Parameters for Quantifying the Interaction with Spoken Dialogue Telephone Services

Sebastian Möller
Institut für Kommunikationsakustik (IKA), Ruhr-Universität Bochum, Germany
sebastian.moeller@ruhr-uni-bochum.de

Abstract

When humans interact with spoken dialogue systems, parameters can be logged which quantify the flow of the interaction, the behavior of the user and the system, and the performance of individual system modules during the interaction. Although such parameters are not directly linked to the quality perceived by the user, they provide useful information for system development, optimization, and maintenance. This paper presents a collection of such parameters which are now considered to be recommended by the International Telecommunication Union (ITU-T) for evaluating telephone-based spoken dialogue services. As an initial evaluation, a case study is described which shows that the parameters correlate only weakly with subjective judgments, but that they still may be used for predicting quality with PARADISE-style regression models.

1 Introduction

Speech technology devices, such as automatic speech recognition (ASR), speaker verification, speech synthesis, or spoken dialogue systems (SDSs), are increasingly used in wireline and mobile telephone networks to provide automatic voice-enabled services. In contrast to simple interactive voice response (IVR) systems with DTMF input, spoken dialogue systems offer the full range of speech interaction capabilities, including the recognition of user speech, the assignment of meaning to the recognized words, the decision on how to continue the dialogue, the formulation of a linguistic response, and the generation of spoken output to the user. In this way, a more-or-less "natural" spoken interaction between user and system is enabled.

Recently, the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) set up a new Recommendation describing subjective evaluation methods for telephone services based on spoken dialogue systems (ITU-T Rec. P.851, 2003). This Recommendation describes methods for conducting subjective evaluation experiments in order to determine quality from a user's point-of-view. For enabling system developers to get rough estimates of quality during the development phase, these methods are foreseen to be complemented by a set of so-called *interaction parameters*. Such parameters help to quantify the flow of the interaction, the behavior of the user and the system, and the performance of the speech technology devices involved in the interaction. They address system performance from a system developer's and service operator's point-of-view, and thus provide complementary information to subjective evaluation data.

The present paper provides an overview of interaction parameters which have been used for evaluating SDSs in the past 15 years, based on theoretical work which is described in Möller (2005). Section 2 presents a brief characterization of the parameters, with respect to the interaction aspect they address and the measurement method which is required to determine the parameter. The parameters are categorized and listed in Section 3. Section 4 presents an initial evaluation of the set of parameters, showing their correlation to subjective quality judgments and their contribution for predicting quality, using PARADISE-style regression models. Section 5 summarizes the main findings and identifies future work to obtain a reduced set of parameters to be recommended by the ITU-T.

2 Characteristics of Interaction Parameters

Interaction parameters can be extracted when real or test users interact with the telephone service under consideration. The extraction is performed on the basis of log files, be it instrumentally or with the help of a transcribing and annotating expert. Parameters which relate to the surface form of the utterances exchanged between user and system, like the duration of the interaction or the number of turns, can usually be measured fully instrumentally. On the other hand, human transcription and annotation is necessary when not only the surface

form (speech signals) is addressed, but also the contents and meaning of system or user utterances (e.g. to determine a word or concept accuracy). Both (instrumental and expert-based) ways of collecting interaction parameters should be combined in order to obtain as much information as possible.

Because interaction parameters are based on data which has been collected in an interaction between user and system, they are influenced by the characteristics of the system, of the user, and of the interaction between both. These influences cannot be separated, because the user's behavior is strongly influenced by that of the system (e.g. the questions asked by the system), and vice-versa (e.g. the vocabulary and speaking style of the user influences the system's recognition and understanding accuracy). Consequently, interaction parameters strongly reflect the characteristics of the user group they have been collected with.

Interaction parameters are either determined in a laboratory test setting under controlled conditions, or in a field test. In the latter case, it may not be possible to extract all parameters, because not all necessary information can be gathered. For example, if the success of a task-oriented interaction (e.g. collection of a train timetable) is to be determined, then it is necessary to know about the exact aims of the user. Such information can only be collected in a laboratory setting, e.g. in the way it is described in ITU-T Rec. P.851 (2003). In case that the fully integrated system is not yet available, it is possible to collect parameters from a so-called "Wizard-of-Oz" simulation, where a human experimenter replaces missing parts of the system under test. The characteristics of such a simulation have to be taken into account when interpreting the obtained parameter values.

Interaction parameters can be calculated on a word level, on a sentence or utterance level, or on the level of a full interaction or dialogue. In case of word or utterance level parameters, average values are often calculated for each dialogue. The parameters collected with a specific group of users may be analyzed with respect to the impact of the system (version), the user group, and the experimental setting (scenarios, test environment, etc.), using standard statistical methods. A characterization of these influences can be found in Möller (2005).

3 Review of Interaction Parameters

Based on a broad literature survey, parameters were identified which have been used in different assessment and evaluation experiments during the past 15 years. The respective literature includes Billi et al. (1996), Boros et al. (1996), Carletta (1996), Cookson (1988), Danieli and Gerbino (1995), Fraser (1997), Gerbino et al. (1993), Glass et al. (2000), Goodine et al. (1992), Hirschman and Pao (1993), Kamm et al. (1998), Polifroni et al. (1992), Price et al. (1992), San-Segundo et al. (2001), Simpson and Fraser (1993), Skowronek (2002), Strik et al. (2000, 2001), van Leeuwen and Steeneken (1997), Walker et al. (1997, 1998), Zue et al. (2000).

The parameters can broadly be classified as follows:

- Dialogue- and communication-related parameters
- Meta-communication-related parameters
- Cooperativity-related parameters
- Task-related parameters
- Speech-input-related parameters

These categories will be briefly discussed in the following sections. The respective parameters are listed in the Appendix, together with a definition, the interaction level addressed by the parameter (word, utterance or dialogue), as well as the measurement method (instrumental or expert annotation).

3.1 Dialogue- and Communication-Related Parameters

Parameters which refer to the overall dialogue and to the communication of information give a very rough indication of how the interaction takes place. They do not specify the communicative function of each individual utterance in detail. These parameters are listed in Table 2 of the Appendix, and include duration-related parameters (overall dialogue duration, duration of system and user turns, system and user response delay), and word- and turn-related parameters (average number of system and user turns, average number of words per system and per user turn, number of system and user questions).

Two parameters which have been proposed by Glass et al. (2000) are worth noting: The *query density* gives an indication of how efficiently a user can provide new information to a system, and the *concept efficiency* describes how efficiently the system can absorb this information from the user. These parameters also refer to the system's language understanding capability, but they have been included in this section because they result from the system's interaction capabilities as a whole, and not purely from the language understanding capabilities.

All parameters in this category are of global character and refer to the dialogue as a whole, although they are partly calculated on an utterance level. Global parameters are sometimes problematic, because the individual differences in cognitive skill may be large in relation to the system-originated differences, and because subjects might learn strategies for task solution which have a significant impact on global parameters.

3.2 Meta-Communication-Related Parameters

Meta-communication, i.e. the communication about communication, is particularly important for the spoken interaction with systems which have limited recogni-

tion, understanding and reasoning capabilities. In this case, correction and clarification utterances or even sub-dialogues are needed to recover from misunderstandings.

The parameters belonging to this group quantify the number of system and user utterances which are part of meta-communication. Most of the parameters are calculated as the absolute number of utterances in a dialogue which relate to a specific interaction problem, and are then averaged over a set of dialogues. They include the number of help requests from the user, of time-out prompts from the system, of user utterances rejected by the system in the case that no semantic content could be extracted (ASR rejections), of diagnostic system error messages, of barge-in attempts from the user, and of user attempts to cancel a previous action.

The ability of the system (and of the user) to recover from interaction problems can be described in two ways: Either explicitly by the correction rate, i.e. the percentage of all (system or user) turns which are primarily concerned with rectifying an interaction problem, or implicitly with the *implicit recovery* parameter, which quantifies the capacity of the system to regain utterances which have partially failed to be recognized or understood.

In contrast to the global measures, most meta-communication-related parameters describe the function of system and user utterances in the communication process. Thus, most parameters have to be determined with the help of an annotating expert. The parameters are listed in Table 3 of the Appendix.

3.3 Cooperativity-Related Parameters

Cooperativity has been identified as a key aspect for a successful interaction with a spoken dialogue system (Bernsen et al., 1998). Unfortunately, it is difficult to quantify whether a system behaves cooperatively or not. Several of the dialogue- and meta-communication-related parameters somehow relate to system cooperativity, but they do not attempt to quantify this aspect.

Direct measures of cooperativity are the contextual appropriateness parameters introduced by Simpson and Fraser (1993). Each system utterance has to be judged by a number of experts as to whether it violates one or more of Grice's maxims for cooperativity, see Grice (1975). These principles have been stated more precisely by Bernsen et al. (1998) with respect to spoken dialogue systems.

The utterances are classified into the categories of appropriate (not violating Grice's maxims), inappropriate (violating one or more maxim), appropriate/inappropriate (the experts cannot reach agreement in their classification), incomprehensible (the content of the utterance cannot be discerned in the dialogue context), or total failure (no linguistic response from the system). It has to be noted that the classification is not always straightforward, and that interpretation principles may be necessary.

3.4 Task-Related Parameters

Current state-of-the-art telephone services enable task-orientated interactions between system and user, and task success is a key issue for the usefulness of a service. Task success may best be determined in a laboratory situation where explicit tasks are given to the test subjects, see Möller (2005). However, realistic measures of task success have to take into account potential deviations from the scenario by the user, either because he/she did not pay attention to the instructions given in the scenario, because of his/her inattentiveness to the system utterances, or because the task was unresolvable and had to be modified in the course of the dialogue.

Modification of the experimental task is considered in most definitions of task success which are reported in the literature. Success may be reached by simply providing the right answer to the constraints set in the instructions, by constraint relaxation from the system or from the user (or both), or by spotting that no solution exists for the defined task. Task failure may be tentatively attributed to the system's or to the user's behavior, the latter however being influenced by the one of the system.

A different approach to determine task success is the κ coefficient. It assumes a speech-understanding approach which is based on attributes (concepts, slots) for which allowed values have to be assigned in the course of the dialogue, resulting in attribute-value-pairs (AVPs). A set of all available attributes together with the values assigned by the task (a so-called attribute-value matrix, AVM) completely describes a task which can be carried out with the help of the system. In order to determine the κ coefficient, a confusion matrix *M(i,j)* is set up for the attributes in the key (scenario definition) and in the reported solution (log file of the dialogue). Then, the agreement between key and solution *P(A)* and the chance agreement *P(E)* can be calculated from this matrix, see Table 5. *M(i,j)* can be calculated for individual dialogues, or for a set of dialogues which belong to a specific system or system configuration.

The κ coefficient relies on the availability of a simple task coding scheme, namely in terms of an AVM. However, some tasks cannot be characterized as easily. In that case, more elaborated approaches to task success are needed, approaches which usually depend on the type of task under consideration.

3.5 Speech-Input-Related Parameters

The speech input capability of a spoken dialogue system is determined by its capability to recognize words and utterances, and to extract the meaning from the recognized string. The speech recognition task can

be categorized into isolated word recognition, keyword spotting, or continuous speech recognition. Speech understanding is often performed on the basis of attribute-value pairs, see the previous section. The parameters described in the following paragraph address both speech recognition and speech understanding.

Continuous speech recognizers generally provide a word string hypothesis which has to be aligned with a reference transcription produced by an annotating expert. On the basis of the alignment, the number of correctly determined words c_w, of substitutions s_w, of insertions i_w, and of deletions d_w is counted. These counts can be related to the total number of words in the reference n_w, resulting in two alternative measures of recognition performance, the word error rate *WER* and the word accuracy *WA*, see Table 6.

Complementary performance measures can be defined on the sentence level, in terms of a sentence accuracy, *SA*, or a sentence error rate, *SER*, see Table 6. In general, *SA* is lower than *WA*, because a single misrecognised word in a sentence impacts the *SA* parameter. It may however become higher than the word accuracy, especially when many single-word sentences are correctly recognized. The fact that *SER* and *SA* penalize a whole utterance when a single misrecognised word occurs has been pointed out by Strik et al. (2000, 2001); the problem can be circumvented with the parameters *NES* and *WES*, see Table 6. When utterances are not separated into sentences, all sentence-related metrics can also be calculated on an utterance instead of a sentence level.

Isolated word recognizers provide an output hypothesis for each input word or utterance. Input and output words can be directly compared, and similar performance measures as in the continuous recognition case can be defined, omitting the insertions. Instead of the insertions, the number of "false alarms" in a time period can be counted, see van Leeuwen and Steeneken (1997). *WA* and *WER* can also be determined for keywords only, when the recognizer operates in a keyword-spotting mode.

For speech understanding assessment, two common approaches have to be distinguished. The first one is based on the classification of system answers to user questions into categories of correctly answered, partially correctly answered, incorrectly answered, or failed answers. The individual answer categories can be combined into measures which have been used in the US DARPA program, see Table 6. The second way is to classify the system's parsing capabilities, either in terms of correctly parsed utterances, or of correctly identified AVPs. On the basis of the identified AVPs, global measures such as the concept accuracy, *CA*, the concept error rate, *CER*, or the understanding accuracy, *UA*, can be calculated. All parameters are listed in Table 6.

3.6 Further Parameters

When separating the quality of an SDS-based service into quality aspects, in the way which is indicated in ITU-T Rec. P.851 (2003, Section 5.3), it can be observed that several aspects of quality are not addressed by interaction parameters. No parameters directly relate to usability, user satisfaction, acceptability, or speech output quality. So far, only very few approaches have been made which address the quality of speech output (be it concatenated or synthesized) in a parametric way. Instrumental measures related to speech intelligibility are defined e.g. in IEC Standard 60268-16 (1998), but they have not been designed for a telephone environment. Concatenation cost measures have been proposed which can be calculated from the input text and the speech database of a concatenative synthesis system (Chu and Peng, 2001). Although they sometimes show high correlations to mean opinion scores obtained in subjective experiments, such measures are very specific to the speech synthesizer and its concatenation corpus.

4 Initial Evaluation of Interaction Parameters

Although interaction parameters as the ones defined in Section 3 are important for system design, optimization and maintenance, they are not directly linked to the quality which is perceived by the human user. Consequently, the collection of interaction parameters should be complemented by a collection of user judgments, as it is described in ITU-T Rec. P.851 (2003).

In order to determine the relationship between subjective user judgments and interaction parameters, a limited case study has been carried out in the frame of the EC-funded IST project INSPIRE (INfotainment management with SPeech Interaction via REmote microphones and telephone interfaces). In this project, a prototype of a spoken dialogue system for controlling domestic devices (lamps, blinds, video recorder, answering machine, etc.) has been set up. The prototype has been evaluated in a controlled laboratory experiment at IKA. Because the speech recognizer was not available when the experiment was carried out, it had to be replaced by a human transcriber, making this a partly Wizard-of-Oz-based experiment.

During this experiment, 24 test users interacted with the system in a realistic home environment, following three scenario-guided interactions, each comprising several tasks. After each interaction, users were asked to fill in a questionnaire with 37 statements which has been designed following the methodology of ITU-T Rec. P.851 (2003). In parallel, the interactions have been logged, transcribed and annotated using a specifically-designed annotation interface (Skowronek, 2002; Möller, 2005). From the annotation, 64 parameters

could be extracted for each interaction which are mainly identical to the ones listed in Section 3. Thus, a set of user judgments on quality and interaction parameters is available for the initial evaluation, reflecting the same set of interactions with a prototypical system. Details on the experiment are described in Möller et al. (2005).

4.1 Correlation between Interaction Parameters and User Judgments

From this database, correlations between interaction parameters and subjective judgments have been calculated. Because several interaction parameters and user judgments do not follow a Gaussian distribution, Spearman rank-order correlations ρ have been chosen. The results were disappointing on a first view: The highest coefficients were around 0.6.

Interestingly, quality-related information seems to be captured mostly in the speech-recognition- and speech-understanding-related parameters. This is astonishing, because the (simulated) recognition accuracy of the INSPIRE system was nearly perfect (mean $WA = 97.2\%$). The recognition-related parameters were shown to have correlations of up to 0.6 with interaction control, up to 0.52 with interaction pleasantness, up to 0.47 with the difficulty of operation, up to 0.43 with system helpfulness, up to 0.42 with dialogue smoothness, and up to 0.40 with error recovery. The correlation between speech-recognition- and speech-understanding-related parameters is only moderate, justifying measuring both types of parameters to obtain a maximum of information. Perceived system understanding correlates only moderately with the measured understanding accuracy, *UA* ($\rho = 0.41$).

With respect to efficiency, humans do not seem to be adequate measurement instruments either. The correlation between the perceived length of a dialogue and *DD* (communication efficiency) is very low, as well as the correlation between annotated and perceived task success (task efficiency).

The subjective judgment on overall quality seems to be mainly dominated by the characteristics of the system turns (*STD*: $\rho = 0.40$), by the understanding accuracy (*UA*: $\rho = 0.39$; *UCT*: $\rho = 0.36$), and by the recognition accuracy (ρ between 0.39 and 0.42). Still, this correlation is not high enough to be able to predict overall system quality on the basis of individual interaction parameters.

4.2 Quality Prediction Models

More sophisticated models have been developed to predict system usability and acceptability from a combination of parameters. The most popular approach is the PARADISE framework developed by Walker et al. (1997, 1998). The model aims at predicting "user satisfaction", which is calculated as an arithmetic mean over several user judgments on different quality aspects, as a linear combination of several interaction parameters. In its original version, Walker et al. used 8-9 interaction parameters as an input to the model, including a subjective judgment on task success. The weighting coefficients of the linear prediction function are determined with the help of a multivariate linear regression analysis, using a database of user judgments and interaction parameters which have been collected under controlled (laboratory) conditions.

From the INSPIRE database, several PARADISE-style models have been calculated, using different user judgments as the prediction target (judgment on "overall quality", "user satisfaction", or the arithmetic mean over all 37 judgments), and several sets of interaction parameters as the input variables (full set of 64 parameters or restricted set of 5 parameters similar to Walker et al., 1997). In particular, two types of parameters have been used for describing task success: Either an expert-derived weighted task success index *TSe* (which is calculated from the *TS* labels of Table 2, assigning a value of one for each sub-task which has been successfully achieved by the user, and a value of zero for all failures), or a user judgment of task success *TSu* (as it was the case in the experiments reported in Walker et al., 1997 and 1998). The regression algorithm used a stepwise (forward-backward) inclusion of parameters (for 64 parameters) or a forced inclusion of all parameters (for 5 parameters only), did not include a constant term, and replaced missing values by their respective means.

Table 1: Regression models.

Input parameters		Target variable	Prediction result	
# par.	Task success		R^2_{corr}	# par.
64	*TSe*	Overall quality	0.247	2
64	*TSe*	User satisfaction	0.409	4
64	*TSe*	Mean of all judgm.	0.420	4
64	*TSu*	Overall quality	0.409	3
64	*TSu*	User satisfaction	0.409	4
64	*TSu*	Mean of all judgm.	0.459	3
5	*TSe*	Overall quality	0.091	5
5	*TSe*	User satisfaction	0.022	5
5	*TSe*	Mean of all judgm.	0.133	5
5	*TSu*	Overall quality	0.310	5
5	*TSu*	User satisfaction	0.086	5
5	*TSu*	Mean of all judgm.	0.305	5

The results are shown in Table 1. Indicated is the amount of variance in the subjective judgments which can be covered by the respective model (R^2_{corr}) and the number of input parameters selected by the regression algorithm. For the large set of input parameters, R^2_{corr} reaches 0.46 in the best case, which is comparable to the prediction accuracy reported by Walker et al. (1997, 1998). However, when using only the restricted set of

parameters as an input to the regression analysis, the prediction accuracy is much lower. The user-derived judgment of task success leads in all cases to better prediction results; it is particularly important when only few input parameters are available. All in all, the prediction accuracy does not depend on the number of input parameters, but on their informative value.

5 Conclusions

An overview has been presented of interaction parameters quantifying the interaction between a user and a spoken dialogue system. Such parameters can be used in the design, implementation, optimization and operation phase of SDS-based services. They provide important information to the system developer, but no direct measures of quality, as it would be perceived by the user of the respective service.

The set of parameters has been evaluated in a pilot experiment carried out with an SDS for controlling domestic devices. The results show that the correlation between individual interaction parameters and subjective user judgments is indeed relatively low; highest correlations were in the area of 0.6, and for overall quality not higher than 0.42. Nevertheless, a combination of parameters can be used to predict overall quality or user satisfaction, based on a linear regression model defined by the PARADISE framework. Such models may capture about 45% of the variance in the subjective data, provided that the right – informative – parameters are selected as an input to the model. Still, this value is too low to replace subjective quality judgments by interaction parameters when the quality of SDS-based services is to be measured.

The collected set of interaction parameters is considered by the ITU-T for a supplement to its P-Series Recommendations, to be approved in late 2005 (ITU-T Del. Contr. D.030, 2005). However, further empirical validation is necessary in order to restrict the full set of available parameters to the ones which are relevant for quality. Such a restricted set of interaction parameters will form the basis for a new Recommendation P.PST which will be developed by ITU-T SG12 in the next 1-2 years. Contributions in this respect are invited by the ITU-T, see the roadmap on http://www.itu.int/ITU-T/studygroups/com12/q12roadmap/index.html.

Acknowledgements

The present work has been performed at IKA, Ruhr-University Bochum, in the context of the EC-funded IST-project INSPIRE (IST-2001-32746), see http://www.knowledge-speech.gr/inspire-project. Partners of INSPIRE were: Knowledge S.A., Patras, and WCL, University of Patras, both Greece; IKA, Ruhr-University Bochum, and ABS Jena, both Germany; TNO Human Factors, Soesterberg and Philips Electronics Nederland B.V., Eindhoven, both The Netherlands; and EPFL, Lausanne, Switzerland. The author would like to thank Rosa Pegam for acting as a Wizard and for reviewing the manuscript; Noha El Mehelmi and Jörn Opretzka for annotating the dialogues; as well as all other INSPIRE partners for their support in the experiments and for fruitful discussions.

References

Bernsen, N.O., Dybkjær, H., Dybkjær, L. (1998). *Designing interactive speech systems: From first ideas to user testing*. Springer, Berlin.

Billi, R., Castagneri, G., Danieli, M. (1996). *Field trial evaluations of two different information inquiry systems*. In: Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96), Basking Ridge NJ, 129-134.

Boros, M., Eckert, W., Gallwitz, F., Gorz, G., Hanrieder, G., Niemann, H. (1996). *Towards understanding spontaneous speech: Word accuracy vs. concept accuracy*. In: Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96), IEEE, Piscataway NJ, 2, 1009-1012.

Carletta, J. (1996). *Assessing agreement of classification tasks: The kappa statistics*, Computational Linguistics, 22(2), 249-254.

Chu, M., Peng, H. (2001). *An objective measure for estimating MOS of synthesized speech*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 - Scandinavia), Aalborg, 3, 2087-2090.

Cookson, S. (1988). *Final evaluation of VODIS - Voice operated data inquiry system*. In: Proc. of Speech'88, 7th FASE Symposium, Edinburgh, 4, 1311-1320.

Danieli, M., Gerbino, E. (1995). *Metrics for evaluating dialogue strategies in a spoken language system*. In: Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAAI Symposium, US-Stanford CA, AAAI Press, Menlo Park CA, 34-39.

Fraser, N. (1997). *Assessment of interactive systems*. In: Handbook on Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, Berlin, 564-615.

Gerbino, E., Baggia, P., Ciaramella, A., Rullent, C. (1993). *Test and evaluation of a spoken dialogue system*. In: Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP'93), 2, 135-138.

Glass, J., Polifroni, J., Seneff, S., Zue, V. (2000). *Data collection and performance evaluation of spoken dialogue systems: The MIT experience*. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), Beijing, 4, 1-4.

Goodine, D., Hirschman, L., Polifroni, J., Seneff, S., Zue, V. (1992). *Evaluating interactive spoken language systems*. In: Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'92), Banff, 1, 201-204.

Grice, H.P. (1975). *Logic and conversation*. In: Syntax and Semantics, Vol. 3: Speech Acts (P. Cole and J.L. Morgan, eds.), Academic Press, New York NY, 41-58.

Hirschman, L., Pao, C. (1993). *The cost of errors in a spoken language system*. In: Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), Berlin, 2, 1419-1422.

IEC Standard 60268-16 (1998). Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index. International Electrotechnical Commission, Geneva.

ITU-T Delayed Contribution D.030 (2005). *Proposal for Parameters Describing the Performance of Speech Technology Devices,* Federal Republic of Germany (Author: S. Möller), ITU-T SG12 Meeting, 18-27 January 2005, CH-Geneva.

ITU-T Rec. P.851 (2003). Subjective quality evaluation of telephone services based on spoken dialogue systems. International Telecomm. Union, Geneva.

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.

Möller, S. (2005). *Quality of telephone-based spoken dialogue systems*. Springer, New York NY.

Möller, S., Smeele, P., Boland, H., Krebber, J. (2005). *Evaluating spoken dialogue systems according to de-facto standards: A case study*, submitted to Computer Speech and Language.

NIST Speech Recognition Scoring Toolkit (2001). *Speech recognition scoring toolkit*. National Institute of Standards and technology, http://www.nist.gov/speech/tools, Gaithersburg MD.

Polifroni, J., Hirschman, L., Seneff, S., Zue, V. (1992). *Experiments in evaluating interactive spoken language systems*. In: Proc. DARPA Speech and Natural Language Workshop, Harriman CA, 28-33.

Price, P.J., Hirschman, L., Shriberg, E., Wade, E. (1992). *Subject-based evaluation measures for interactive spoken language systems*. In: Proc. DARPA Speech and Natural Language Workshop, Harriman CA, 34-39.

San-Segundo, R., Montero, J.M., Colás, J., Gutiérrez, J., Ramos, J.M., Pardo, J.M. (2001). *Methodology for dialogue design in telephone-based spoken dialogue systems: A Spanish train information system*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 - Scandinavia), Aalborg, 3, 2165-2168.

Simpson, A., Fraser, N.M. (1993). *Black box and glass box evaluation of the SUNDIAL system*. In: Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), Berlin, 2, 1423-1426.

Skowronek, J. (2002). *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität, Bochum.

Strik, H., Cucchiarini, C., Kessens, J.M. (2001). *Comparing the performance of two CSRs: How to determine the significance level of the differences*. In: Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 - Scandinavia), Aalborg, 3, 2091-2094.

Strik, H., Cucchiarini, C., Kessens, J.M. (2000). *Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test*. In: Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000), Beijing, 4, 740-743.

van Leeuwen, D., Steeneken, H. (1997). *Assessment of recognition systems*. In: Handbook on Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, Berlin, 381-407.

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(3), 317-347.

Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A. (1997). *PARADISE: A framework for evaluating spoken dialogue agents*. In: Proc. of the 35th Ann. Meeting of the Assoc. for Computational Linguistics, Madrid, 271-280.

Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J., Hetherington, L. (2000). *JUPITER: A telephone-based conversational interface for weather information*. IEEE Trans. Speech and Audio Processing, 8(1), 85-96.

Appendix A. Definition of Interaction Parameters

Table 2: Dialogue- and communication-related interaction parameters.

Abbr.	Name	Definition	Int. level	Meas. meth.
DD	dialogue duration	Overall duration of a dialogue in [ms], see e.g. Fraser (1997).	dial.	instr.
STD	system turn duration	Average duration of a system turn, from the system starting speaking to the system stopping speaking, in [ms]. A turn is an utterance, i.e. a stretch of speech spoken by one party in the dialogue. (Fraser, 1997)	utter.	instr.
UTD	user turn duration	Average duration of a user turn, from the user starting speaking to the user stopping speaking, in [ms]. (Fraser, 1997)	utter.	instr.
SRD	system response delay	Average delay of a system response, from the user stopping speaking to the system starting speaking, in [ms]. (Fraser, 1997)	utter.	instr.
URD	user response delay	Average delay of a user response, from the system stopping speaking to the user starting speaking, in [ms]. (Fraser, 1997)	utter.	instr.
# turns	number of turns	Overall number of turns uttered in a dialogue. (Walker et al., 1998)	dial.	instr./ expert.
# system turns	number of system turns	Overall number of system turns uttered in a dialogue. (Walker et al., 1998)	dial.	instr./ expert.
# user turns	number of user turns	Overall number of user turns uttered in a dialogue. (Walker et al., 1998)	dial.	instr./ expert.
WPST	words per system turn	Average number of words per system turn in a dialogue. (Cookson, 1988)	utter.	instr./ expert.
WPUT	words per user turn	Average number of words per user turn in a dialogue. (Cookson, 1988)	utter.	instr./ expert.
# system questions	number of system questions	Overall number of questions from the system per dialogue.	dial.	expert.
# user questions	number of user questions	Overall number of questions from the user per dialogue. (Goodine et al., 1992; Polifroni et al., 1992)	dial.	expert.
QD	query density	Average number of new concepts (slots, see Section 3.4) introduced per user query. Being n_d the number of dialogues, $n_q(i)$ the total number of user queries in the i^{th} dialogue, and $n_u(i)$ the number of unique concepts correctly "understood" by the system in the i^{th} dialogue, then $$QD = \frac{1}{n_d}\sum_{i=1}^{n_d}\frac{n_u(i)}{n_q(i)}$$ A concept is not counted to $n_u(i)$ if the system already understood it in one of the previous utterances. (Glass et al., 2000)	set of dial.	expert.
CE	concept efficiency	Average number of turns which are necessary for each concept to be "understood" by the system. Being n_d the number of dialogues, $n_u(i)$ the number of unique concepts correctly "understood" by the system in the i^{th} dialogue, and $n_c(i)$ the total number of concepts in the i^{th} dialogue, then $$CE = \frac{1}{n_d}\sum_{i=1}^{N_d}\frac{n_u(i)}{n_c(i)}$$ A concept is counted whenever it was uttered by the user and was not already understood by the system. (Glass et al., 2000)	set of dial.	expert.

Table 3: Meta-communication-related interaction parameters.

Abbr.	Name	Definition	Int. level	Meas. meth.
# help request	number of help requests from the user	Overall number of user help requests in a dialogue. A user help request is labeled by the annotation expert if the user explicitly asks for help. This request may be formulated as a question (e.g. "What are the available options?") or as a statement ("Give me the available options!"). (Walker et al., 1998)	utter.	expert.
# system help	number of diagnostic system help messages	Overall number of help messages generated by the system in a dialogue. A help message is a system utterance which informs the user about available options at a certain point in the dialogue.	utter.	instr./ expert.
# time-out	number of time-out prompts	Overall number of time-out prompts, due to no response from the user, in a dialogue. (Walker et al., 1998)	utter.	instr.
# ASR rejection	number of ASR rejections	Overall number of ASR rejections in a dialogue. An ASR rejection is defined as a system prompt indicating that the system was unable to "hear" or to "understand" the user, i.e. that the system was unable to extract any meaning from a user utterance. (Walker et al., 1998)	utter.	instr.
# system error	number of diagnostic system error messages	Overall number of diagnostic error messages from the system in a dialogue. A diagnostic error message is defined as a system utterance in which the system indicates that it is unable to perform a certain task or to provide a certain information. (Price et al., 1992)	utter.	instr./ expert.
# barge-in	number of user barge-in attempts	Overall number of user barge-in attempts in a dialogue. A user barge-in attempt is counted when the user intentionally addresses the system while the system is still speaking. In this definition, user utterances which are not intended to influence the course of the dialogue (laughing, expressions of anger or politeness) are not counted as barge-ins. (Walker et al., 1998)	utter.	expert.
# cancel	number of user cancel attempts	Overall number of user cancel attempts in a dialogue. A user turn is classified as a cancel attempt if the user tries to restart the dialogue from the beginning, or if he/she explicitly wants to step one or several levels backwards in the dialogue hierarchy. (Kamm et al., 1998; San-Segundo et al., 2001)	utter.	expert.
SCT, *SCR*	number of system correction turns, system correction rate	Overall number (*SCT*) or percentage (*SCR*) of all system turns in a dialogue which are primarily concerned with rectifying a "trouble", thus not contributing new propositional content and interrupting the dialogue flow. A "trouble" may be caused by speech recognition or understanding errors, or by illogical, contradictory, or undefined user utterances. In case that the user does not give an answer to a system question, the corresponding system answer is labeled as a system correction turn, except when the user asks for an information or action which is not supported by the current system functionality. (Simpson and Fraser, 1993; Gerbino et al., 1993)	utter.	expert.
UCT, *UCR*	number of user correction turns, user correction rate	Overall number (*UCT*) or percentage (*UCR*) of all user turns in a dialogue which are primarily concerned with rectifying a "trouble", thus not contributing new propositional content and interrupting the dialogue flow (see *SCT*, *SCR*). (Simpson and Fraser, 1993; Gerbino et al., 1993)	utter.	expert.
IR	implicit recovery	Capacity of the system to recover from user utterances for which the speech recognition or understanding process partly failed. Determined by labeling the partially parsed utterances (see definition of *PA:PA* in Section 3.5) as to whether the system response was "appropriate" or not: $IR = \frac{\#\text{ utterances with appropriate system answer}}{PA:PA}$ For the definition of "appropriateness" see Grice (1975) and Bernsen et al. (1998). (Danieli and Gerbino, 1995)	utter.	expert.

Table 4: Cooperativity-related interaction parameters.

Abbr.	Name	Definition	Int. level	Meas. meth.
CA:AP, *CA:IA*, *CA:TF*, *CA:IC*, *%CA:AP*, *%CA:IA*, *%CA:TF*, *%CA:IC*	contextual appropriateness	Overall number or percentage of system utterances which are judged to be appropriate in their immediate dialogue context. Determined by labeling utterances according to whether they violate one or more of Grice's maxims for cooperativity: *CA:AP*: Appropriate, not violating Grice's maxims, not unexpectedly conspicuous or marked in some way. *CA:IA*: Inappropriate, violating one or more of Grice's maxims. *CA:TF*: Total failure, no linguistic response. *CA:IC*: Incomprehensible, content cannot be discerned by the annotation expert. For more details see Simpson and Fraser (1993) and Gerbino et al. (1993); the classification is similar to the one adopted in Hirschman and Pao (1993).	utter.	expert.

Table 5: Task-related interaction parameters.

Abbr.	Name	Definition	Int. level	Meas. meth.
TS	task success	Label of task success according to whether the user has reached his/her goal by the end of a dialogue, provided that this goal could be reached with the help of the system. The labels indicate whether the goal was reached or not, and the assumed source of problems: *S*: Succeeded (task for which solutions exist) *SCs*: Succeeded with constraint relaxation by the system *SCu*: Succeeded with constraint relaxation by the user *SCsCu*: Succeeded with constraint relaxation both from the system and from the user *SN*: Succeeded in spotting that no solution exists *Fs*: Failed because of the system's behavior, due to system adequacies *Fu*: Failed because of the user's behavior, due to non-cooperative user behavior See also Fraser (1997), Simpson and Fraser (1993) and Danieli and Gerbino (1995).	dial.	expert.
κ	kappa coefficient	Percentage of task completion according to the kappa statistics. Determined on the basis of the correctness of the result AVM reached at the end of a dialogue with respect to the scenario (key) AVM. A confusion matrix *M(i,j)* is set up for the attributes in the result and in the key, with *T* the number of counts in *M*, and t_i the sum of counts in column *i* of *M*. Then $\kappa = \frac{P(A) - P(E)}{1 - P(E)}$ with *P(A)* the proportion of times that the AVM of the actual dialogue and the key agree, $P(A) = \sum_{i=1}^{n} \frac{M(i,i)}{T}$. *P(E)* can be estimated from the proportion of times that they are expected to agree by chance, $P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2$. (Carletta, 1996; Walker et al., 1997)	dial. or set of dial.	expert.

Table 6: Speech-input-related interaction parameters.

Abbr.	Name	Definition	Int. level	Meas. meth.
WER, *WA*	word error rate, word accuracy	Percentage of words which have been correctly recognized, based on the orthographic form of the hypothesized and the (transcribed) reference utterance, and an alignment carried out with the help of the "sclite" algorithm, see NIST (2001). Designating n_w the overall number of words from all user utterances of a dialogue, and s_w, d_w and i_w the number of substituted, deleted and inserted words, respectively, then the word error rate and word accuracy can be determined as follows: $$WER = \frac{s_w + i_w + d_w}{n_w}$$ $$WA = 1 - \frac{s_w + i_w + d_w}{n_w} = 1 - WER$$ See Simpson and Fraser (1993); details on how these parameters can be calculated in case of isolated word recognition are given in van Leeuwen and Steeneken (1997).	word	instr./ expert.
SER, *SA*	sentence error rate, sentence accuracy	Percentage of entire sentences which have been correctly identified. Denoting n_s the total number of sentences, and s_s, i_s and d_s the number of substituted, inserted and deleted sentences, respectively, then: $$SER = \frac{s_s + i_s + d_s}{n_s}$$ $$SA = 1 - \frac{s_s + i_s + d_s}{n_s} = 1 - SER$$ (Simpson and Fraser, 1993)	utter.	instr./ expert.
NES	number of errors per sentence	Average number of recognition errors in a sentence. Being $s_w(k)$, $i_w(k)$ and $d_w(k)$ the number of substituted, inserted and deleted words in sentence k, then $$NES(k) = s_w(k) + i_w(k) + d_w(k)$$ The average *NES* can be calculated as follows: $$NES = \frac{\sum_{k=1}^{\#\ user\ turns} NES(k)}{\#\ user\ turns} = \frac{WER \cdot \#\ user\ words}{\#\ user\ turns}$$ (Strik et al., 2001)	utter.	instr./ expert.
WES	word error per sentence	Related to *NES*, but normalized to the number of words in sentence k, $w(k)$: $$WES(k) = \frac{NES(k)}{w(k)}$$ The average *WES* can be calculated as follows: $$WES = \frac{\sum_{k=1}^{\#\ user\ turns} WES(k)}{\#\ user\ turns}$$ (Strik et al., 2001)	word	instr./ expert.
AN:CO, *AN:IN*, *AN:PA*, *AN:FA*, *%AN:CO*, *%AN:IN*, *%AN:PA*, *%AN:FA*	number or percentage of correct/ incorrect/ partially correct/ failed system answers	Overall number or percentage of questions from the user which are • correctly (*AN:CO*) • incorrectly (*AN:IC*) • partially correctly (*AN:PA*) • not at all (*AN:FA*) answered by the system, per dialogue, see Polifroni et al. (1992), Goodine et al. (1992) and Hirschman and Pao (1993).	utter.	expert.

Abbr.	Name	Definition	Int. level	Meas. meth.
$DARPA_s$, $DARPA_{me}$	DARPA score, DARPA modified error	Measures according to the DARPA speech understanding initiative, modified by Skowronek (2002) to account for partially correct answers: $$DARPA_s = \frac{AN:CO - AN:IC}{\#\ user\ questions}$$ $$DARPA_{me} = \frac{AN:FA + 2 \cdot (AN:IC + AN:PA)}{\#\ user\ questions}$$ (Polifroni et al., 1992; Goodine et al., 1992; Skowronek, 2002)	utter.	expert.
PA:CO, *PA:PA*, *PA:IC*, *%PA:CO*, *%PA:PA*, *%PA:IC*	number of correctly/ partially correctly/ incorrectly parsed user utterances	Evaluation of the number of concepts (attribute-value pairs, AVPs) in an utterance which have been extracted by the system: *PA:CO*: All concepts of a user utterance have been correctly understood by the system. *PA:PA*: Not all but at least one concept of a user utterance has been correctly understood by the system. *PA:IC*: No concept of a user utterance has been correctly understood by the system. Expressed as the overall number or percentage of user utterances in a dialogue which have been parsed correctly/ partially correctly/ incorrectly. (Danieli and Gerbino, 1995)	utter.	expert.
CA, *CER*	concept accuracy, concept error rate	Percentage of correctly understood semantic units, per dialogue. Concepts are defined as attribute-value pairs (AVPs), with n_{AVP} the total number of AVPs, and s_{AVP}, i_{AVP} and d_{AVP} the number of substituted, inserted and deleted AVPs. The concept accuracy and the concept error rate can then be determined as follows: $$CA = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$$ $$CER = \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$$ (Gerbino et al., 1993; Simpson and Fraser, 1993; Boros et al., 1996 ; Billi et al., 1996)	utter.	expert.
UA	understanding accuracy	Percentage of user utterances in which all semantic units (AVPs) have been correctly extracted: $$UA = \frac{PA:CO}{\#\ user\ turns}$$ (Zue et al., 2000)	utter.	expert.