Modeling User Satisfaction with Hidden Markov Models

Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, Sebastian Möller
Deutsche Telekom Laboratories, Quality & Usability Lab, TU Berlin,
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{Klaus-Peter.Engelbrecht,Florian.Goedde,
Hamed.Ketabdar,Sebastian.Moeller}@telekom.de
Felix.Hartard@Berlin.de

Abstract

Models for predicting judgments about the quality of Spoken Dialog Systems have been used as overall evaluation metric or as optimization functions in adaptive systems. We describe a new approach to such models, using Hidden Markov Models (HMMs). The user's opinion is regarded as a continuous process evolving over time. We present the data collection method and results achieved with the HMM model.

1 Introduction

Spoken Dialog Systems (SDSs) are now widely used, and are becoming more complex as a result of the increased solidity of advanced techniques, mainly in the realm of natural language understanding (Steimel et al. 2008). At the same time, the evaluation of such systems increasingly demands for testing the entire system, as components for speech recognition, language understanding and dialog management are interacting more deeply. For example, the system might search for web content on the basis of meaning extracted from an n-best list, and generate the reply and speech recognition grammars depending on the content found (Wootton et al. 2007). The performance of single components strongly depends on each other component in this case.

While performance parameters become less meaningful in such a system, the system's overall quality, which can only be measured by asking the user (Jekosch 2005), gains interest for the evaluation. Typically, users fill out questionnaires after the interaction, which cover various perceptional dimensions such as efficiency, dialog smoothness, or the overall evaluation of the system (Hone and Graham, 2001; ITU-T Rec. P.851, 2003; Möller 2005a). Judgments of the system's overall quality can be used to compare systems with respect to a single measure, which however comprises all relevant aspects of the interaction. Thus, the complexity of the evaluation task is reduced.

In addition, user simulation is increasingly used to address the difficulty of foreseeing all possible problems a user might encounter with the system (e.g. Ai and Weng, 2008; Engelbrecht et al., 2008a; Chung, 2004; López-Cózar et al., 2003). In order to evaluate results from such simulations, some approaches utilize prediction models of user judgments (e.g. Ai and Weng, 2008; Engelbrecht et al., 2008a).

Currently, prediction models for user judgments are based on the PARADISE framework introduced by Walker et al. (1997). PARADISE assumes that user satisfaction judgments describe the overall quality of the system, and are causally related to task success and dialog costs, i.e. efficiency and quality of the dialog. Therefore, a linear regression function can be trained with interaction parameters describing dialog costs and task success as predictors, and satisfaction ratings as the target. The resulting equation can then be used to predict user satisfaction with unseen dialogs.

In follow-up studies, it could be shown that such models are to some degree generalizable (Walker et al., 2000). However, also limitations of the models in predicting judgments for other user groups, or for systems with different levels of ASR performance, were reported (Walker et al., 1998). In the same study, prediction

functions for user satisfaction were proposed to serve as optimization function in a system adapting its dialog strategy during the interaction. This idea is taken up by Rieser and Lemon (2008).

The prediction accuracy of PARADISE functions typically lies around an R^2 of 0.5, meaning that 50% of the variance in the judgments is explained by the model. While this number is not absolutely satisfying, it could be shown that mean values for groups of dialogs (e.g. with a specific system configuration) can be predicted more accurately than single dialogs with the same models (Engelbrecht and Möller, 2007). Low R^2 for the predictions of ratings of individual dialogs seems to be due to inter-rater differences at least to some degree. Such differences have been described, and may concern the actual perception of the judged issue (Guski, 1999), or the way the perception is described by the participant (Okun and Weir, 1990; Engelbrecht et al., 2008b)

We have tested the PARADISE framework extensively, using different classifier models and interaction parameters. Precise and general models are hard to achieve, even if the set of parameters describing the interaction is widely extended (Möller et al., 2008). In an effort to improve such prediction models, we developed two ideas:

- Predict the distribution of ratings which can be expected for a representative group of users given the same stimulus. This takes into account that in most cases the relevant user characteristics determining the judgment cannot be tracked, or even are unknown.
- Consider the time relations between events by modeling user opinion as a variable evolving over the course of the dialog. This way, time relations like co-occurrence of events, which affect quality perception, attention, or memory can be modeled most effectively.

In this paper, we present a new modeling approach considering these ideas. In Section 2, we introduce the topology of the model. Following this, we report how training data for the model were obtained from user tests in Section 3. Evaluation results are presented in Section 4 and discussed in Section 5, before we conclude with some remarks on follow-up research.

2 Modeling Judgments with HMMs

Hidden Markov Models (HMMs) are often used for classifying sequential stochastic processes, e.g. in computational linguistics or bio-informatics. An HMM models a sequence of events as a sequence of states, in which each state emits certain symbols with some probability. In addition, the transitions between states are probabilistic. The model is defined by a set of state symbols, a set of emission symbols, the probabilities for the initial state, the state transition matrix, and the emission matrix. The transition matrix contains the probabilities for transitions from each state to each other state or itself. The emission matrix contains the probabilities for each emission symbol to occur at each state.

While the sequence of emissions can be observed, the state sequence is hidden. However, given an emission sequence, standard algorithms defined for the HMM allow to calculate the probability of each state at each point in the sequence. The probability for the model to be in a state is dependent on the previous state and the emissions observed at the current state.

As illustrated by Figure 1, the development of the users' opinions can be modelled as an HMM. The user judgment about the dialog is modelled as states, each state representing a specific judgment (think of it as "emotional states"). A prediction is made at each dialog turn. In the model depicted, the user judgment can either be "bad" or "good". Each judgment has a probabilistic relation to the current events in the dialog. In the picture, the events are described in the form of understanding errors and confirmation types, i.e. there are two features which can take a number of different values, each with a certain probability.

Although the judgments do not "emit" the events at each turn (the causal relation is opposite), the probabilistic relation between them can be captured and evaluated with the HMM and the associated algorithms.

Apart from the dialog events, the current judgment is also determined by the previous judgment. For example, we expect that the judgments are varying smoothly, i.e. the probability for a transition becomes lower with increasing (semantic) distance between the state labels.

Although events in previous turns cannot impact the current judgment given this model topology, it is possible to incorporate dialog

history by creating features with a time lag. E.g., a feature could represent the understanding error in the previous turn. Also, simultaneity of different events affecting the quality perception can be evaluated by calculating probabilities for each judgment given the observed combination of features. If the features are interacting (i.e. the probability of one feature changes in dependence of another feature), this is modelled by directly specifying the emission probabilities for each combination of features. We call this a layer of emissions. Additional layers with other features can be created. In this case, the likelihood of each judgment given probabilities from each layer can be calculated by multiplication of the probabilities from each layer.

For the calculation of state probabilities, we can use forward recursion (Rabiner, 1989). The algorithm proceeds through the observed sequence, and at each step calculates the probability for each state given the probabilities of the observation, the probabilities of each state at the previous step, and the transition probabilities.

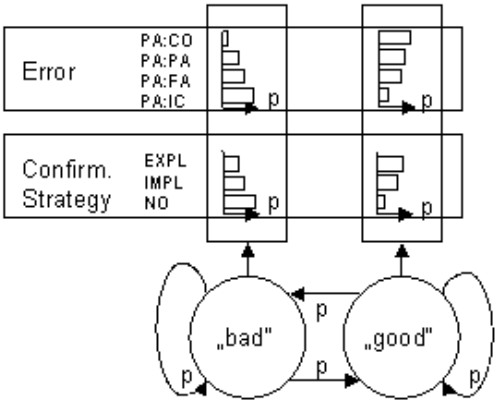

Figure 1. Topology of an HMM to model user judgments ("good" or "bad") in their probabilistic relation to dialog events (error and confirmation strategy) and the previous rating.

3 Data Collection

In order to train the model, data is needed in which the association between dialog events and judgments at each turn is represented. Therefore, we conducted an experiment in which users interacted with an SDS and judged the interaction after each dialog turn. In addition, all users followed the same five dialogs, which was possible by using the Wizard-of-Oz technique (Fraser and Gilbert 1991) and relatively strict scenario descriptions.

As the interactions were predefined, care was taken that a representative sample of relevant usability problems were passed in the interactions. Such problems were collected in a brainstorming session, and comprise understanding errors, task failure, prompt wording issues (such as incomplete information), missing control options, or not being able to input one of the criteria specified in the scenario description. As further features of the dialog, which might impact the quality judgment, the number of concepts which can be specified, as well as the confirmation strategy, should vary along the interactions.

We then designed a dialog system with a consistent strategy, which however allowed for the quality-relevant issues mentioned above to occur in the dialogs. The system was derived from the BoRIS restaurant information system (Möller 2005b), which enables the search of restaurants in the German city Bochum according to the criteria day, time, location, type of cuisine and pricing. For our system, we simplified the dialog strategy to system initiative, however, at the first turn, the users could say both day and time. If both were recognized, explicit confirmation was requested. If just one constraint was understood, confirmation was implicit. The last constraint specified (pricing) was not confirmed, and instead the restaurant information was output directly. If no restaurant was found according to the specified criteria, the user could change one of the constraints. Table 1 shows an example of a dialog with the system.

We also reformulated some prompts to fit the changed dialog strategy, and partly to incorporate potential wording problems. Prompts were recorded with a non-professional speaker,

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

Table 1. Example dialog with the BoRIS restaurant information system, version as used in the experiment.

using high-quality audio equipment. During the interactions, the wizard simply replayed the prompt foreseen at the current state of the predefined dialog script. In addition to the foreseen prompts, the wizard had at hand no-input and help prompts in case the user would behave unexpectedly.

25 users (13 females, 12 males), recruited from the campus, but not all students, participated in the experiment. Participants were aged between 20 and 46 years (M=26.5; STD=6.6). Ratings were given on a 5-point scale, where the points were labeled "bad", "poor", "fair", "good", and "excellent". Ratings were input through a number pad attached to the scale. Each participant rehearsed the procedure with a test dialog. Before the experiment, all users filled out a questionnaire measuring their technical affinity.

As the data collected in the described experiment are all needed to train the prediction model for as many combinations of feature values as possible, we conducted a second experiment to generate test data. For this test, we asked 17 persons from our lab to conduct two dialogs with the system mock-up. The test setup was the same as in the previous experiment, except that new dialogs were created without particular requirements or restrictions.

In both experiments, not all users behaved as we hoped. Therefore, not all of the predefined dialog scripts were judged by all participants (N=15…23 for training corpus, N=9…13 for test corpus; N: number of valid dialogs). For one dialog script in the training corpus, the deviating interactions were all equal (N=9), so distributions of ratings per turn could be calculated for comparison with the predicted distributions for this dialog. For the training and calculation of initial state probabilities, all dialogs in the training corpus were used.

Feature	Values
understanding errors	PA:PA (partially correct) PA:FA (failed) PA:IC (incorrect)
confirmation strategy	explicit implicit none
system speech act	ask for 2 constraints ask for 1 constraint ask for selection of a constraint provide info
user speech act	provide info repeat info confirm meta communication no-input
contextual appropriateness (Grice's maxims)	manner quality quantity relevance
task success	success failure

Table 2. Annotated dialog features.

The model derived from the data includes five possible states (one for each rating). For a list of features annotated in the dialogs see Figure 2.

4 Results

In order to evaluate the modeling approach, we first searched for the best model given the training data from the first experiment. We then applied this model to the test data from the second experiment in order to evaluate the model accuracy given unseen data. Afterwards, we examined if another model trained on the training set can predict the test set better, i.e. we "optimized" the model on the test data. Finally, we cross-check how well the model optimized on the test data performs on the training data, which gives a glimpse at how much the model is biased towards the test data.

As the criterion for the optimization, we determined the mean squared error (*MSE*), and averaged across all dialog script in the corpus on which the model was optimized. For each dialog script, all 5 probabilities (ratings "bad" to "excellent") at each dialog turn were taken into account, i.e. the squared prediction errors were added. If *rate* is the rating, then

$$MSE_{dial} = \frac{\sum_{turn=1}^{n} \sum_{rate=1}^{5} \left[p_{emp}(rate) - p_{pred}(rate) \right]^2}{n}$$

As this measure, in the particular way we applied it here, is not easily comparable to other results, we add two pictures illustrating the accuracy represented either by a rather low or by a rather high *MSE*. In addition, we report the mean absolute error (MAE_{max}) of the models in predicting the most likely rating at each state (mean rating if two ratings with equal probability) and the baseline performance when the unconditional distribution of ratings is predicted.

We first optimized a model on the training data, meaning that we selected parameters, trained the HMM with these parameters on the training data and then predicted results for all 6 dialog scripts contained in the training set (top of

Table 3). The optimized model was chosen as the one returning the smallest *MSE* (mean of all tasks). The best model included understanding errors interacting with confirmation type at each turn, and interacting with task success. As we analyzed the prediction results, we found that whenever the system changed from asking two constraints at a time to just one (which is done in order to avoid multiple errors in a row), the predictions were too positive. We therefore introduced a new feature, which is annotated whenever the system asks for a single constraint which has been asked in a more complex question before ("dummy"). In the model optimized on training data, this parameter was included on a separate feature layer. That is, this feature impacts quality perception independent of the other features' values.

We then used this model to predict the test data collected in the second experiment (top of Table 4). As expected, the *MSE* clearly increases; however, this was partly due to the difference in the sample of participants. As in the second experiment participants were recruited from our lab, their technical affinity was relatively high. Therefore, we retrained the HMM with only those 50% of the users from the training set who got the highest score on the technical affinity questionnaire. With this model, the prediction of test data improved.

In a next step, we optimized the model on the test set meaning that we searched for the parameter combination achieving the best result on the two test dialogs. However, the model was still trained on the training data from the first experiment. As expected, the *MSE* could be improved. However, only minor changes in the feature configuration are necessary: Still, errors and confirmation type are interacting on the same layer. However, task success is included as independent variable on a second layer, and instead, the error in the previous turn determines the impact of errors and confirmation on the ratings. Again, we tested if the prediction can be

Predicted: training dialogs	**Dial 1**	**Dial 2**	**Dial 3**	**Dial 4**	**Dial 5**	**Dial 6**	**Mean (basel.)**
Optimized on training dialogs	*Layer 1: Error, Confirm, Task Success*						
	Layer 2: Dummy						
MSE:	0.0185	0.0307	0.0166	0.0216	0.0333	0.0477	0.0281 (0.1201)
MAE_{max}:	0.7000	0.5714	0.2857	0.0556	0.3636	0.3333	0.3849 (0.6167)
Optimized on test dialogs	*Layer 1: Errors, Errors_lag, Confirmation*						
	Layer 2: TaskSuccess						
MSE:	0.0272	0.0358	0.0247	0.0374	0.0400	0.0574	0.0371 (0.1201)
MAE_{max}:	0.5000	0.4286	0.4286	0.3889	0.4545	0.3333	0.4223 (0.6167)
Number of valid dialogs (N):	22	15	23	17	17	9	

Table 3. Evaluation of predictions of training dialogs (mean squared error and mean absolute error in predicting the most probable state at each turn). Baseline results are given in brackets. The feature combinations with which results were obtained are also reported.

Predicted: test dialogs	**Dial 1**	**Dial 2**	**Mean (baseline)**
Optimized on training dialogs	*Layer 1: Error, Confirm, Task Success*		
	Layer 2: Dummy		
MSE:	0.1039	0.0429	0.0734 (0.1583)
MAE_{max}:	0.4444	0.6250	0.5347 (0.6944)
Optimized on training dialogs (tah)	*Layer 1: Error, Confirm, Task Success*		
	Layer 2: Dummy		
MSE:	0.0957	0.0387	0.0672 (0.1636)
MAE_{max}:	0.3333	0	0.1667 (0.6944)
Optimized on test dialogs (rf)	*Layer 1: Errors, Errors_lag, Confirm*		
	Layer 2: TaskSuccess		
MSE:	0.0789	0.0349	0.0569 (0.1583)
MAE_{max}:	0.4444	0.6250	0.5347 (0.6944)
Optimized on test dialogs (tah; rf)	*Layer 1: Errors, Confirm*		
MSE:	0.0860	0.0374	0.0617 (0.1636)
MAE_{max}:	0.3333	0	0.1667 (0.6944)
Number of valid dialogs (N):	9	13	

Table 4. Evaluation of predictions of training dialogs (tah=model trained on users with high technical affinity; rf=user speech act feature exclude from analysis)

improved by considering differences between the users' technical affinity. However, repeating the procedure for only those users with high technical affinity did not improve the result this time. Concerning the parameters, error and confirmation type were confirmed to be significant predictors of quality judgments. The dummy parameter created to improve the accuracy on training data was not proven useful for the prediction of the test set ratings.

In order to cross-check the validity of the model optimized on test data, we finally predicted the ratings of the 6 dialogs from the training set with the same model (bottom of Table 3). As can be seen, the prediction is worse than that from the model optimized on the training set. However, the quality of the prediction is still reasonable, showing that the two datasets do not demand for completely different models. All predictions are above the baseline.

5 Discussion

In the previous section, we presented results achieved with our models in terms of *MSE*. In order to gain meaning to the values of *MSE*, we added the mean absolute error of predicting the most probable judgment at each state. A closer look at the relation between *MSE* and MAE_{max} reveals that both measures are not strictly correlated (see e.g. the first two models in Table 4). While the *MSE* measures the distance at each measurement point in the distribution, the MAE_{max} is a rough indicator of the similarity of the shape of the predicted and observed probability curves. The results for MAE_{max} are promising, as predictions of test data are in the range of predictions of training data and better than the baseline. Also, predictions made from participants with high technical affinity achieve better results on the test data in all cases, which was expected, but not found for the *MSE* results.

Figure 2 presents examples of prediction results graphically. We chose one example of an average, and one of a relatively bad prediction, to allow extrapolation to other results presented. The pictures show that even a relatively high *MSE* corresponds to a fair quality of the prediction. The probability curves are mostly similar, mainly smoother than the observed probability distributions. Sometimes the predictions are too optimistic, however, usually the change in judgments is predicted, just not the extent of this change. We can only hypothesize

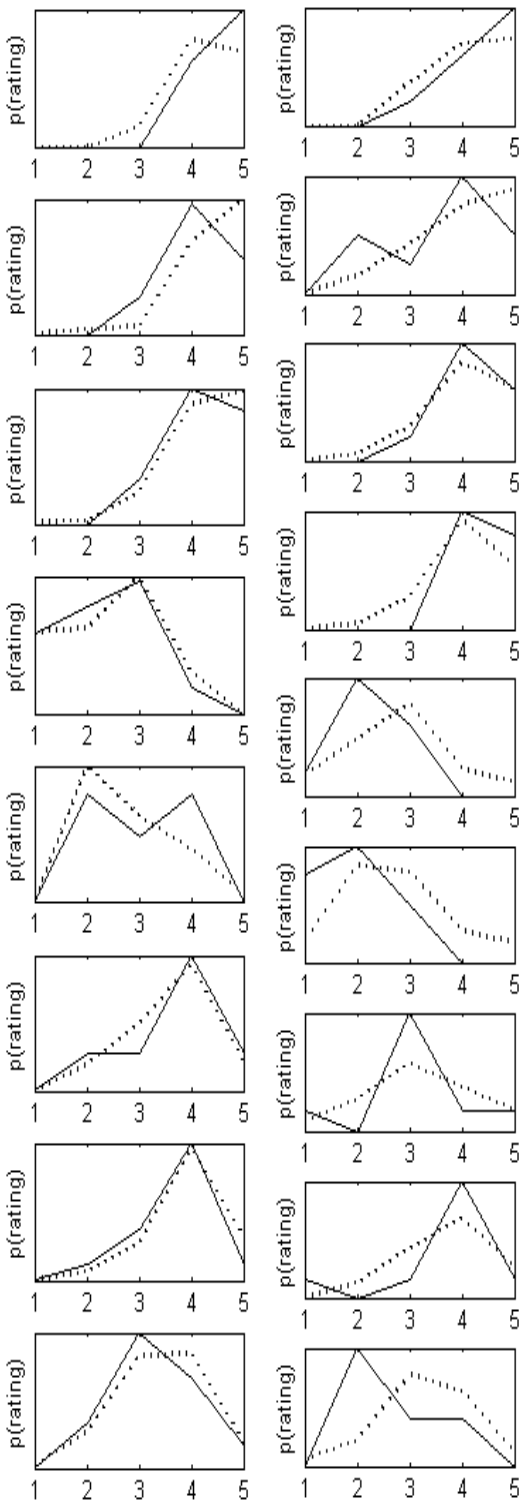

Figure 2. Examples of predictions on test data made with the model, illustrating the meaning of *MSE* values. Depicted are two dialogs (columns) with 9 (left) and 8 (right) turns (rows). For each turn, the empirical (solid line) and predicted (dotted line) rating distributions are given. Left: *MSE*=0.0957; N(emp)=9. Right: *MSE*=0.0349; N(emp)=13.

about the reasons for the participants to judge the respective dialog worse than predicted by the model. A possible reason is that users more easily decrease their judgments when the dialog has a longer history of problematic situations. According to our data, the users were relatively forgiving and increased their judgments if the dialog went well, even if previously errors had occurred. However, the errors might not really be forgot, and be reflected in the judgment of later problems and errors. Unfortunately, for reasons of data scarcity, the wider dialog history cannot be considered in the models.

Another source of prediction error might be the sample size available for the predicted dialogs. If sample size (N) and *MSE* values are compared among the dialogs, it can be observed that both values are correlated. This might be due to less smooth probability distribution curves if few ratings are available at each turn. While the curves depicted in Figure 2 are sometimes spiky, with increasing sample size normal distribution should be more likely. This might to some degree explain the clearly higher *MSE* for the test data predictions despite the relatively small error in predicting the most probable ratings.

6 Conclusion

In this paper, we presented a new approach to the prediction of user judgments about SDSs, using HMMs. The approach allows predicting the users' judgments at each step of a dialog. In predicting the distribution of ratings of many users, the approach takes into account differences between the users' judgment behaviors. This increases the usefulness of the model for a number of applications. E.g., in adaptive systems, the decision process can take into account differences between the users which cannot be attributed to user characteristics known to the system. If the model is applied to automatically generated dialogs, e.g. in the MeMo workbench (Engelbrecht et al., 2008a), a more detailed prediction of user satisfaction is enabled, allowing analysis on a turn-by-turn basis.

In addition, the approach facilitates the analysis of models and features affecting the quality ratings, as results can be compared to the empirical ratings with more detail. We hope to gain further insight into the relations between interaction parameters and user judgments by running simulations under different assumptions of relations between these entities.

A drawback of the approach is the generation of training data. The models presented in this paper cannot be assumed to be general, and in particular are lacking important parameters reflecting the timing in the dialogs. Therefore, as a next step the acquisition of judgments should be improved to be less disruptive for the interaction. In addition, it would be interesting to find a method for deriving the correct distributions of ratings at each dialog turn from a corpus of different dialogs, e.g. by grouping situations which are comparable. At the moment, we are also investigating if judgments can be acquired after the interactions without a loss in validity.

After all, the results we achieved with the model suggest that HMMs are suitable for modeling the users' quality perception of dialogs with SDSs. Further research on the topic will hopefully show if the dialog history has to be considered to a wider degree than in our present models.

Concerning dialog features and their relation to the judgments, the role of understanding errors in combination with the confirmation type could be established so far. More rich data are needed to work towards a general model for judgment predictions, including all relevant parameters. If judgments can be acquired after the interactions, we will be able to easily get the data needed for a better (and maybe complete) model. In any case, we are confident that the approach taken will allow a deeper analysis of the quality judgment process, which will enable progress by more analytical methods, such as formulating and testing hypotheses about this process.

References

Hua Ai, Fuliang Weng. 2008. *User Simulation as Testing for Spoken Dialog Systems.* Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, Ohio.

Grace Chung. 2004. *Developing a flexible spoken dialog system using simulation*. Proc. of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

Klaus-Peter Engelbrecht, Sebastian Möller. 2007. *Pragmatic Usage of Linear Regression Models for the Prediction of User Judgments*. Proc. of 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium.

Klaus-Peter Engelbrecht, Michael Kruppa, Sebastian Möller, Michael Quade. 2008a. *MeMo Workbench*

for Semi-Automated Usability Testing. Proc. of 9th Interspeech, Brisbane, Australia.

Klaus-Peter Engelbrecht, Sebastian Möller, Robert Schleicher, Ina Wechsung. 2008b. *Analysis of PARADISE Models for Individual Users of a Spoken Dialog System*. Proc. of ESSV 2008, Frankfurt/Main, Germany.

Klaus-Peter Engelbrecht, Felix Hartard, Florian Gödde, Sebastian Möller. 2009. *A Closer Look at Quality Judgments of Spoken Dialog Systems*, submitted to Interspeech 2009.

Norman M. Fraser, G. Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.

Rainer Guski. 1999. Personal and Social Variables as Co-determinants of Noise Annoyance. *Noise & Health*, 3:45-56.

Kate S. Hone, Robert Graham. 2001. *Subjective Assessment of Speech-system Interface Usability*. Proc. of EUROSPEECH, Aalborg, Denmark.

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

Ute Jekosch. 2005. *Voice and Speech Quality Perception. Assessment and Evaluation*, Springer, Berlin, Germany.

Ramón López-Cózar, Ángel de la Torre, José C. Segura and Antonio J. Rubio. 2003. Assessment of Dialogue Systems by Means of a New Simulation Technique. *Speech Communication*, 40(3):387-407.

Sebastian Möller. 2005a. *Perceptual Quality Dimensions of Spoken Dialog Systems: A Review and New Experimental Results*, Proc. of Forum Acusticum, Budapest, Hungary.

Sebastian Möller. 2005b. *Quality of Telephone-based Spoken Dialog Systems*. Springer, New York.

Sebastian Möller, Klaus-Peter Engelbrecht, Robert Schleicher. 2008. Predicting the Quality and Usability of Spoken Dialogue Services, *Speech Communication*, 50:730-744.

Morris A. Okun, Renee M. Weir. 1990. Toward a Judgment Model of College Satisfaction. *Educational Psychological Review*, 2(1):59-76.

Lawrence R. Rabiner. 1989. A tutorial on HMM and selected applications in speech recognition. *Proc. IEEE*, 77(2):257-286.

Verena Rieser, Oliver Lemon. 2008. *Automatic Learning and Evaluation of User-Centered Objective Functions for Dialogue System Optimisation*. Proc. of LREC'08, Marrakech, Morocco.

Bernhard Steimel, Oliver Jacobs, Norbert Pfleger, Sebastian Paulke. 2008. *Testbericht VOICE Award 2008: Die besten deutschsprachigen Sprachapplikationen*. Initiative Voice Business, Bad Homburg, Germany.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, Alicia Abella. 1997. *PARADISE: A Framework for Evaluating Spoken Dialogue Agents.* Proc. of ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics, Madrid, Spain.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, Alicia Abella. 1998. Evaluating Spoken Dialog Agents with PARADISE: Two Case Studies. *Computer Speech and Language*, 12:317-347.

Marilyn Walker, Candace Kamm, Diane Litman. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*, 6(3-4):363-377.

Craig Wootton, Michael McTear, Terry Anderson. 2007. *Utilizing Online Content as Domain Knowledge in a Multi-Domain Dynamic Dialogue System*. Proc. of Interspeech 2007, Antwerp, Belgium.