Quality prediction models for telephone-based spoken dialogue systems

Sebastian Möller

IKA, Ruhr-Universität, D-44780 Bochum, Germany, Email: sebastian.moeller@ruhr-uni-bochum.de

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 synthesised 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.

In order to estimate the overall quality experienced by the user, quality prediction models have been set up. They calculate an index related to user satisfaction on the basis of parameters which can be logged during the interaction. These *interaction parameters* can be extracted either instrumentally (e.g. dialogue duration, number of system and user turns, response delay), or they require a transcription and annotation by a human expert (e.g. word error rate, classification of utterances according to pre-defined categories). The most popular modelling approach is the PARADISE model developed by Walker et al. [1]. PARADISE uses a weighted linear combination of a number of interaction parameters to predict the mean value of several *user judgements*. The modelling function and the weighting coefficients once being determined in a controlled laboratory experiment with human test subjects, it is possible to use the prediction model for optimising the system and its components, in the sense of reaching optimum (predicted) user satisfaction, and not just optimum performance of individual components.

In this paper, experimental results are presented which illustrate the modelling principle as well as the prediction accuracy which can be reached. Starting from a taxonomy of quality aspects which have demonstrated relevance for telephone-based spoken dialogue systems, it is shown that the PARADISE approach is still limited in predicting global quality aspects like overall user satisfaction. Better results may be obtained for limited quality aspects, depending on the information which is available with the input parameters.

Quality Aspects and Prediction Models

Despite the efforts made to predict quality from instrumentally or expert-derived parameters, the quality of an interaction with a spoken dialogue system depends on the perception of the user. Quality is the result of a perception and a judgement process, in which the perceiving subject establishes a relationship between the perceptive event, and what he or she expects or desires from the interaction [2]. The perceptions of the user are of highly multidimensional nature; thus, it is difficult to summarise them under the global label "overall quality" or "user satisfaction".

In an earlier publication [3], the most relevant *quality aspects* (i.e. nameable components of quality) have been organised with the help of a taxonomy. The taxonomy identifies the characteristics of the system, of the task the system has been designed for, of the physical environment the system is used in (transmission channel, background noise, room acoustics, etc.), of the non-physical context of use (costs, availability, opening hours), and of the user (linguistic background, experience, motivation, attitude, etc.) which are relevant to quality. The taxonomy shows how these characteristics are related to quality aspects such as speech input and output quality, dialogue cooperativity, dialogue symmetry, communication efficiency, comfort, task efficiency, usability, user satisfaction, and finally to the acceptability of the system.

From the taxonomy, it is obvious that "user satisfaction" can hardly be interpreted as an average value of user judgements on characteristics like intelligibility, perceived system understanding, task ease, interaction pace, or an expected future use of the system. Nevertheless, the mean value of such judgements forms the target (output) variable of the PARADISE model. As an input, the model combines normalised interaction parameters describing "dialogue costs" (number of utterances, dialogue duration, word error rate, etc.) and task success (expressed as the κ coefficient, see [1]):

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

with US the estimated "user satisfaction" index, κ the task success coefficient, c_i the interaction parameters related to "dialogue costs", α and w_i the weighting coefficients, and N the Z-score normalisation function. The coefficients α and w_i are determined by a multivariate linear regression analysis for a set of data obtained in a controlled laboratory experiment. Once these coefficients have been determined, the model can be used to estimate US from unseen interaction parameter values. The amount of variance in the subjective data which is covered by the model's prediction is expressed by the R^2 value of the regression analysis.

Experimental Results

The PARADISE model has been applied to experimental data obtained with a telephone-based system for restaurant information [4][5]. The system has been tested in various configurations, differing with respect to the recognition rate (simulated by a transcribing expert), the confirmation strategy, and the speech output (natural vs. synthesised speech).

40 test subjects (11 f, 29 m, 23-51 years) interacted with five different system configurations. The interactions have been logged and annotated by an expert. In this way, about 40 interaction parameters per dialogue could be extracted. After each dialogue, the test subjects rated 26 statements related to

different quality aspects (part B of the questionnaire). The test was closed by an additional questionnaire (part C) with 18 questions reflecting the overall experience with the dialogue system gained during the experiment. Detailed test results are discussed in [4][5][6].

On the basis of the subjective ratings and the interaction parameters, several quality prediction models have been derived. The models all follow the linear combination approach described by equation (1). They differ, however, with respect to the target variable reflecting quality (*US*) and the interaction parameters used as the input (κ and c_i). *US* has been calculated either as the mean value over 9 subjective ratings as described by Walker et al. (US_w), as the mean over all subjective ratings of part B of the questionnaire (Mean(B)), or as a rating on statement B23 ("Overall, you are satisfied with the dialogue"). As an input, either the full set of 40 interaction parameters (set 2) or the limited set used by Walker et al. (set 1, containing 4 parameters) was selected. In addition, either the κ coefficient or a subjective rating reflecting task success (question B1, similar to the rating replacing κ in [1]) were used.

Input parameters		Target variable	R^2 (# input parameters)
Dialogue cost	Task success		
set 1	κ	US_w	0.14 (5)
set 1	B1	US_w	0.25 (5)
set 1	B1	B23	0.48 (5)
set 1	B1	Mean(B)	0.52 (5)
set 2	κ	US_w	0.41 (9)
set 2	B1	US_w	0.35 (4)
set 2	B1	B23	0.46 (4)
set 2	B1	Mean(B)	0.59 (5)

Table 1: Prediction models for target variables related to "user satisfaction"

Table 1 shows the percentage of variance in the subjective judgements which can be covered by the models. In most cases, R^2 does not exceed 0.5, showing that about half of the influencing factors on user satisfaction are not yet covered. R^2 depends on the available input parameters (both for "dialogue costs" and task success) and on the target variable. It has to be noted that the regression algorithm selects a different number of input variables to be included in equation (1), cf. the last column of Table 1; for set 1, a forced-inclusion of all 5 variables is used, for set 2 a stepwise (forward-backward) inclusion method.

Apart from global estimates of "user satisfaction", it is possible to predict subjective ratings which are related to individual quality aspects of the described taxonomy. Table 2 shows some examples in this respect. As an input, the set 2 interaction parameters and expert-based parameters related to task success (no subjective ratings!) were used. The target variable is the mean of all subjective ratings addressing a specific quality aspect, selected intuitively with the help of the taxonomy. It can be seen that the model coverage is worst for global quality aspects like usability, user satisfaction and acceptability; better results are obtained for the lower-level aspects, except for speech input/output quality. The latter finding will mainly be due to the absence of speech-output-related interaction parameters.

Target variable	R^2 (# input parameters)
Speech input/output quality	0.25 (6)
Dialogue cooperativity	0.42 (7)
Dialogue symmetry	0.31 (6)
Communication efficiency	0.51 (10)
Comfort	0.41 (9)
Task efficiency	0.40 (9)
Usability	0.12 (4)
User satisfaction	0.28 (7)
Acceptability	0.12 (4)

Table 2: Prediction models for individual quality aspects

Discussion and Conclusions

The results show that prediction models for the interaction with spoken dialogue systems miss about half of the characteristics which are relevant for the quality from a user's point of view (more precisely: half of the variance in the judgements). It is expected that information is still missing in the interaction parameters which are used as an input to the models. On the other hand, prediction success largely depends on the predicted target variable. By simply averaging subjective ratings related to different quality aspects, it is assumed that each addressed aspect has the same importance for the user. Both theoretical and empirical evidence for this assumption is missing. Due to the lack of independent test data, the models have been tested on the training data; a leave-one-test-subject-out experiment suggests that the amount of covered variance is even lower for unseen data.

Acknowledgement

The study was carried out at IKA, Ruhr-Universität Bochum (PD U. Jekosch, Prof. R. Martin), and supported by the EC project INSPIRE (IST-2001-32746). The author would like to thank Janto Skowronek for his support in the experiment.

References

[1] Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A. (1997). *PARADISE: A framework for evaluating spoken dialogue agents*, in: Proc. of the ACL/EACL 35th Ann. Meeting, ES-Madrid, 271-280.

[2] Jekosch, U. (2000). *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, Habilitation thesis, Universität/GH, D-Essen.

[3] Möller, S. (2002). *A new taxonomy for the quality of telephone services based on spoken dialogue systems*, in: Proc. 3rd SIGdial Workshop on Discourse and Dialogue, USA-Philadelphia PA, 142-153.

[4] Skowronek, J. (2002). *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*, Diploma thesis, IKA, Ruhr-Universität, D-Bochum.

[5] Möller, S. (2003). *Quality of telephone-based spoken dialogue systems*, Habilitation thesis (submitted), Ruhr-Universität, D-Bochum.

[6] Möller, S., Skowronek, J. (2003). *Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue system*, in: Proc. EUROSPEECH 2003, CH-Geneva, Vol. 3, 1953-1956.