 18 and 53 years old, with a mean of 29 years, and were paid for their service. The majority of subjects did not have any experience with spoken dialogue services, but most of them knew the town of Bochum and some of the local restaurants.

Test participants had to follow four scenarios which provided criteria for a restaurant search; in the fifth interaction, the participants were asked to define criteria on their own before interacting with BoRIS. After each interaction, the participants had to fill in a questionnaire with 26 items relating to different aspects of the system, including its overall quality. The interactions were logged, and log-files were transcribed and annotated by a human expert after the experiment. Details on the experimental design, the questionnaire and the annotation procedure are given in Möller (2005). On the basis of the annotation, 52 interaction parameters were extracted for each dialogue, quantifying user and system behavior and performance. In a second step, interaction problems and consequences have been annotated according to the classification scheme of Oulasvirta et al. (2006). This resulted in another 6 problem frequencies and 3 consequence frequencies per dialogue which have been annotated and counted. The parameters used for this investigation are listed in Section 3.1.

2.2 INSPIRE Smart-home System

The INSPIRE smart-home system gives a spoken-dialogue access to domestic devices such as lamps, blinds, fans, TVs, video recorders, electronic program guides, and answering machines. It has been set up in the frame of the EU-funded IST project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces; IST 2001-32746) in a living-room environment at Ruhr-University Bochum. The system is implemented with the help of generic dialogue nodes (Rajman et al., 2003) associated with each attribute to be provided by the user in order to carry out a specific task. Dialogue nodes contain standard meta-communication capabilities (misunderstanding, non-understanding, help) and are connected through a global branching logic.

As in the BoRIS system, the speech recognizer has been replaced by a transcribing wizard. Whereas the wizard's transcriptions have been kept unchanged in the first experiment, controlled amounts of recognition errors have been generated in the second one to simulate realistic speech input behavior, as it is described in Trutnev et al. (2004). On the system output side, speech has been generated by concatenating pre-recorded phrases. In the first experiment, the system was embodied either in terms of an avatar displayed on a screen in the living room (talking-head metaphor), via loudspeakers mounted close to the devices to-be-operated (intelligent devices metaphor), or via ceiling loudspeakers generating a more-or-less diffuse sound field (ghost metaphor). In the second experiment, only the ghost metaphor was used. A more detailed analysis of the effect of the system metaphor can be found in Möller et al. (2005d).

The two experiments had been carried out in two different states of the system development process, reflecting two system versions. Apart from the metaphor, system versions differed with respect to the vocabulary (which was extended and improved from the first to the second experiment), the speech understanding capabilities (optimized keyword-matching), and the system prompts (shorter and less ambiguous prompts in some nodes). In addition, the system was extended with a 'macro function', i.e. combinations of actions could now be triggered with a single command (e.g. "everything off" to switch off all devices). This function was explained to the participants before the experiment.

During the first experiment (INSPIRE 1), each of 24 participants (10 f, 14 m) carried out three scenario-guided interactions with the system. Each interaction consisted of 9-11 tasks which were linked in a kind of short story, combining all devices to create interactions of comparable length and complexity. Test participants were mostly students or employees of the university and were between 19 and 29 years old (mean: 23.7 years). The majority had some prior knowledge of SDSs. The second (INSPIRE 2) experiment followed mainly the same test protocol and involved 28 participants (14 f, 14 m, 19-50 years, mean 26.4 years). Participants were paid for their service.

Participants had to compile questionnaires with 37 items after each interaction. The questionnaires were adapted from the respective BoRIS questionnaires and contained some additional items to reflect functionality differences of the INSPIRE system compared to BoRIS. The interactions were logged, transcribed and annotated, resulting in 53 interaction parameters, 6 interaction problem frequencies and 3 consequence

frequencies describing each dialogue. Details on the experimental set-up and the questionnaire are given in Möller et al. (2007).

2.3 Resulting Databases

All data have been processed for statistical analysis in SPSS (Statistical Package for the Social Sciences, SPSS Inc). For each experiment, data obtained with different system versions have been merged, except in Section 4.5 where individual systems versions are explicitly addressed. For the BoRIS experiment, the data includes 26 quality judgments and 52 interaction parameters for each of 197 dialogues from 40 test subjects; 3 dialogues had to be stopped due to system breakdowns. For INSPIRE 1, the set contains 37 quality judgments and 53 interaction parameters for each of 68 dialogues carried out by 24 subjects; 4 dialogues could not be annotated due to logging problems. Unfortunately, not all subjects answered all questions, leading to slightly less ratings (66) in some cases in the INSPIRE 1 experiment. The database of the INSPIRE 2 experiment contains 37 quality judgments and 51 interaction parameters for each of the 84 dialogues from 28 test subjects.

The subjective ratings have been transformed into numbers between -2 and +2, the latter corresponding to the most positive (shortest, quickest) rating. All judgments and interaction parameters have been z-score-normalized to show a mean of zero and unity variance.

3. Prediction Models

From the described databases, prediction models have been calculated using different sets of input parameters, target variables, as well as modeling algorithms. The variants of each of these parts of the model are outlined in Section 3.1 to 3.3. Section 3.4 defines criteria for assessing the prediction performance for such models.

3.1 Input Parameters

Input parameters quantify the interaction behavior and performance of user and system, as well as any problems occurring during the interaction. Four sets of parameters have been used here to predict subjective quality ratings:

- *Set 1* contains interaction parameters which are defined in ITU-T Suppl. 24 to P-Series Rec. (2005) and which could be collected with the BoRIS and INSPIRE systems. From the full set of 53 (BoRIS) and 51 (INSPIRE) parameters, parameters which showed zero-values in more than 95% of all cases were eliminated, as well as parameters which are correlated with each other by their definition. For example, the number of turns (*#Turns*), the number of system turns (*#System Turns*) and the number of user turns (*#User Turns*) were correlated because of the strict alternation of turns in our system. The final set includes 27 parameters for the experiment with the BoRIS system, and 30 parameters for both experiments carried out with the

INSPIRE system. These parameters are listed in Table 2, and exact definitions can be found in Möller (2005) and ITU-T Suppl. 24 to P-Series Rec. (2005).

- *Set 2* is a restricted set of interaction parameters, namely those which were used by Walker et al. (1997) in the definition of PARADISE, provided that they could be measured in our experiments, see Table 2. In addition to these "dialogue cost" parameters, two options were used for describing task success:
  - *Set 2a*: Expert annotation of task success, in terms of *TSw* and κ. κ is determined on the basis of attribute-value pairs provided at the end of the dialogue, and corrected for chance agreement. *TSw* has been calculated by classifying task success according to Fraser (1997) into succeed, succeed with constraint relaxation by the system or by the user or both, succeed in spotting that no answer exists, or failure due to system or user behavior, and then weighting the individual task success labels as described in Möller (2005).
  - *Set 2b*: User judgment on task success, i.e. the rating on the statement "the system provided the desired information" for BoRIS, or the statement "the system did not always do what I wanted" for INSPIRE 1 and 2.

  Despite being in contrast to the idea of quality prediction models – namely to get independent of direct user judgments – subjective ratings on task success have frequently been used with PARADISE models. We include Set 2b for comparison with the figures cited in Section 1.
- *Set 3* contains interaction problem classes annotated according to the scheme described by Oulasvirta et al. (2006), and modified by Engelbrecht (2006) to form the following classes: Goal-level errors (i.e. the system does not posses the function or capability assumed in the user's request), task-level errors (i.e. the user does not understand how to reach the goal in the interaction with the system), representation-level errors (i.e. the user issues a command that would be valid if the system represented the "world" in a different way), command-level errors (i.e. the user makes use of linguistic variations like synonyms or grammar which are not understood by the system), technical errors (interaction failures which even a completely cooperative user cannot influence, e.g. ASR errors), as well as other errors not captured by these classes. In addition to the errors, the consequences of each error (stagnation, regression, partial progress despite the error) have been counted, leading to a set of 9 input parameters.
- *Set 4* contains both the full set of interaction parameters and the error and consequence frequencies, i.e. the joint Set 1 and Set 3.

Table 2: Input parameters available in the BoRIS and INSPIRE experiments.

Variable			*Available in experiment*		*Used in parameter set*	
Abbreviation	*Description*	*Unit*	*BoRIS*	*INSPIRE*	*1*	*2*
DD	Dialogue duration	ms	X	X	X	X
STD/UTD	System/user turn duration	ms	X	X	X	
SRD/URD	System/user response delay	ms	X	X	X	
#Turns	Number of turns	1	X	X	X	X
WPST/WPUT	Words per system/user turn	1	X	X	X	
%Barge-Ins	Percentage of user barge-in attempts	1[%]	X	X	X	X
%System Error Messages	Percentage system error messages	1[%]	X	X	X	
%System Questions	Percentage system questions	1[%]	X	X	X	
%User Questions	Percentage user questions	1[%]	X	X	X	
%PA:CO	Perc. correctly parsed user utterances	1[%]	X	X	X	
%PA:PA	Perc. partially parsed user utterances	1[%]	X	X	X	
%PA:PA	Perc. failed-to-be-parsed user utterances	1[%]	X	X	X	
SCR / UCR	Perc. system /user correction turns	1[%]	X	X	X	
%CA:AP	Perc. appropriate system prompts	1[%]	X	X	X	
%CA:IA	Perc. inappropriate system prompts	1[%]	X	X	X	
IR	Perc. of implicitly recovered problems	1[%]	X	X	X	
IC	Understanding accuracy on a concept level	1[%]	X	X	X	X
UA	Understanding accuracy on an utterance level	1[%]	X	X	X	
WA	Word accuracy	1[%]	X	X	X	
NEU	Number of errors per utterance	1	X	X	X	
WEU	Word error per utterance	1	X	X	X	
κ	Task success measure *kappa*	1	X			X
TSw	Weighted task success per dialogue	1	X	X	X	X
%ASR Rejections	Perc. speech recognizer rejections	1[%]		X	X	
%System Help Messages	Percentage system help messages	1[%]		X	X	
%Help Requests	Percentage help request from the user	1[%]		X	X	
%Cancel Attempts	Perc. cancel attempts from the user	1[%]		X	X	

3.2 Target Variables

The prediction models estimate user judgments related to perceived quality and usability. A large number of judgments have been collected in both experiments, which evidently

are not uncorrelated to each other. In fact, a factor analysis of the BoRIS data reveals 5 underlying perceptual dimensions (Möller, 2005; Engelbrecht, 2006), and for the INSPIRE 1 data, 8 dimensions could be extracted. These dimensions are not identical, but a large-scale comparison of different systems described in Möller (2005c) shows that overall acceptability, communication efficiency and cognitive effort are amongst the ones explaining most of the overall variance in the judgments.

While the use of individual ratings may introduce a significant amount of noise in the prediction, averaging over different ratings may lead to a loss of information. As a compromise, it was decided to use different prediction target variables:

- The arithmetic mean of all positively- and negatively-aligned user judgments in the respective questionnaire (*AM*).
- The user's direct judgment on overall quality, obtained at the beginning of the questionnaire on a continuous rating scale (*OQ*).
- The perceptual dimensions "overall acceptability" (*ACC*), "efficiency" (*EFF*) and "cognitive effort" (*COE*), calculated by averaging the subjective ratings of the questionnaire items which loaded higher than 0.6 on the respective factor.

3.3 Prediction Algorithms

Considering the complex interdependence of system components and the dependency of the input parameters, a simple linear modeling approach like in PARADISE may not be an optimum way to predict perceived quality and usability. However, more complex modeling algorithms may increase the likelihood that the model optimizes on the specific set of training data, and is less reliable in predicting new – unseen – data (overfitting of the model).

A first analysis of non-linear relationships carried out by Compagnoni (2006) did not reveal any clear relationship between input parameters and target variables which might have been used for a non-linear regression. Neural networks, in particular multi-layer perceptron nets, showed a prediction performance similar to the one of linear regression models; however the resulting model cannot be checked for plausibility because the underlying rules are not accessible.

As a consequence, it was decided to consider the following types of models for the analysis:

- LR: A multivariate linear regression (LR) model as in the PARADISE framework. The respective input parameters (see Section 3.1) were z-score-normalized, and if a Poisson distribution was detected for an individual parameter, the square root was taken instead of the parameter value. Relevant parameters were selected with a stepwise (forward-backward) inclusion algorithm, and missing values were replaced by the overall means in the analysis.
- Decision trees: Such trees allow simple and interpretable rules to be derived from training data, and then being tested on independent data. Two different approaches were used: Classification And Regression Trees (CARTs) and Chi-squared Automatic Interaction Detection (CHAID). Both procedures aim at splitting the original sample

in as homogenous subgroups as possible. With CART, each split results in two children nodes, whereas CHAID allows for multiple children nodes, i.e. non-binary splits. To avoid overfitting, the minimum terminal node size was set to 6 cases. Additionally, pruning (removal of meaningless nodes) was done manually based on plausibility considerations once the trees were calculated. As for the LR model, input parameters were z-score-normalized, and the square root was taken in case of Poisson data distribution.

- MLP: Neural networks allow classifiers to be built without requiring knowledge about the meaning of the respective input parameters. We opted for a very simple Multi-Layer Perceptron (MLP) model which has proven successful in speech recognition. The network consists of one hidden layer, compound of twelve neurons with log-sigmoid transfer function, and one output neuron with a linear transfer function. The network was trained with a feed-forward backpropagation training function. In order to prevent overfitting, a MATLAB (The MathWorks Inc.) routine was used which smoothens the prediction function by employing Bayesian regularization (MacKay, 1992), keeping the weights and biases in the network small. In order to determine the optimum model, a stepwise inclusion method has been implemented: The first training is done for each single parameter separately; in the next iteration, the parameter showing the highest correlation remains fixed and the best additional parameter is determined by trying each of the remaining ones[1]. Input parameters were z-score-normalized; however, no parameters were square-rooted, because the neural network can cope with both versions of the parameter equally well.

3.4 Performance Evaluation

Both within-data (interpolation) performance and out-of-data (extrapolation) performance have been analyzed. As the amount of data is limited due to the effort required in carrying out subjective tests, three cases have been distinguished:

- All cases (ALL): Training and test sets are identical, i.e. the performance related to the coverage of the training data.
- Leave-one-out (L1O): Data from one user is omitted in the training, and the respective data is taken for testing a model trained on the remaining n-1 users. This procedure is repeated for all n users in order to make optimum use of the available data, and the obtained performance indices are averaged over all analyses. Because we assume that the interaction behavior and subjective judgments may be very user-specific, we prefer to omit a user completely instead of just omitting individual interactions.
- Cross-experiment extrapolation (CEE): A model is trained on all data from one experiment, and then used to predict the data obtained in a different experiment. This may lead to a cross-system extrapolation between BoRIS and INSPIRE.

[1] Because the calculations for the leave-one-out case (L1O, see Section 3.4) with a large number of predictors results in very long processing times, the set was restricted in this case to parameters which had proven useful in the ALL case as a first sub-set (Set ALL), and parameters included in the respective LR model as a second subset (Set LR).

The performance of the obtained models has been evaluated by means of the Pearson correlation coefficient r and by the prediction error E_p:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\cdot\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad (1)$$

and

$$E_p = \sqrt{\left(\frac{1}{N-d}\sum_{i=1}^{N}(X_i - Y_i)^2\right)} \quad (2)$$

In these equations, X_i is the subjective (target) judgment or the combination of judgments for dialogue i, Y_i is the estimated (predicted) judgment for dialogue i, $\bar{X}$ the arithmetic mean over all target variables, $\bar{Y}$ the arithmetic mean of all predictions, N the number of considered dialogues, and d the degree of freedom of the model (d = 1 in our case).

For the LR models, the amount of covered variance R^2 is more common to indicate the model performance, where $R^2 = r^2$ in our case. R^2 can be adjusted for the number of predictors k in the model:

$$R^2{}_{adj} = 1 - (1 - R^2)\cdot\left(\frac{N-1}{N-k-1}\right) \quad (3)$$

In this way, the performance of a model with many predictors – which is likely to better cover the training data, but increases the risk of overfitting – is made comparable to the R^2 of a model with fewer predictors. We have not analyzed the individual predictors here, but only cite the overall performance of the models; the interested reader is referred to Möller (2005) for an in-depth analysis of LR predictors for the BoRIS system.

4. Prediction Results

In order to compare the performance of our models to the standard PARADISE approach, a baseline has been established for each experiment in Section 4.1. We then varied the input parameter set, the target variable, as well as the modeling algorithm, see Sections 4.2, 4.3 and 4.4. We also checked the performance of the models for extrapolating across system versions and across systems in Sections 4.5 and 4.6.

4.1 Baseline Performance

The PARADISE framework proposes the use of a limited number of interaction parameters (Set 2) to predict an averaged quality judgment (*AM*) using an LR model. Whereas the original formula of PARADISE suggests the use of κ as an expert-derived measure of task success as an input to the model (Walker et al., 1997), most of the performance indicators cited in Section 1 are based on a user judgment on task success.

Consequently, we use both Set 2a and Set 2b for the comparison. The performance for each experiment is given in Table 3.

Table 3: Baseline performance on training (ALL) and independent test data (L1O) with Set 2 input parameters, *AM* target variables, and LR models. Maximum values of r, R^2 and R^2_{adj} and minimum values of E_p are marked in bold.

Configuration			*Model performance*			
Experiment	*Input*	*Training*	r	R^2	R^2_{adj}*	E_p
BoRIS	Set2a	ALL	0.468	0.219	0.207	0.892
BoRIS	Set2b	ALL	0.663	0.440	0.434	0.758
INSPIRE 1	Set2a	ALL	0.704	0.496	0.464	0.720
INSPIRE 1	Set2b	ALL	**0.753**	**0.567**	**0.532**	**0.666**
INSPIRE 2	Set2a	ALL	0.558	0.311	0.302	0.849
INSPIRE 2	Set2b	ALL	0.668	0.446	0.430	1.007
BoRIS	Set2a	L1O	0.412	0.170	0.157	0.913
BoRIS	Set2b	L1O	0.637	0.406	0.400	0.775
INSPIRE 1	Set2a	L1O	0.474	0.225	0.183	0.936
INSPIRE 1	Set2b	L1O	0.576	0.332	0.282	0.853
INSPIRE 2	Set2a	L1O	0.451	0.203	0.193	0.924
INSPIRE 2	Set2b	L1O	0.557	0.310	0.294	0.859

* To compute the adjusted R^2_{adj} for L1O, k in Formula (3) was set to the average number of variables used in the individual regressions.

For the training data (ALL), R^2 is in the range 0.22…0.50 for Set 2a, and 0.44…0.57 for Set 2b. The latter values are slightly higher than the ones cited in Section 1 for a comparable model setting. In all cases, using a subjective rating of task success (Set 2b) provides better performance than using an expert-derived one (Set 2a). However, this requires subjective ratings to be collected from the test participants, which is contrary to the aim of a prediction model – namely to get independent of direct user judgments and assess quality in advance of a user test. The adjusted R^2_{adj} values are slightly lower, but still comparable to the uncorrected ones. The prediction error is usually in the range 0.7…0.9, which is comparable to the variance observed in the subjective ratings; only for the INSPIRE 2 experiment and Set 2b it is slightly higher.

When testing the models on unseen test data (L1O), the performance of the model decreases in all cases. This shows that the models are better in interpolating the training data than in extrapolating to unseen test data. The degradation is slightly smaller for BoRIS compared to INSPIRE 1 and 2.

4.2 Impact of Input Parameters

The selection of input parameters determines which type of information is available for the prediction, and should therefore be crucial to model performance. Table 4 indicates

the performance for the BoRIS and the INSPIRE 1 experiment; INSPIRE 2 has been omitted here to save space, but shows a similar picture as INSPIRE 1.

Table 4: Performance on training (ALL) and independent test data (L1O) with different input parameter sets, *AM* target variables, and LR models.

Configuration			*Model performance*			
Experiment	*Input*	*Training*	r	R^2	R^2_{adj}	E_p
BoRIS	Set 1	ALL	0.712	0.507	0.471	0.881
BoRIS	Set 2a	ALL	0.468	0.219	0.207	0.892
BoRIS	Set 2b	ALL	0.663	0.440	0.434	0.758
BoRIS	Set 3	ALL	0.509	0.259	0.243	0.861
BoRIS	Set 4	ALL	0.712	0.507	0.477	0.866
INSPIRE 1	Set 1	ALL	0.749	0.561	0.525	0.672
INSPIRE 1	Set 2a	ALL	0.704	0.496	0.464	0.720
INSPIRE 1	Set 2b	ALL	**0.753**	0.567	0.532	**0.666**
INSPIRE 1	Set 3	ALL	0.617	0.381	0.341	0.799
INSPIRE 1	Set 4	ALL	0.748	**0.583**	**0.549**	0.673
BoRIS	Set 1	L1O	0.439	0.193	0.168	0.977
BoRIS	Set2a	L1O	0.412	0.170	0.157	0.913
BoRIS	Set2b	L1O	0.637	0.406	0.400	0.775
BoRIS	Set 3	L1O	0.463	0.214	0.198	0.888
BoRIS	Set 4	L1O	0.450	0.203	0.209	0.970
INSPIRE 1	Set 1	L1O	0.480	0.230	0.172	0.991
INSPIRE 1	Set2a	L1O	0.474	0.225	0.183	0.936
INSPIRE 1	Set2b	L1O	0.576	0.332	0.282	0.853
INSPIRE 1	Set 3	L1O	0.300	0.090	0.046	1.002
INSPIRE 1	Set 4	L1O	0.430	0.185	0.117	1.029

The comparison of input parameter sets for each experiment shows that the prediction performance on the training data (ALL) increases when augmenting the set of input parameters from Set 2a to Set 1, in that r and R^2_{adj} increase and the prediction error decreases. For the INSPIRE 1 experiment, a model with Set 1 input parameters nearly reaches the performance of a Set 2 model, but here without relying on a subjective judgment of task success. For independent test data (L1O), r and R^2_{adj} still increase, however E_p also increases from Set 2a to Set 1. Here, the performance of the Set 1 model is significantly lower than the one of a Set 2b model.

On the basis of information on error and consequence frequencies alone (Set 3), the performance of the LR model is slightly better than with Set 2a input parameters for the BoRIS experiment, but considerably worse for the INSPIRE 1 experiment. Adding this information to the interaction parameters (Set 4) increases the performance on the training data of the BoRIS experiment, but not of the INSPIRE 1 experiment. On the test data, these models do not perform significantly better than Set 1 or Set 3 models alone.

Overall, Set 1 and Set 4 seem to be the best combinations of input parameters not including subjective judgments (as in Set 2b). However, the extension of the input variables by the error and consequences classes in Set 4 does not cause a considerable improvement for the prediction. Therefore, to reach an optimum prediction performance, we have considered only the Set 1 input parameters for the subsequent experiments.

4.3 Impact of the Target Variable

The PARADISE framework estimates "user satisfaction" which is defined as the arithmetic mean of 8-10 user judgments on items as diverse as task completion, TTS performance, ASR performance, task ease, interaction pace, or system transparency. Still, there is no reason to assume that these items contribute equally to the user's overall satisfaction. The latter might also be quantified on a global scale labeled "overall quality" (*OQ*), as it has been included at the beginning of our questionnaires. Taking a single item response as a prediction target may however be disadvantageous, as single responses may contain more judgment-noise than averaged items. As an alternative, factor analyses of the subjective judgments have been carried out, and the three factors "overall acceptability" (*ACC*), "efficiency" (*EFF*) and "cognitive effort" (*COE*) have been used as prediction targets. The results for predictions of different target variables in the BoRIS and INSPIRE 1 experiments are listed in Table 5.

It can be seen that the prediction performance varies largely for the different target variables. The general tendency is that targets which are calculated as means from several judgments can be predicted with a higher correlation than the single questionnaire item *OQ*. R^2_{adj} follows r in this tendency. Interestingly, however, the E_p for the predictions of the perceptual dimensions is higher in some cases despite a higher r. Thus, it is not the reduction in spread caused by averaging across items that is responsible for the increased correlation.

The common finding across all evaluation metrics is that for both systems, using the L1O as well as the ALL method, the mean of all judgments (*AM*) could better be predicted than all other targets. Considering the perceptual dimensions, acceptability (*ACC*) seems to be better predictable (in terms of a higher R^2) than efficiency (*EFF*) and cognitive effort (*COE*), an exception being BoRIS predicted with the L1O method. It has to be noted that acceptability is the most important factor found in the BoRIS as well as the INSPIRE 1 judgments, and in both cases it was the factor with the highest number of correlated judgments. As the target values of the perceptual dimensions have been calculated by averaging the correlated judgments, *ACC* is more similar to *AM* than the factors *EFF* and *COE*. The similarity to *AM* might be the reason why we observed the dimension predictions getting better the more judgments are averaged in a target.

From the data, it is clear that *AM* provides an optimum prediction target. Therefore, only *AM* is retained as the prediction target for the subsequent analyses.

Table 5: Performance on training (ALL) and independent test data (L1O) with Set 1 input parameters, different target variables, and LR models.

Configuration			*Model performance*			
Experiment	*Output*	*Training*	r	R^2	R^2_{adj}	E_p
BoRIS	*AM*	ALL	0.712	0.507	0.471	0.881
BoRIS	*OQ*	ALL	0.475	0.226	0.208	0.960
BoRIS	*ACC*	ALL	0.676	0.457	0.425	0.869
BoRIS	*EFF*	ALL	0.571	0.326	0.302	0.962
BoRIS	*COE*	ALL	0.627	0.393	0.364	0.892
INSPIRE 1	*AM*	ALL	**0.749**	**0.561**	**0.525**	**0.672**
INSPIRE 1	*OQ*	ALL	0.583	0.340	0.308	0.820
INSPIRE 1	*ACC*	ALL	0.662	0.439	0.403	0.886
INSPIRE 1	*EFF*	ALL	0.574	0.329	0.287	0.910
INSPIRE 1	*COE*	ALL	0.513	0.263	0.228	0.948
BoRIS	*AM*	L1O	0.439	0.193	0.168	0.977
BoRIS	*OQ*	L1O	0.219	0.048	0.039	1.025
BoRIS	*ACC*	L1O	0.310	0.096	0.072	1.060
BoRIS	*EFF*	L1O	0.324	0.105	0.090	0.982
BoRIS	*COE*	L1O	0.400	0.160	0.143	0.948
INSPIRE 1	*AM*	L1O	0.480	0.230	0.172	0.990
INSPIRE 1	*OQ*	L1O	0.149	0.022	-0.02	1.091
INSPIRE 1	*ACC*	L1O	0.291	0.085	0.033	1.137
INSPIRE 1	*EFF*	L1O	0.093	0.009	0.046	1.083
INSPIRE 1	*COE*	L1O	0.121	0,015	0.102	1.113

4.4 Impact of the Modeling Algorithm

So far, only multivariate linear regression models were considered. A linear approach, however, presupposes that each parameter contributes in a "the-more-the-better" or "the-less-the-better" way to perceived quality. Obviously, this is not always true. For example, one can expect that the number of turns exchanged between user and system for reaching a specified goal has a non-zero optimum value, at least for a novice user: If the dialogue is too short, something might have gone wrong or some information might be missing, while a large number of utterances may be closely linked to problems occurring during the interaction. The most natural (in the sense of human-like) interaction will have an optimum value in-between these extremes.

In order to take such non-linearities into account, decision trees and neural networks have been considered as an alternative to LR models. The models have been calculated for the Set 1 input parameters, using *AM* as the prediction target, as this set and prediction target turned out in Sections 4.2 and 4.3 to be most promising. The results are summarized in Table 6. While linear regression (LR) and decision trees (CART, CHAID) yield comparable correlations in many cases, neural nets (MLP) achieve very high correlations

when predicting the same data they were trained on, but that fit may be restricted to that specific data set. For leave-1-out prediction (L1O), the predictive power of all classifiers decreases. Here again, neural networks appear to provide the best results. However, the risk of over-specialization still persists for neural networks in particular. For that reason we adhere to linear regression for the following analyses. The capability of neural networks to extrapolate to completely unseen data is further examined in Section 4.6.

Unpruned versions of decision trees obtain good results, but this is also most likely due to overfitting i.e. representing the data in an oversized tree which incorporates peculiarities of the specific data set. At the same time, the prediction quality does not increase after manual pruning for L1O testing.

Table 6: Performance on training (ALL) and independent test data (L1O) with different modeling algorithms, Set 1 input parameters and *AM* target variable. LR = Linear Regression, CART= Classification And Regression Trees, CHAID = Chi-squared Automatic Interaction Detection (tree), MLP = Multi-Layer Perceptron networks.

Configuration				*Model performance*		
Experiment	*Model*	*Note*	*Training*	r	R^2	E_p
BoRIS	LR	--	ALL	0.712	0.507	0.881
BoRIS	CART	unpruned	ALL	0.626	0.392	0.780
BoRIS	CART	pruned	ALL	0.439	0.193	0.898
BoRIS	CHAID	unpruned	ALL	0.571	0.326	0.821
BoRIS	MLP	--	ALL	0.922	0.849	0.157
INSPIRE 1	LR	--	ALL	0.749	0.561	0.672
INSPIRE 1	CART	unpruned	ALL	0.753	0.567	0.666
INSPIRE 1	CART	pruned	ALL	0.678	0.460	0.735
INSPIRE 1	CHAID	unpruned	ALL	0.578	0.334	0.827
INSPIRE 1	MLP	--	ALL	**0.981**	**0.962**	**0.041**
BoRIS	LR	--	L1O	0.439	0.193	0.977
BoRIS	CART	unpruned	L1O	0.276	0.076	1.038
BoRIS	CART	pruned	L1O	0.277	0.077	1.024
BoRIS	CHAID	unpruned	L1O	0.283	0.080	1.004
BoRIS	MLP	Set LR	L1O	0.443	0.197	0.804
BoRIS	MLP	Set ALL	L1O	0.500	0.250	0.753
INSPIRE 1	LR	--	L1O	0.480	0.230	0.991
INSPIRE 1	CART	unpruned	L1O	0.498	0.248	0.940
INSPIRE 1	CART	pruned	L1O	0.471	0.222	0.962
INSPIRE 1	CHAID	unpruned	L1O	0.072	0.005	1.129
INSPIRE 1	MLP	Set LR	L1O	0.628	0.394	0.617
INSPIRE 1	MLP	Set ALL	L1O	0.322	0.104	0.939

4.5 Impact of System Configuration

So far, the models have been tested on data collected with the same system and system version. In real-live use, however, the models should be able to perform predictions from one system version to another, e.g. when predicting the impact of a change in a system module.

In the BoRIS experiment, ten system versions were tested, differing with respect to the speech recognizer (simulated recognition performance), the speech output module (naturally-produced vs. synthesized speech), and the dialogue manager (confirmation strategy), see Table 1. For estimating the extrapolation performance in case of such system changes, we trained LR models for all system versions with one setting of these modules, and tested them on the remaining settings. The results are given in Table 7.

Table 7: Performance of LR models when extrapolating across BoRIS system versions. Set 1 input parameters, *AM* target variable.

Training			*Testing*		
System configuration	R^2	E_p	*System configuration*	R^2	E_p
All	0.507	0.881	All (L1O)	0.193	0.977
WA = 100% (ALL)	0.681	0.531	WA = 90% (ALL)	0.085	1.020
			WA = 80% (ALL)	0.312	0.938
			WA = 70% (ALL)	0.068	0.986
			WA = 60% (ALL)	**0.490**	0.929
Natural voice (ALL)	0.542	0.532	TTS (ALL)	*	1.661
			Mixed(ALL)	0.011	**0.798**
Confirmation (ALL)	0.621	0.540	No confirmation (ALL)	0.258	0.865

* R^2 has not been computed because the correlation is negative.

In accordance with our expectations, the prediction of test data resulted in poorer R^2s than calculated for the training data. However, it turned out that some predictions across system configurations were more accurate than those calculated with the L1O method on the complete data set. Surprisingly, relatively high R^2s were achieved when predicting the rating of dialogs with either 60% or 80% WA, while for WA = 70% or WA = 90% the R^2s are very low. There is no reasonable explanation for this finding except that the training as well as the calculation of R^2 was performed on only 20 cases each. This number is too small to guarantee a valid model (in this case, only the percentage of partially correctly parsed user utterances was taken into the model as a predictor) and the correlation coefficient is more sensitive to small changes in the data structure. Considering E_p of the predictions, it can be noted that they follow the general tendency of the R^2s, but the values do not differ as clearly.

A further high R^2 was found for the prediction of dialogs without confirmation of acquired information with an equation trained on dialogs with explicit confirmation. In contrast to that, worst results were obtained when predicting dialogs with different system voices. Obviously, the effect of the system voice is not covered by the interaction

parameters which form the input for quality prediction; in turn, the effect of a confirmation strategy might well be reflected by the interaction parameter values, and can thus be taken into account by the model.

We also tried to predict judgments of a newer system version with an equation trained on data from an earlier experiment. For this, we took advantage of the INSPIRE system having been tested at two different points in its development cycle. Between the first and the second test, the system was mainly improved with respect to the vocabulary and the speech understanding component, as well as the system prompt wording. Table 8 provides performance values when different models are trained on one of the experimental databases and tested on the other one.

Table 8: Performance of different models when extrapolating across INSPIRE experiments. Set 1 input parameters, *AM* target variable. LR = Linear Regression, CART= Classification And Regression Trees, CHAID = Chi-squared Automatic Interaction Detection (tree), MLP = Multi-Layer Perceptron networks.

Training				*Testing*		
Experiment	*Model*	R^2	E_p	*Experiment*	R^2	E_p
INSPIRE 1 Set 1	LR	0.561	0.672	INSPIRE 1 (L1O)	0.230	0.991
				INSPIRE 2 (ALL)	0.026	1.126
	CART, unpruned	0.567	0.666	INSPIRE 1 (L1O)	0.248	0.940
				INSPIRE 2 (ALL)	0.145	1,026
	CART, pruned	0.460	0.735	INSPIRE 1 (L1O)	0.221	0.962
				INSPIRE 2 (ALL)	0.141	0.981
	CHAID	**0.578**	0.827	INSPIRE 1 (L1O)	0.005	1.129
				INSPIRE 2 (ALL)	0.220	0.895
	MLP, Set LR (L1O)	0.394	0.617	INSPIRE 2 (ALL)	0.001	1.237
	MLP, Set ALL (L1O)	0.104	0.939	INSPIRE 2 (ALL)	0.174	**0.832**
INSPIRE 1 Set 2a parameters	LR	0.496	**0.518**	INSPIRE 1 (L1O)	0.225	0.936
				INSPIRE 2 (ALL)	0.008	
INSPIRE 1 Set 2b parameters	LR	0.567	0.693	INSPIRE 1 (L1O)	**0.332**	0.853
				INSPIRE 2 (ALL)	0.098	

Considering the Set 1 input parameters, the prediction performance decreases significantly when moving from one system version to the other. Best predictions for INSPIRE 2 data were obtained with the CHAID trees and the MLP trained on the parameter set from the ALL method. However, in both CHAID and MLP cases, prediction accuracy on INSPIRE 2 data was better than on INSPIRE 1 data with the L1O method. This indicates that the good values are somewhat accidental, as generalizability

towards another system (configuration) can hardly be guaranteed without validity of the model on the corpus it was trained on.

For the remaining models, the R^2s do not reach a level which would justify the model to be used for predictions. This finding is in opposition to what was found by Walker et al. (2000), who could predict test cases from other systems equally well as the training data themselves, reaching R^2s up to 0.55 on the test cases. We became suspicious that the big amount of input parameters of Set 1 might result in too specialized models and therefore repeated the procedure with the same method as utilized by Walker et al. (2000), which corresponds to LR modeling with Set 2 input parameters. As can be seen in the last two rows of Table 8, we found that the poor generalizability of our models is not due to the enriched parameter Set 1. Even if a subjective judgment on task success is included in the model (Set 2b), prediction accuracy on test data from another system (configuration) is considerably lower than the one reported by Walker or achieved by us on the test data. Thus, these values seem to represent the limit of cross-configuration prediction for the INSPIRE experiments, and not a bias specific for the classifier or input parameter set.

It should be noted that – besides the system version – also the participant group changed between INSPIRE 1 and 2. As the two system versions have not been tested with the same participant group, it cannot be decided whether the system changes or the changes in the user group affect the prediction results most.

4.6 Cross-system Prediction

An ideal prediction model would be applicable to a large number of SDSs. As there are three databases available which have been annotated according to the same principle, a cross-system prediction – from BoRIS to INSPIRE and vice-versa – becomes possible. However, as it has already been observed in the previous section, the data has been obtained from different test participants, casting doubt on whether any differences stem from the tested system or the participant group. Table 9 summarizes results for models trained on the INSPIRE 1 and the BoRIS datasets and applied to all other data.

Table 9: Performance of LR models when extrapolating across systems, *AM* target variable.

Training				*Testing*		
Experiment	*Input*	R^2	E_p	*Experiment*	R^2	E_p
INSPIRE 1	Set 1	0.561	0.672	INSPIRE 1 (L1O)	0.230	0.991
				BoRIS	0.029	1.134
	Set 2a	0.496	0.720	INSPIRE 1 (L1O)	0.225	0.936
				BoRIS (ALL)	0.077	1.073
	Set 2b	**0.567**	**0.666**	INSPIRE 1 (L1O)	0.332	0.853
				BoRIS (ALL)	0.132	1.013
	Set 3	0.381	0.799	INSPIRE 1 (L1O)	0.090	1.002
				BoRIS (ALL)	0.032	1.292
BoRIS	Set 1	0.507	0.881	BoRIS (L1O)	0.193	0.977

Training				*Testing*		
Experiment	*Input*	R^2	E_p	*Experiment*	R^2	E_p
				INSPIRE 1 (ALL)	0.004	1.810
				INSPIRE 2 (ALL)	0.092	1.687
	Set 2a	0.219	0.892	BoRIS (L1O)	0.170	0.913
				INSPIRE 1 (ALL)	0.120	0.970
				INSPIRE 2 (ALL)	0.263	0.879
	Set 2b	0.440	0.758	BoRIS (L1O)	**0.406**	**0.775**
				INSPIRE 1 (ALL)	0.321	0.843
				INSPIRE 2 (ALL)	0.377	0.821
	Set 3	0.259	0.861	BoRIS (L1O)	0.214	0.888
				INSPIRE 1 (ALL)	0.072	1.021
				INSPIRE 2 (ALL)	0.116	0.964

In all cases, the prediction performance decreases significantly when moving from one system to the other. The decrease is slightly less strong when using Set 2b input parameters, i.e. the set including the subjective judgment on task success. Nevertheless, it can be stated that a cross-system extrapolation is not possible with the existing models.

5. Discussion

In the previous section, we analyzed different model configurations with respect to their performance in describing both known training and unknown test data. Test data was either taken from the same experiments, using a leave-one-user-out approach, or from a different system configuration or system (i.e. experiment).

The performance of our baseline model on the training data fulfilled or even surpassed the expectations set by the literature (e.g. R^2 = 0.41 in Kamm et al., 1998; R^2 = 0.47 in Walker et al, 1998; R^2 = 0.39…0.56 in Walker et al., 2000; R^2 = 0.51 in Larsen, 2003), as long as a subjective judgment on task success is included in the input parameter set. This approach – although frequently followed in the literature – is however prohibitive if real prediction of user judgment is of interest. Such predictions are important for operating services when interactions can be logged easily, but when service operators are not willing to ask their customers about their impression. Therefore, we extended the set of input parameters and reached a similar performance (R^2 = 0.30…0.58) even when omitting subjective judgments at the input to the models. Either a large set of standard interaction parameters (e.g. the ones recommended in ITU-T Suppl. 24 to P-Series Rec., 2005) or a combination of interaction parameters and error frequencies turned out to be best.

The prediction performance on unseen training data is in all cases significantly lower than the one on the training data. This is not in accordance with the findings of Walker et al. (2000), who were able to predict independent test data (even taken from other experiments) with a similar accuracy as the training data. However, the systems examined in this study came in several configurations, differing significantly in the

(simulated) recognition performance, speech output, and/or in the metaphor represented to the user. Thus, we expected a lower performance on the testing data. Still, whereas the general tendency of this finding was expected, the amount of the decrease was not: R^2 dropped to 0.09…0.23 when omitting subjective judgments at the input to the model. This shows that the models have significant problems in predicting unseen data.

The situation did not change when using different target variables, e.g. a single-item overall quality judgment, or perceptual quality dimensions derived by a factor analysis. Other model algorithms have been investigated which do not presuppose a linear relationship between input and output variables, namely Classification And Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID), and Multi-Layer Perceptron networks (MLP). The results were very similar to the ones obtained with the linear regression models.

When changing the system configuration (e.g. individual system modules in the BoRIS system, or developmental changes of the INSPIRE system), prediction accuracy decreased once again. This seems to be a significant limitation, because prediction models often are meant to provide estimations of perceived quality and usability for unknown system configurations. The ultimate aim, namely to provide valid and reliable predictions for new systems, seems to be far out of reach of the modeling approaches used here.

The poor extrapolation performance of our models compared to the ones given in Walker et al. (2000) may have several underlying reasons. Firstly, our models did not contain any user judgment on task success. The absence of a user-derived predictor – which seems to be well correlated with the to-be-predicted quality judgment – may have a detrimental effect on the generalizability of a model. Apparently, the larger set of input parameters could not provide the same *general* information as user-derived parameters can. This result however cannot be generalized: there might be other interaction parameters capturing the same type of general information as user judgments do, but we simply did not find them. Secondly, because our models could select from a larger set on input parameters (Set 1 or Set 4), they may have become more tailored to the system they have been trained on. However, the cross-system extrapolation results with the Set 2a and Set 2b input parameters in Table 8 show that this was not decisive in our case.

6. Application Examples

Nevertheless, we decided to reconsider the aim of our investigation: In order to estimate the usability of a system, it is not necessary to provide valid and reliable estimations of quality judgments for *individual interactions* or *individual users*. On the contrary, system developers optimize the system for a large group of potential users, taking into account mean values or the distribution of judgments. Thus, it may be sufficient to be able to estimate an average user judgment for a set of dialogues, e.g. in order to take decisions in the system design process.

An analysis of scatter-plots showed that indeed there is a relationship between some parameters and overall user judgments, however not in the form of a 1-to-1 mapping: While for extremely “poor” (i.e. very high or very low) values of a parameter the

judgment can often be predicted to be negative, for average and "good" values of the parameter judgments spread across a broad range. Despite this large spread, there seems to be an interpretable relationship between several parameters and *overall* user judgments. Thus, it may be possible to estimate the average ratings for different system configurations.

We investigated the capability of our prediction models to estimate average judgments taking a typical system design problem. A critical point in the development of most SDSs is the question of whether the recognition performance of the system is good enough to allow for high-quality interactions. Because we varied the recognition performance in the INSPIRE 2 experiment in a controlled way, we are able to compare the subjective and the predicted averaged quality judgments for the system configurations differing in (simulated) ASR performance.

Figure 1 shows the comparison between the subjective data and the predictions with the ALL and L1O method. The predictions have been multiplied by the ratings' standard deviation, and the mean value of the ratings was added, in order to reverse the z-transformation performed on the target variable before the training. While the distributions of the predictions are considerably narrower than in the subjective data, the medians show the same relation to the ASR performance. Therefore, for decisions made on the basis of the median, the predictions can be of use. In our example, the developer might draw the same conclusions from the predictions as from the subjective data, namely that ASR performance matters for the perceived quality, however, even with 100% recognition accuracy the system is rated only average. She could further read from the graph, that the system was judged better than average (3.0) only for recognition rates higher than 86%.

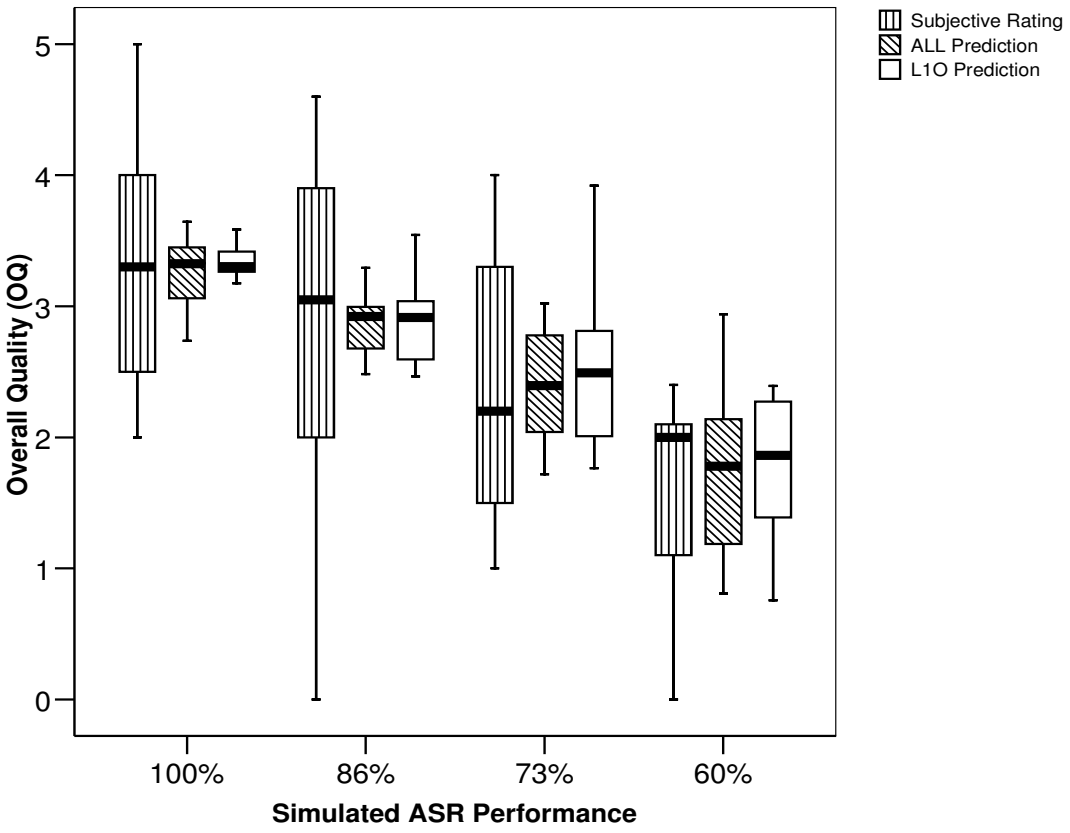


Figure 1: Subjective and predicted *OQ* values for the INSPIRE 2 experiment, LR prediction model with ALL and with L1O training, Set 1 input parameters. Indicated are the median (thick line), the 25-75% quartile, and the 5-95% range.

It is particularly remarkable that the quality of the predictions does not decrease for predictions from unseen test data as compared to predictions of the training data. In the figure, medians predicted with the L1O method are very similar to those predicted with the ALL method, and less accurate only in case of the 73% median. Generalizability, which is a key requirement for practical applications of the prediction model, seems to be sufficient for the application shown here.

7. Final Conclusions and Future Work

The prediction performance which we determined in our experiments is significantly lower that the one known from other application areas. For examples, the PESQ model recommended in ITU-T Rec. P.862 (2001) for predicting the quality of telephone-transmitted speech on the basis of a comparison of two signals showed a correlation with subjective judgments (single-item judgments as for *OQ*, but obtained on a slightly different scale) of 0.935 on 22 training databases, and the same correlation on 8 unknown test databases (Rix et al., 2006). The correlation for the P.563 model on predicting speech transmission quality on the basis of an individual degraded signal – a more difficult task – is still 0.88 on 24 training databases (Rix et al., 2006). The same model may be used to estimate the quality of synthesized speech, by automatically generating a reference signal with the help of a vocal tract analysis and integrating several perceptual dimensions to form an overall quality judgment. For this task, correlations range between 0.59 and 0.74 on two test databases which are definitely outside the original scope of the model (Möller and Heimansberg, 2006).

In this study, the *training* correlations of our baseline LR model are between 0.47 and 0.70 for BoRIS and between 0.62 and 0.75 for INSPIRE 1, when omitting subjective ratings at the input to the model. Correlations for the unknown *test* data were even lower with between 0.41 and 0.46 for BoRIS, and between 0.30 and 0.48 for INSPIRE 1. In other terms, we can cover approximately half of the variance in the training data, and a significantly lower amount of the test data variance.

Still, the models seem to do a good job in estimating mean overall quality judgments instead of ratings for individual users and dialogues. Such mean values may be taken as decision criteria in the system design process. For example, we showed that the decision for the minimum required word accuracy in order to reach an acceptable level of system performance can be determined with the help of such models, i.e. without directly asking test participants about their opinion. In this way, it is possible to just log interactions – carried out e.g. with remote participants – and not require participants to come to a test lab, or to answer surveys.

In order to improve the prediction accuracy of the presented models, it may be necessary to identify better – i.e. more informative – input parameters. For example, no parameters are yet available for describing speech output quality. First approaches are described in Möller and Heimansberg (2006) and in ITU-T Contr. COM 12-47 (2007), but they are not yet satisfying, in the sense of a too low correlation (see above). Further information may be derived from the user's speech signal, e.g. with respect to his/her emotional state (which may be linked to the speech level, or to prosodic features), or from system characteristics (e.g. a confusable vocabulary or grammar may lead to more misunderstandings in the dialogue). Work in this direction is underway to provide additional input parameters for PARADISE-style or other types of models.

From the described experiments, it is not yet clear whether a larger improvement of prediction performance can be achieved by selecting better input parameters, or by using better modeling algorithms. Non-linear models, including the ones used here, may incorporate relationships between input parameters. In addition, threshold effects can be modeled, in the sense that a certain parameter matters for overall quality if it is above a

threshold, but that it does not matter any more if it is below that threshold. Unfortunately, it is difficult to design experiments exploring such relationships, because the parameter values are largely influenced by the (spontaneous) behavior of the user.

In a similar vein, the choice of an appropriate target variable has also to be considered as a potential way of improving prediction models and their utility for system design. We tried to account for this from a modeling point-of-view, see Section 4.3. The results obtained here suggest that the mean of all subject ratings is superior to predicting a single overall quality rating or perceptual dimensions underlying the corresponding questionnaires. Still, other combinations, e.g. a weighted sum of individual answers, are conceivable. From a system design perspective, it has to be checked whether the target variables we proposed are informative for the purpose of the evaluation; this can be decided only in a particular evaluation situation.

In the future, quality prediction models may form a part of (semi-) automatic system development tools. With the help of such tools, a system developer is able to estimate the quality and usability of his system during the development process without a direct involvement of human test participants. For example, user behavior may be modeled to simulate interactions between user and system, as described in the MeMo workbench (Möller et al., 2006). Prediction models like the ones described here can be used in conjunction with such tools, as long as they provide valid and reliable estimations of average (not necessarily individual) user judgments.

Acknowledgement

The described work has been carried out at Deutsche Telekom Laboratories (T-Labs), Berlin University of Technology, based on data which has been collected at Ruhr-University Bochum in the frame of the first author's habilitation thesis and the EU-funded IST project INSPIRE (IST 2001-32746). It is partly related to a diploma thesis carried out by Bernardo Compagnoni in conjunction with IfN, TU Braunschweig (Tim Fingscheidt) referenced in the text. The authors would like to thank all colleagues at T-Labs for their valuable discussions, Tim Fingscheidt and Bernardo Compagnoni for their impact on the described work, as well as two anonymous reviewers for helpful comments on an earlier version of this paper.

References

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

Compagnoni, B., 2006. Development of Prediction Models for the Quality of Spoken Dialogue Systems. Diploma thesis (unpublished), Deutsche Telekom Laboratories, TU Berlin / Institut f. Nachrichtentechnik, TU Braunschweig.

Constantinides, P.C., Rudnicky, A.I., 1999. Dialogue Analysis in the Carnegie Mellon Communicator, in : Proc. 6th Europ. Conf. on Speech Communication and Technology (Eurospeech'99), Budapest, 1, 243-246.

Engelbrecht, K.-P., 2006. Fehlerklassifikation und Benutzbarkeits-Vorhersage für Sprachdialogdienste auf der Basis von mentalen Modellen (Error Classification and Usability Prediction for Spoken Dialogue Services on the Basis of Mental Models). Magister thesis (unpublished), Deutsche Telekom Laboratories, TU Berlin.

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

Hastie, H.W., Prasad, R., Walter, M., 2002. Automatic Evaluation: Using a DATE Dialogue Act Tagger for User Satisfaction and Task Completion Prediction, in: Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), Las Palmas, 2, 641-648.

Hone, K.S., Graham, R., 2000. Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4), 287-303.

Hone, K.S., Graham, R., 2001. Subjective Assessment of Speech-system Interface Usability, in : Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia), Aalborg, 3, 2083-2086.

ISO 9241-11, 1998. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 11: Guidance on Usability. International Organization for Standardization, Geneva.

ITU-T Contribution COM 12-47, 2007. Quality Estimation for Transmitted Synthesized Speech with Single-Ended Models: Comparison of Step 1 Results, Federal Republic of Germany, Alcatel-Lucent, Psytechnics (Authors: S. Möller, D.-S. Kim, L. Malfait and B. Kleijn), ITU-T SG12 Meeting, 16-25 Jan. 2007, Geneva.

ITU-T Recommendation P.851, 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.

ITU-T Recommendation P.862, 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva.

ITU-T Supplement 24 to P-Series Recommendations, 2005. Parameters Describing the Interaction with Spoken Dialogue Systems. International Telecommunication Union, Geneva.

Jekosch, U., 2005. Voice and Speech Quality Perception. Assessment and Evaluation. Springer, Berlin.

Kamm, C.A., Litman, D.J., Walker, M.A., 1998. From Novice to Expert: The Effect of Tutorials on User Expertise with Spoken Dialogue Systems, in: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98), Sydney, 4, 1211-1214.

Larsen, L.B. (2003). Issues in the Evaluation of Spoken Dialogue Systems Using Objective and Subjective Measures, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03), 209-214.

MacKay, D.J.C., 1992. Bayesian Interpolation. Neural Computation 4(3), 415-447.

Möller, S., 2005. Quality of Telephone-based Spoken Dialogue Systems. Springer, New York NY.

Möller, S., 2005b. 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.

Möller, S., 2005c. Perceptual Quality Dimensions of Spoken Dialogue Systems: A Review and New Experimental Results, in: Proc. 4th European Congress on Acoustics (Forum Acusticum Budapest 2005), Budapest, 2681-2686.

Möller, S., Krebber, J., Smeele, P., 2005d. Evaluating the Speech Output Component of a Smart-Home System. Speech Communication 48, 1-27.

Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N., 2006. MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations, in: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh PA, 1786-1789.

Möller, S., Heimansberg, J., 2006. Estimation of TTS Quality in Telephone Environments Using a Reference-free Quality Prediction Model, in: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 56-60.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating Spoken Dialogue Systems According to De-facto Standards: A Case Study. Computer Speech and Language 21, 26-53.

Oulasvirta, A., Möller, S., Engelbrecht, K., Jameson, A., 2006. The Relationship of User Errors to Perceived Usability of a Spoken Dialogue System, in: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 61-67.

Rajman, M., Rajman, A., Seydoux, F., Trutnev, A., 2003. Assessing the Usability of a Dialogue Management System Designed in the Framework of a Rapid Dialogue Prototyping Methodology, in: Proc. ISCA Tutorial and Research Workshop on Auditory Quality of Systems (AQS 2003), Mont Cenis, 126–133.

Rix, A.W., Beeerends, J.G., Kim, D.-S., Kroon, P., Ghitza, O., 2006. Objective Assessment of Speech and Audio Quality – Technology and Applications. IEEE Trans. Audio, Speech, Lang. Process. 14, 1890–1901.

Simpson, A., Fraser, N.M., 1993. Black Box and Glass Box Evaluation of the SUNDIAL System, in: Proc. 3rd European Conf. on Speech Communication and Technology (Eurospeech'93), Berlin, 2, 1423-1426.

Sutton, S., Cole, R., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, M., Cohen, M., 1998. Universal Speech Tools: The CSLU Toolkit, in: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98), Sydney, 7, 3221–3224.

Trutnev, A., Ronzenknop, A., Rajman, M., 2004. Speech Recognition Simulation and its Application for Wizard-of-Oz Experiments, in: Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004), Lisbon, 611-614.

Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A., 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents, in: Proc. ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics, Madrid, 271-280.

Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A., 1998. Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. Computer Speech and Language 12, 317-147.

Walker, M., Kamm, C., Litman, D., 2000. Towards Developing General Models of Usability with PARADISE. Natural Language Engineering 6, 363-377.

Walker, M., Kamm, C., Boland, J., 2000b. Developing and Testing General Models of Spoken Dialogue System Performance, in: Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC 2000), Athens, 1, 189-196.