A Framework for Model-based Evaluation of Spoken Dialog Systems

Sebastian Möller
Deutsche Telekom Laboratories
Technische Universität Berlin
10587 Berlin, Germany
sebastian.moeller@telekom.de

Nigel G. Ward
Computer Science Department
University of Texas at El Paso
El Paso, Texas 79968, USA
nigelward@acm.org

Abstract

Improvements in the quality, usability and acceptability of spoken dialog systems can be facilitated by better evaluation methods. To support early and efficient evaluation of dialog systems and their components, this paper presents a tripartite framework describing the evaluation problem. One part models the behavior of user and system during the interaction, the second one the perception and judgment processes taking place inside the user, and the third part models what matters to system designers and service providers. The paper reviews available approaches for some of the model parts, and indicates how anticipated improvements may serve not only developers and users but also researchers working on advanced dialog functions and features.

1 Introduction

Despite the utility of many spoken dialog systems today, the user experience is seldom satisfactory. Improving this is a matter of great intellectual interest and practical importance. However improvements can be difficult to evaluate effectively, and this may be limiting the pace of innovation: today, valid and reliable evaluations still require subjective experiments to be carried out, and these are expensive and time-consuming. Thus, the needs of system developers, of service operators, and of the final users of spoken dialog systems argue for the development of additional evaluation methods.

In this paper we focus on the prospects for an *early* and *model-based* evaluation of dialog systems. Doing evaluation as early as possible in the design and development process is critical for improving quality, reducing costs and fostering innovation. Early evaluation renders the process more efficient and less dependent on experience, hunches and intuitions. With the help of such models predicting the outcome of user tests, the need for subjective testing can be reduced, restricting it to that subset of the possible systems which have already been vetted in an automatic or semi-automatic way.

Several approaches have already been presented for semi-automatic evaluation. For example, the PARADISE framework (Walker et al., 1997) predicts the effects of system changes, quantified in terms of interaction parameters, on an average user judgment. Others (Araki and Doshita, 1997; López-Cózar et al., 2003; Möller et al., 2006) have developed dialog simulations to aid system optimization. However the big picture has been missing: there has been no clear view of how these methods relate to each other, and how they might be improved and joined to support efficient early evaluation.

The remainder of this paper is organized as follows. Section 2 gives a brief review of different evaluation purposes and terminology, and outlines a new tripartite decomposition of the evaluation problem. One part of our framework models the behavior of user and system during the interaction, and describes the impact of system changes on the interaction flow. The second part models the perception and judgment processes taking place inside the user, and tries to predict user ratings on various perceptual dimensions. The third part models what matters to system designers and service providers for

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 182–189,
Columbus, June 2008.

a specific application. Sections 3, 4, and 5 go into specifics on the three parts of the framework, discussing which components are already available or conceivable. Finally, Section 6 discusses the potential impact of the approach, and Section 7 lists the issues to be resolved in future work.

2 Performance, Quality, Usability and Acceptability Evaluation

Developers tend to use indices of *performance* to assess their systems. The performance indicates the "ability of a system to provide the function it has been designed for" (Möller, 2005). The function and an appropriate measure for quantifying the degree of fulfillment may easily be determined for certain components — e.g. word accuracy for a speech recognizer or concept error rate for a speech understanding module — but it is harder to specify for other components, such as a dialog manager or an output generation module. However, definitive measures of component quality are not always necessary: what matters for such a module is its contribution to the quality of the entire interaction, as it is perceived by the user.

We follow the definition of the term *quality* as introduced by Jekosch (2000) and now accepted for telephone-based spoken dialog services by the International Telecommunication Union in ITU-T Rec. P.851 (2003): "Result of judgment of the perceived composition of an entity with respect to its desired composition". Quality thus involves a perception process and a judgment process, during which the perceiving person compares the perceptual event with a (typically implicit) reference. It is the comparison with a reference which associates a user-specific value to the perceptual event. The perception and the comparison processes take place in a particular context of use. Thus, both perception and quality should be regarded as "events" which happen in a particular personal, spatial, temporal and functional context.

Usability is one sub-aspect of the quality of the system. Following the definition in ISO 9241 Part 11 (1998), usability is considered 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". Usability is degraded when interaction problems occur. Such problems influence the perceptual event of the user interacting with the system, and consequently the quality s/he associates with the system as a whole. This may have consequences for the *acceptability* or the system or service, that is, how readily a customer will use the system or service. This can be quantified, for example as the ratio of the potential user population to the size of the target group.

It is the task of any evaluation to quantify aspects of system performance, quality, usability or acceptability. The exact target depends on the purpose of the evaluation (Paek, 2007). For example, the system developer might be most interested in quantifying the performance of the system and its components; s/he might further need to know how the performance affects the quality perceived by the user. In contrast, the service operator might instead be most interested in the acceptability of the service. S/he might further want to know about the satisfaction of the user, influenced by the usability of the system, and also by other (e.g. hedonic) aspects like comfort, joy-of-use, fashion, etc. Different evaluation approaches may be complementary, in the sense that metrics determined for one purpose may be helpful for other purposes as well. Thus, it is useful to describe the components of different evaluation approaches in a single framework.

Figure 1 summarizes our view of the evaluation landscape. At the lower left corner is what we can change (the dialog system), at the right is what the service operator might be interested in (a metric for the value of the system). In between are three components of a model of the processes taking place in the evaluation. The behavior model describes how system and user characteristics determine the flow of the interaction and translate this to quantitative descriptors. The perception and judgment model describes how the interaction influences the perceptual and quality events felt by the user, and translates these to observable user judgments. Finally the value model associates a certain value to the quality judgments, depending on the application. The model properties have been grouped in three layers: aspects of the user and his/her behavior, aspects of the system in its context-of-use, and the work of an external observer (expert) carrying out the evalua-

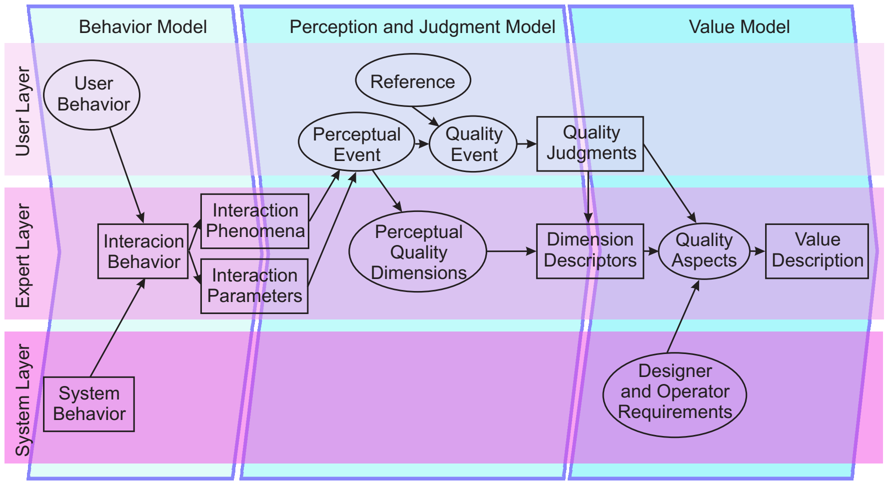

Figure 1: Tripartite view of a model-based evaluation. Observable properties are in boxes, inferred or hidden properties are in ovals. The layers organize the properties as mostly user-related, mostly system-related, and mostly expert-related, and mostly system-related.

tion. They have further been classified as to whether they are observable (boxes) or hidden from the evaluator (ovals).

The next three sections go through the three parts of the model left-to-right, explaining the needs, current status, and prospects.

3 Behavior Model

The behavior model translates the characteristics of the system and the user into predicted interaction behavior. In order to be useful, the representations of this behavior must be concise.

One way to describe dialog behavior is with *interaction parameters* which quantify the behavior of the user and/or the system during the interaction. Such parameters may be measured instrumentally or given by expert annotation. In an attempt to systematize best practice, the ITU-T has proposed a common set of interaction parameters suitable for the evaluation of telephone-based spoken dialog systems in ITU-T Suppl. 24 (2005). These parameters have been developed bottom-up from a collection of evaluation reports over the last 15 years, and include metrics related to dialog and communication in general, meta-communication, cooperativity, task, and speech-input performance (Möller, 2005). Unfortunately, it as is yet unclear which of these parameters relate to quality from a user's point-of-view. In addition, some metrics are missing which address critical aspects for the user, e.g. parameters for the quality and attractiveness of the speech output.

Another manageable way to describe system behavior is to focus on *interaction phenomena*. Several schemes have been developed for classifying such phenomena, such as system errors, user errors, points of confusion, dead time, and so on (Bernsen et al., 1998; Ward et al., 2005; Oulasvirta et al., 2006). Patterns of interaction phenomena may be reflected in interaction parameter values, and may be identified on that basis. Otherwise, they have to be determined by experts and/or users, by means of observation, interviews, thinking-aloud, and other techniques from usability engineering. (Using this terminology we can understand the practice of usability testing as being the identification of interaction phenomena, also known as "usability events" or "critical incidences", and using these to estimate specific quality aspects or the overall value of the system.)

Obtaining the interaction parameters and classifying the interaction phenomena can be done, obviously, from a corpus of user-system interactions. The challenge for early evaluation is to obtain these without actually running user tests. Thus, we would like to have a *system behavior model* and a *user behavior model* to simulate *interaction behavior*, and to map from system parameters and user properties to interaction parameters or phenomena. The value of such models for a developer is clear: they could enable estimation of how a change in the system (e.g. a change in the vocabulary) might affect the interaction properties. In addition to the desired effects, the side-effects of system changes are also important. Predicting such side-effects will substantially decrease the risk and uncertainty involved in dialogue design, thereby decreasing the gap between research and commercial work on dialog system usability (Heisterkamp, 2003; Pieraccini and Huerta, 2005).

Whereas modeling system behavior in response to user input is clearly possible (since in the last resort it is possible to fully implement the system), user behavior can probably not be modeled in closed form, because it unavoidably relates to the intricacies of the user and reflects the time-flow of the interaction. Thus, it seems necessary to employ a simulation of the interaction, as has been proposed by Araki and Doshita (1997) and López-Cózar et al. (2003), among others.

One embodiment of this idea is the MeMo workbench (Möller et al., 2006), which is based on the idea of running models of the system and of the user in a dedicated usability testing workbench. The system model is a description of the possible tasks (system task model) plus a description of the system's interaction behavior (system interaction model). The user model is a description of the tasks a user would want to carry out with the system (user task model) plus a description of the steps s/he would take to reach the goal when faced with the system (user interaction model). Currently the workbench uses simple attribute-value descriptions of tasks the system is able to carry out. From these, user-desired tasks may be derived, given some background knowledge of the domain and possible tasks. The system interaction model is described by a state diagram which models interactions as paths through a number of dialog states. The system designer provides one or several 'intended paths' through the interaction, which lead easily and/or effectively to the task goal.

The user's interaction behavior will strongly depend on the system output in the previous turn. Thus, it is reasonable to build the user interaction model on top of the system interaction model: The user mainly follows the 'intended path', but at certain points deviations from this path are generated in a probabilistic rule-based manner. For example, the user might deviate from the intended path, because s/he does not understand a long system prompt, or because s/he is irritated by a large number of options. Each deviation from the intended path has an associated probability; these are calculated from system characteristics (e.g. prompt length, number of options) and user characteristics (e.g. experience with dialog systems, command of foreign languages, assumed task and domain knowledge).

After the models have been defined, simulations of user-system interactions can be generated. These interactions are logged and annotated on different levels in order to detect interaction problems. Usability predictions are obtained from the (simulated) interaction problems. The simulations can also support reinforcement learning or other methods for automatically determining the best dialog strategy.

Building user interaction models by hand is costly. As an alternative to explicitly defining rules and probabilities, simulations can be based on data sets of actual interactions, augmented with annotations such as indications of the dialog state, current subtask, inferred user state, and interaction phenomena. Annotations can be generated by the dialog participants themselves, e.g. by re-listening after the fact (Ward and Tsukahara, 2003), or by top communicators, decision-makers, trend-setters, experts in linguistics and communication, and the like. Machine learning techniques can help by providing predictions of how users tend to react in various situations from lightly annotated data.

4 Perception and Judgment Model

Once the interaction behavior is determined, the evaluator needs to know about the impact it has on the quality perceived by the user. As pointed out in Section 2, the perception and judgments processes take place in the human user and are thus hidden from the observer. The evaluator may, however, ask the user to describe the *perceptual event* and/or the *quality event*, either qualitatively in an open form or quantitatively on rating scales. Provided that the experiment is properly planned and carried out, user *quality judgments* can be considered as direct quality measurements, reflecting the user's quality perception.

Whereas user judgments on quality will reflect the internal *reference* and thus depend heavily on the specific context and application, it may be assumed that the characteristics of the perceptual event are more universal. For example, it is likely that samples of observers and/or users would generally agree on whether a given system could be characterized as responsive, smooth, or predictable, etc. regardless of what they feel about the importance of each such quality aspect. We may take advantage of this by defining a small set of universal *perceptual quality dimensions*, that together are sufficient for predicting system value from the user's point-of-view.

In order to quantify the quality event and to identify perceptual quality dimensions, psychometric measurement methods are needed, e.g. interaction experiments with appropriate measurement scales. Several attempts have been made to come up with a common questionnaire for user perception measurement related to spoken dialog systems, for example the SASSI questionnaire (Hone and Graham, 2000) for systems using speech input, and the ITU-standard augmented framework for questionnaires (ITU-T Rec. P.851, 2003) for systems with both speech-input and speech-output capabilities. Studies of the validity and the reliability of these questionnaires (Möller et al., 2007) show that both SASSI and P.851 can cover a large number of different quality and usability dimensions with a high validity, and mainly with adequate reliability, although the generalizability of these results remains to be shown.

On the basis of batteries of user judgments obtained with these questionnaires, *dimension descriptors* of the perceptual quality dimensions can be extracted by means of factor analysis. A summary of such multidimensional analyses in Möller (2005b) reveals that users' perceptions of quality and usability can be decomposed into around 5 to 8 dimensions. The resulting dimensions include factors such as overall acceptability, task effectiveness, speed, cognitive effort, and joy-of-use. It should be noted that most such efforts have considered task-oriented systems, where effectiveness, efficiency, and success are obviously important, however these dimensions may be less relevant to systems designed for other purposes, for example tutoring or "edutainment" (Bernsen et al., 2004), and additional factors may be needed for such applications.

In order to describe the impact of the interaction flow on user-perceived quality, or on some of its sub-dimensions, we would ideally model the human perception and judgment processes. Such an approach has the clear advantage that the resulting model would be generic, i.e. applicable to different systems and potentially for different user groups, and also analytic, i.e. able to explain why certain interaction characteristics have a positive or negative impact on perceived quality. Unfortunately, the perception and judgment processes involved in spoken-dialog interaction are not yet well understood, as compared, for example, to those involved in listening to transmitted speech samples and judging their quality. For the latter, models are available which estimate quality with the help of peripheral auditory perception models and a signal-based comparison of representations of the perceptual event and the assumed reference (Rix et al., 2006). They are able to estimate user judgments on "overall quality" with an average correlation of around 0.93, and are widely used for planning, implementing and monitoring telephone networks.

For interactions with spoken dialog systems, the situation is more complicated, as the perceptual events depend on the interaction between user and systems, and not on one speech signal alone. A way out is not to worry about the perception processes, and instead to use simple linear regression models for predicting an average user judgment from various interaction parameters. The most widely used framework designed to support this sort of early evaluation is PARADISE (Walker et al., 1997). The target variable of PARADISE is an average of several user judgments (labeled "user satisfaction") of different system and interaction aspects, such as system voice, perceived system understanding, task ease, interaction pace, or the transparency of the interaction. The interaction parameters are of three types, those relating to efficiency (including elapsed time and the number of turns), those relating to "dialog quality" (including mean recognition score and the number of timeouts and rejections), and a measure of effectiveness (task success). The model can be trained on data, and the results are readily interpretable: they can indicate which features of the interaction are most critical for improving user satisfaction.

PARADISE-style models can be very helpful tools for system developers. For example, a recent investigation showed that the model can be used to effectively determine the minimum acceptable recognition rate for a smart-home system, leading to the same critical threshold as that obtained from user judgments (Engelbrecht and Möller, 2007). However, experience also shows that the PARADISE framework does not reliably give valid predictions of *individual* user judgments, typically covering only around 40-50% of the variance in the data it is trained on. The generality is also limited: cross-system extrapolation works sometimes but other times has low accuracy (Walker et al., 2000; Möller, 2005). These limitations are easy to understand in terms of Figure 1: over-ambitious attempts to directly relate interaction parameters to a measure of overall system value seem unlikely to succeed in general. Thus it seems wise to limit the scope of the perception and judgment component to the prediction of values on the perceptual quality dimensions.

In any case, there are several ways in which such models could be improved. One issue is that a linear combination of factors is probably not generally adequate. For example, parameters like the number of turns required to execute a specific task will have a non-zero optimum value, at least for inexperienced users. An excessively low number of turns will be as sure a sign of interaction problems as an excessively large number. Such non-linear effects cannot be handled by linear models which only support relationships like "the-more-the-better" or "the-less-the-better". Non-linear algorithms may overcome these limitations. A second issue is that of temporal context: instead of using a single input vector of interaction parameters for each dialog, it may be possible to apply a sequence of feature vectors, one for each exchange (user-system utterance pair). The features may consist not only of numeric measures but also of categories encoding interaction phenomena. Using this input one could then perhaps use a neural network or Hidden-Markov Model to predict various user judgments at the end of the interaction.

5 Value Model

Even if a model can predict user judgments of "overall quality" with high validity and reliability, this is not necessarily a good indicator of the acceptability of a service. For example, systems with a sophisticated and smooth dialog flow may be unacceptable for frequent users because what counts for them is effectiveness and efficiency only. Different users may focus on different quality dimensions in different contexts, and weight them according to the task, context of use, price, etc.

A first step towards addressing this problem is to define *quality aspects* that a system developer or service operator might be concerned about. There can be many such, but in usability engineering they are typically categorized into "effectiveness", "efficiency" and "satisfaction". A more detailed taxonomy of quality aspects can be found in Möller (2005). On the basis of this or other taxonomizations, value prediction models can be developed. For example, a system enabling 5-year old girls to "talk to Barbie" might ascribe little importance to task completion, speech recognition accuracy, or efficiency, but high importance to voice quality, responsiveness, and unpredictability. The value model will derive a *value description* which takes such a weighting into account. A model for systems enabling police officers on patrol to obtain information over the telephone would have very different weights.

Unfortunately, there appear to be no published descriptions of value prediction models, perhaps because they are very specific or even proprietary, depending on a company's business logic and customer base. Such models probably need not be very complex: it likely will suffice to ascribe weights to the perceptual quality dimensions, or to quality aspects derived from system developer and/or service operator requirements. Appropriate weights may be uncovered in stakeholder workshops, where designers, vendors, usability experts, marketing strategists, user representatives and so on come together and discuss what they desire or expect.

6 Broader Impacts

We have presented a tripartite evaluation framework which shows the relationship between user and system characteristics, interaction behavior, perceptual and quality events, their descriptions, and the final value of the system or service. In doing so, we have mainly considered the needs of system developers. However, an evaluation framework that supports judgments of perceived quality could provide additional benefits for users. We can imagine user-specific value models, representing what is important to specified user groups. These could be solicited for an entire group, or inferred from each user's own personal history of interactions and decisions, e.g, through a personalization database available to the service operator. The models could also be used to support system selection, or to inform real-time system customization or adaptation.

Better evaluation will also support the needs of the research community. With the help of model-based evaluation, it will become easier for researchers not only to do evaluation more efficiently, but also to to produce more meaningful evaluation results; saying not just "this feature was useful" but also providing quantitative statements of how much the feature affects various interaction parameters, and from that how much it impacts the various quality dimensions, and ultimately the value itself. This will make evaluation more meaningful and make it easy for others to determine when an innovation is worth adopting, speeding technology transfer.

One might worry that a standardized framework might only be useful for evaluating incremental improvements, thereby discouraging work on radically different dialog design concepts. However well-designed evaluation components should enable this framework to work for systems of any type, meaning that it may be easier to explore new regions of the design space. In particular it may enable more accurate prediction of the value of design innovations which in isolation may not be effective, but which in combination may be.

7 Future Work

Although examples of some model components are available today, notably several interaction simulations and the PARADISE framework for predicting user judgments from interaction parameters, these are limited. To realize a complete and generally useful evaluation model will require considerable work, for example, on:

- *User behavior model*: Of the three components, perhaps the greatest challenges are in the development of user behavior models. We need to develop methods which produce simulated behavior which is realistic (congruent to the behavior of real users), and/or which produce interaction parameters and/or quality indicators comparable to those obtained by subjective interaction experiments. It is yet unclear whether realistic user behavior can also be generated for more advanced systems and domains, such as computer games, collaborative problem solving systems, or educational systems. We also need to develop models that accurately represent the behavior patterns of various user groups.

- *Interaction parameters*: Several quality aspects are still not reflected in the current parameter sets, e.g. indices for the quality of speech output. Some approaches are described in Möller and Heimansberg (2006), but the predictive power is still too limited. In addition, many parameters still have to be derived by expert annotation. It may be possible to automatically infer values for some parameters from properties of the user's and system's speech signals, and such analyses may be a source for new parameters, covering new quality aspects.

- *Perceptual* and *quality events* and *reference*: These items are subject of ongoing research in related disciplines, such as speech quality assessment, sound quality assessment, and product sound design. Ideas for better, more realistic modeling may be derived from cooperations with these disciplines.

- *Quality judgments* and *dimension descriptors*: In addition to the aspects covered by the SASSI and P.851 questionnaires, psychologists have defined methods for assessing cognitive load, affect, affinity towards technology, etc. Input from such questionnaires may provide a better basis for developing value models.

Although a full model may be out of reach for the next decade, a more thorough understanding of human behavior, perception and judgment processes is not only of intrinsic interest but promises benefits enough to make this a goal worth working towards.

Acknowledgments

This work was supported in part by NSF Grant No. 0415150.

References

M. Araki, and S. Doshita. 1997. Automatic Evaluation Environment for Spoken Dialogue Systems. *Dialogue Processing in Spoken Language Systems, ECAI'96 Workshop Proceedings*, Springer Lecture Notes in Artificial Intelligence No. 1236, 183-194, Springer, Berlin.

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

N. O. Bernsen, L. Dybkjær, L., and S. Kiilerich. 2004. Evaluating Conversation with Hans Christian Andersen. *Proc. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004)*, 3, pp. 1011-1014, Lisbon.

K.-P. Engelbrecht, and S. Möller. 2007. Using Linear Regression Models for the Prediction of Data Distributions. *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, pp. 291-294.

P. Heisterkamp. 2003. "Do not attempt to light with match!": Some Thoughts on Progress and Research Goals in Spoken Dialog Systems. *Proc. 8th Europ. Conf. on Speech Communication and Technology (Eurospeech 2003 – Switzerland).*

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

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

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

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

U. Jekosch. 2000. *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*. Habilitation thesis (unpublished), Universität/Gesamthochschule, Essen.

R. López-Cózar, A. De la Torre, J. Segura, and A. Rubio. 2003. Assessment of Dialog Systems by Means of a New Simulation Technique. *Speech Communication*, 40: 387-407.

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

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

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

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

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

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

T. Paek. 2007. Toward Evaluation that Leads to Best Practices: Reconciling Dialog Evaluation in Research and Industry. *Bridging the Gap: Academic and Industrial Research in Dialog Technologies Workshop Proceedings*, Rochester, pp. 40-47.

R. Pieraccini, J. Huerta. 2005. Where Do We and Commercial Spoken Dialog Systems. *Proc. 6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, pp. 1-10.

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

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

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

N. Ward, A. G. Rivera, K. Ward, and D. G. Novick. 2005. Root Causes of Lost Time and User Stress in a Simple Dialog System. *Proc. 9th European Conf. on Speech Communication and Technology (Interspeech 2005)*, Lisboa.

N. Ward, and W. Tsukahara. 2003. A Study in Responsiveness in Spoken Dialogue. *International Journal of Human-Computer Studies*, 59: 603-630.