A New ITU-T Recommendation on the Evaluation of Telephone-Based Spoken Dialogue Systems

Sebastian Möller

Institute of Communication Acoustics (IKA), Ruhr-University Bochum
D-44780 Bochum, Germany
sebastian.moeller@ruhr-uni-bochum.de

Abstract

This article describes efforts which have recently been undertaken by the International Telecommunication Union (ITU-T) to agree on common methods for evaluating telephone services based on spoken dialogue systems. As a result of these efforts, a new ITU-T Recommendation P.851 (2003) has been approved. It summarizes on the one hand the factors of the system, of the service and of the user which influence the service quality. On the other hand, guidelines are presented on how to evaluate services with the help of subjective interaction experiments, in order to determine the user's quality perceptions. The relationships between influencing factors and perceived quality dimensions are displayed with the help of a taxonomy. This taxonomy puts different quality aspects into a logical relationship, and shows which factors have to be taken into account in the experimental set-up. The article discusses what has been reached in the new Recommendation, but also what is still missing in order to get more analytic information about the performance of system characteristics and their influence on overall service quality.

1. Introduction

Spoken dialogue systems (SDSs) become increasingly part of modern telephone networks. They enable access to databases and transactions via the phone, e.g. for obtaining train or air timetable information, stock exchange rates, tourist information, or for performing bank account operations or hotel reservations. In order to guarantee an adequate service quality for their users, both dialogue system designers as well as telephone network operators need to agree on standards for the assessment of individual system components, and for the evaluation of the entire system and the offered service. The agreement is particularly necessary when customers do not accept or even complain about a service, because the service operator is often not identical with the system designer.

As a first step into this direction, the Telecommunication Standardization Sector of the International Telecommunication Union, ITU-T, recently defined a new Recommendation on the subjective evaluation of telephone services which are based on spoken dialogue systems. It describes methods for evaluating the quality from a user's point of view, taking the SDS as a black box. These methods are based on laboratory experiments in which test subjects interact with a prototypical service in order to perform a specific, pre-defined task. The Recommendation has been agreed upon by the members of the ITU-T, and it is now available on the ITU-T web site (http://www.itu.int/rec/recommendation.asp).

Following a specific definition of quality and its underlying aspects (see Section 2), the Recommendation identifies a number of factors which exert an influence on the quality of the service, as it is experienced by its users. These factors relate to the characteristics of the underlying system, to the physical environment the service is used in, to the task which can be carried out with the help of the service, to the (non-physical) context of use, as well as to the characteristics of the individual user. They are discussed in Section 3. The factors can be displayed in terms of a taxonomy which shows the relationship between factors and the influenced quality aspects. It is important to identify such relationships, because they have to be taken into account in subjective evaluation experiments, in order to get valid measurements of service quality. The Recommendation provides some guidelines for carrying out such evaluation experiments, and it gives examples of questionnaires which can be used for this purpose (Section 4). Examples for guideline application are briefly referenced in Section 5.

It is the intention of the author, who is also the main author of the Recommendation, to give an overview of what has been reached, but also of what is still missing after approval of Rec. P.851. In particular, an evaluation on a global level will not provide sufficiently analytical information on the performance of individual system components and on their contribution to overall system quality and acceptability. Methods which are necessary to get a more complete picture of service quality are addressed in Section 6, and it is expected that they can be defined in the respective ITU-T Study Group within the next years.

2. Quality of SDS-Based Services

Applying a definition of quality developed by Jekosch (2000) to the current object of investigation, the *quality of an SDS-based service* is the

> "result of appraisal of the perceived composition of the service with respect to its desired composition." (ITU-T Rec. P.851, 2003)

Thus, the quality perceived by the user is a compromise between what he/she expects or desires, and the characteristics he/she perceives while using the service. It is highly dependent on the situation in which the perception and judgment take place. This fact has to be taken into account when quality is to be quantified in subjective interaction experiments, namely by creating a more-or-less natural test situation and a realistic test user motivation.

Because it is the user who perceives and judges, quality can ultimately only be measured by performing subjective evaluation experiments with human test subjects. General principles for such evaluation experiments are given e.g. in Fraser (1997), and a detailed application example is described in Bernsen et al. (1998). The methods addressed

in this article and in ITU-T Rec. P.851 (2003) follow these principles, and describe the practical implications which result from the chosen definition of quality.

Both *effectiveness* and *efficiency* are related to the performance in achieving the task goal the service has been built for. Effectiveness is an absolute index which describes to what extent the goal was reached, with respect to the accuracy and completeness of the goals, see e.g. ETR 095 (1993):

> "*Effectiveness*: The accuracy and completeness with which specified users can achieve specified goals in particular environments."

Efficiency, on the other hand, is a relative measure of goal achievement in relation to the resources used (ETR 095, 1993):

> "*Efficiency*: The resources expended in relation to the accuracy and completeness of goals achieved."

Effectiveness and efficiency are criteria characterizing a service with which a user is able to achieve his or her task goals. *Usability*, however, is generally defined in a much broader sense. It describes the capability of the service to be understood, learned and used by specified users under specified conditions. It indicates the suitability of the service to fulfil the user's requirements, includes effectiveness and efficiency of the system, and results in user satisfaction (Möller, 2003). *User satisfaction* is an indicator of the service's perceived usefulness and usability for the intended user group. It includes the information whether the user achieves the tasks he/she wants to achieve, is comfortable with the service, and can achieve the task within an acceptable time (Maier et al., 1997).

3. Factors Influencing Service Quality

In the human-machine interaction situation, the transmission channel will be a limiting factor of system performance, and consequently also of the quality experienced by the user. Degradations introduced by the channel include linear frequency distortions, non-linear codec distortions, circuit noise, time-variant degradations from lost frames or packets (in particular in mobile and IP-based networks), as well as background noise picked up by the user's telephone set. These degradations primarily affect the speech recognition and subsequently the speech understanding performance of the system, but they also influence the system's output speech and the interaction as a whole. They have consequently to be taken into account when evaluating quality under realistic circumstances.

Apart from these *environmental factors* (transmission channel, background noise, room acoustic influences), there are a number of other factors which influence different aspects of system and service quality. The major impact will be exercised by the dialogue system and its behavior in the human-machine interaction. The respective factors have been subsumed under the label *agent factors*, and they comprise the behavior of all system components involved in the interaction with the user (speech and potentially speaker recognition, language understanding, dialogue management, response generation, and speech output). The behavior of the resulting system may then be described in terms of the applied dialogue strategy, the flexibility offered by this strategy, as well as the dialogue- and language-related knowledge which is implemented in the system.

Services which are based on SDSs are normally used to carry out a specified task, and the characteristics of the task (*task factors*) are important for the quality and usability of the service. They include the coverage of potential sub-tasks and the domain by the system, and the flexibility offered by the system in resolving the task. In addition, non-physical *contextual factors* may get decisive for the quality and the acceptability of a service. For example, an automatic speech-based system may be preferred over a human operator or a web interface only in case that it is easier accessible, cheaper, or in case of extended opening hours.

Humans are the users of services which are based on SDSs. Thus, factors characterizing the human user (*user factors*) have to be taken into account when the expectations towards the service and the degree of their fulfillment are determined. User factors include the user's background (linguistic background, task/domain knowledge, experience), his motivation and goals, as well other personal factors like attitude or emotions. These factors may directly influence the flow of the interaction, and they will be reflected in the user's judgments on all quality aspects.

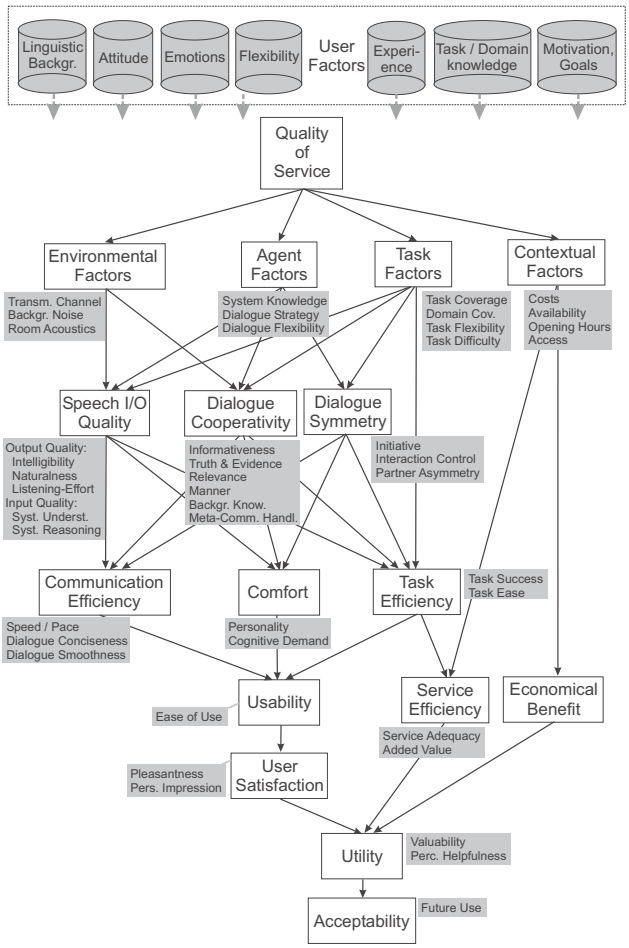

Figure 1: Quality aspects and influencing factors

The mentioned factors have been classified according to a taxonomy which has first been presented in Möller (2002), see Figure 1. Environmental, agent, and task factors carry an influence on the *speech input and output quality*, on the *cooperativity* of system behavior, and on the *symmetry* of

the dialogic interaction. Speech input and output quality includes aspects like intelligibility, naturalness, listening-effort required to understand the system messages, or the perceived system understanding. Cooperativity is defined here in the sense of non-violation of principles for cooperative dialogue behavior, as defined by Grice (1975). It includes the aspects of informativeness, truth and evidence, relevance, manner, background knowledge, and meta-communication handling (i.e. confirmation, clarification, repair and recovery from communication errors), see Bernsen et al. (1998). The partner asymmetry aspect (differences in interaction behavior to be attributed to the asymmetry of the interaction partners) is covered by a category called dialogue symmetry. This category also includes the effects of dialogue initiative and interaction control capabilities.

The mentioned quality aspects result in a (more or less) efficient communication (interaction) and in an efficient solution of the task to be carried out. *Communication efficiency* is related to the speed or pace of the interaction, to dialogue conciseness, and to dialogue smoothness. *Task efficiency*, on the other hand, is linked to task success and task ease. Two additional quality aspects are important: The "personality" of the machine agent (politeness, friendliness, naturalness of behavior) and the effort required from the human user for the interaction (ease of communication, stress/fluster, etc.). These aspects have been subsumed under the term *comfort*.

Communication efficiency, task efficiency and comfort all contribute to the service usability, for which user satisfaction can be seen as an indicator. *Service efficiency*, on the other hand, is influenced by both task efficiency and contextual factors. It is important for the adequacy of the service (for fulfilling the desired task), and for the added value attributed to the service (e.g. in comparison to similar methods for obtaining the same information, like a web interface or a newsticker). Usability, service efficiency, and *economical benefit* result in the *utility* of the service, and finally in its *acceptability*.

4. Subjective Evaluation Experiments

In order to get quantitative information about these quality aspects, guidelines for subjective evaluation methods are given in ITU-T Rec. P.851. They consist in laboratory experiments to be carried out under controlled conditions. In these experiments, test users interact either with a prototype system or with a Wizard-of-Oz simulation (Fraser, 1997). The latter may be necessary in case that the system is not yet fully set up, e.g. for taking decisions on system components during the design process. Wizard-of-Oz simulations may also be used to extrapolate quality for component performances which are beyond the current state-of-the-art. The interactions are usually logged, and interaction parameters can be extracted from the log files with the help of expert annotations. These parameters may be somehow related to the quality aspects, but they are no direct quality measurements (which can only be obtained from the user).

It has been pointed out that user factors may be critical to system quality. Consequently, they have to be taken into account in subjective evaluation experiments. Some of these factors are responsible for the acoustic and linguistic characteristics of the speech produced by the user, namely age and gender, physical status, speaking rate, vocal effort, native language, dialect, or accent. Because these factors may be very critical for the speech recognition and understanding performance, quality judgments obtained from a user group differing in the acoustic and language characteristics might not reflect the quality which can be expected for the target user group. User groups are however variable and ill-defined. A service which is open to the general public will sooner or later be confronted with a large range of different users. Testing with specified users outside the target user group will therefore provide a measure of system robustness with respect to the user characteristics.

A second group of user factors is related to the experience and expertise with the system, the task, and the domain. Several investigations show that user experience affects a large range of speech and dialogue characteristics. For example, it has been reported that users have the tendency to solve more problems per call when they get used to the system, and that the interaction gets shorter (Delogu et al., 1993). Other investigations showed that the number of in-vocabulary utterances increased when the users became familiar with the system. At the same time, the task completion rate increased (Kamm et al., 1997). System familiarity may also lead to a reduced number of user inputs and help messages, and to a reduced transaction time (Lamel et al., 2002).

Because of the lack of a real motivation, laboratory tests often make use of experimental tasks which the subjects have to carry out. The experimental task provides an explicit goal, but this goal should not be confused with a goal which a user would like to reach in a real-life situation. Because of this discrepancy, valid user judgments on system helpfulness and acceptability cannot easily be obtained in a laboratory test set-up. Examples for experimental tasks are included in ITU-T Rec. P.851.

Before the experiment, after each interaction, and at the end of the whole experiment, test subjects have to fill in questionnaires which aim at capturing most of the quality aspects indicated above. Questionnaires given before the experiment usually contain questions related to the user's background, namely some personal information (age, gender, profession, etc.), some task-related information (frequency of the task, usual approach to resolve the task, alternative interfaces, etc.), as well as system-related information (experience with DTMF systems or SDSs).

Questionnaires relating to an individual interaction may cover a number of quality aspects from Figure 1, including:

- Information obtained from the system: Availability, accuracy, completeness, consistency, reliability, clarity, and truth of the obtained information, etc.
- Speech input/output capability: Perceived system understanding, frequency of system errors, perceived system reasoning, listening-effort required to understand the system's messages, perceived intelligibility, perceived comprehensibility, etc.
- System's interaction behavior: Transparency of the interaction, congruence with the user's expectations, flexibility of the interaction, perceived reliability of system processing, distribution of initiative, interaction control capability, confirmation and correction

capabilities, recovery from interaction problems, naturalness of the interaction, length of the dialogue, perceived system speed, smoothness of the dialogue, etc.
- Perceived system personality: Friendliness, politeness, etc.
- Impression on the user: Perceived naturalness of the user's own behavior, pleasantness, cognitive demand put on the user, stress, fluster, etc.
- Perceived task fulfillment: Task success, reliability of task results.

Questionnaires given after the experiment may use similar questions, but the judgments then relate to the overall experience made with the service so far. In this case, very analytic questions should be avoided. Example questionnaires are given in ITU-T Rec. P.851.

5. Evaluation of the New Recommendation

So far, the procedure described in the new Recommendation has been applied to two different systems at the Institute of Communication Acoustics (IKA). The first is an experimental service for restaurant information over the phone, and some results are described in Möller (2003) and Möller and Skowronek (2003). The second is a smart-home system which has been built under the EC-funded IST project INSPIRE. An overview of the evaluation methodology applied for this system is presented in an accompanying paper (Möller et al., 2004). Still, the guidelines and in particular the questionnaires provided in ITU-T Rec. P.851 are far from being fully evaluated. Thus, further application examples of the information contained in the Recommendation are welcome to stabilize the methodology.

6. Discussion and Conclusions

Subjective judgments which are obtained in the way described e.g. in the new Recommendation are the only means to obtain valid information about quality, because quality can ultimately only be judged by the service users. They lack however some analytical power, in that it is not always possible to identify the system component which is responsible for provoking the observed quality percepts.

For this reason, the direct judgments should be amended by collecting interaction parameters, i.e. quantitative indications of system and/or user behavior during the interaction. Such parameters can either be extracted instrumentally, or they rely on a transcription and annotation by an expert. A number of commonly used parameters are summarized in Möller (2003), and they have been classified with respect to the described taxonomy of Figure 1. Correlations between these parameters and subjective judgments are however generally too weak to be able to replace one type of metric by the other. Thus, there is a need for further Recommendations on more complete ways of evaluating telephone-based SDSs. ITU-T Rec. P.851 can only be seen as a first step into this direction.

Acknowledgement

The present work has been performed at IKA (Prof. R. Martin, PD U. Jekosch, Prof. J. Blauert). The development of the taxonomy was partly enabled by the European IST project INSPIRE (IST-2001-32746).

References

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

Delogu, C., di Carlo, A., Sementina, C., Stecconi, S. (1993). A Methodology for Evaluating Human-Machine Spoken Language Interaction. In Proc. 3rd Europ. Conf. on Speech Communication and Technology (Vol. 2, pp. 1427-1430). D-Berlin.

ETSI Technical Report ETR 095 (1993). Human Factors (HF); Guide for Usability Evaluations of Telecommunication Systems and Services. F-Sophia Antipolis: European Telecommunications Standards Institute.

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

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

ITU-T Rec. P.851 (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. CH-Geneva: International Telecommunication Union.

Jekosch, U. (2000). Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitation thesis (unpublished), University/GH, D-Essen.

Kamm, C., Narayanan, S., Dutton, D., Ritenour, R. (1997). Evaluating Spoken Dialogue Systems for Telecommunication Services. In Proc. 5th Europ. Conf. on Speech Communication and Technology (Vol. 4, pp. 2203-2206). GR-Rhodes.

Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H., Temem, J. N. (2002). User Evaluation of the MASK Kiosk. Speech Communication, 38, 131-139.

Maier, E., Mast, A., LuperFoy, S. (1997). Overview. In: Dialogue Processing in Spoken Language Systems. Proc. of the ECAI'96 Workshop, H-Budapest. Lecture Notes in Artificial Intelligence No. 1236 (Maier, E., Mast, M., LuperFoy, S., eds., pp. 1-13). D-Berlin: Springer.

Möller, S. (2003). Quality of Telephone-Based Spoken Dialogue Systems. Habilitation thesis, Institute of Communication Acoustics, Ruhr-University, D-Bochum (to appear).

Möller, S. (2002). A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems. In Proc. 3rd SIGdial Workshop on Discourse and Dialogue (pp. 142-153). USA-Philadelphia PA.

Möller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, A., Schuchardt, D., Fakotakis, N., Ganchev, T., Potamitis, I. (2004). INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control. In Proc. LREC 2004. P-Lisbon.

Möller, S., Skowronek, J. (2003). Quantifying the Impact of System Characteristics on Perceived Quality Dimensions of a Spoken Dialogue Service. In Proc. 8th European Conf. on Speech Communication and Technology (Vol. 3, pp. 1953-1956). CH-Geneva: Int. Speech Com. Ass. ISCA.