A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems

Sebastian Möller
Institute of Communication Acoustics
Ruhr–University Bochum
D–44780 Bochum, Germany
moeller@ika.ruhr-uni-bochum.de

Abstract

This document proposes a new taxonomy for describing the quality of services which are based on spoken dialogue systems (SDSs), and operated via a telephone interface. It is used to classify instrumentally or expert–derived dialogue and system measures, as well as quality features perceived by the user of the service. A comparison is drawn to the quality of human–to–human telephone services, and implications for the development of evaluation frameworks such as PARADISE are discussed.

1 Introduction

Telephone services which rely on spoken dialogue systems (SDSs) have now been introduced at a large scale. For the human user, when dialing the number it is often not completely clear that the agent on the other side will be a machine, and not a human operator. Because of this fact, and because the interaction with the SDS is performed through the same type of user interface (e.g. the handset telephone), comparisons will automatically be drawn to the quality of human–human communication over the same channel, and sometimes with the same purpose. Thus, while acknowledging the differences in behaviors from both — human and machine – sides, it seems justified to take the human telephone interaction (HHI) as one reference for telephone–based human–machine interaction (HMI).

The quality of interactions with spoken dialogue systems is difficult to determine. Whereas structured approaches have been documented on how to design spoken dialogue systems so that they adequately meet the requirements of their users (e.g. by Bernsen et al., 1998), the quality which is perceived when interacting with SDSs is often addressed in an intuitive way. Hone and Graham (2001) describe efforts to determine the underlying dimensions in user quality judgments, by performing a multidimensional analysis on subjective ratings obtained on a large number of different scales. The problem obviously turned out to be multi–dimensional. Nevertheless, many other researchers still try to estimate "overall system quality", "usability" or "user satisfaction" by simply calculating the arithmetic mean over several user ratings on topics as different as perceived TTS quality, perceived system understanding, and expected future use of the system. The reason is the lack of an adequate description of quality dimensions, both with respect to the system design and to the perception of the user.

In this paper, an attempt is made to close this gap. A taxonomy is developed which allows quality dimensions to be classified, and methods for their measurement to be developed. The starting point for this taxonomy was a similar one which has fruitfully been used for the description of human–to–human services in telecommunication networks (e.g. traditional telephony, mobile telephony, or voice over IP), see Möller (2000). Such a taxonomy can be helpful in three respects: (1) system elements which are in the hands of developers, and responsible for specific user perceptions, can be identified, (2) the

dimensions underlying the overall impression of the user can be described, together with adequate (subjective) measurement methods, and (3) prediction models can be developed to estimate quality – as it would be perceived by the user – from purely instrumental measurements. While we are still far from the last point in HMI, examples will be presented of the first two issues.

The next section will discuss what is understood by the term "quality", and will present the taxonomy for HMI. In Section 3, quality features underlying the aspects of the taxonomy are identified, and dialogue- and system-related measures for each aspect are presented in Section 4, based on measures which are commonly documented in literature. Section 5 shows the parallels to the original taxonomy for HHI. The outlook gives implications for the development of evaluation and prediction models, such as the PARADISE framework.

2 Quality of Service Taxonomy

It is obvious that quality is not an entity which could be measured in an easy way, e.g. using a technical instrument. The quality of a service results from the perceptions of its user, in relation to what they expect or desire from the service. In the following, it will thus be made use of the definition of quality developed by Jekosch (2000):

> "Quality is the result of the judgment of a perceived constitution of an entity with regard to its desired constitution. [...] The perceived constitution contains the totality of the features of an entity. For the perceiving person, it is a characteristic of the identity of the entity."

The entity to be judged in our case is the service the user interacts with (through the telephone network), and which is based on a spoken dialogue system. Its quality is a compromise between what s/he expects or desires, and the characteristics s/he perceives while using the service.

At this point, it is useful to differentiate between *quality elements* and *quality features*, as it was also proposed by Jekosch. Whereas the former are system or service characteristics which are in the hands of the designer (and thus can be optimized to reach high quality), the latter are perceptive dimensions forming the overall picture in the mind of the user. Generally, no stable relationship which would be valid for all types of services, users and situations can be established between the two. Evaluation frameworks such as PARADISE establish a temporary relationship, and try to reach some cross–domain validity. Due to the lack of quality elements which can really be manipulated in some way by the designer, however, the framework has to start mostly from dialogue and system measures which cannot be directly controlled. These measures will be listed in Section 4.

The quality of a service (QoS) is often addressed only from the designer side, e.g. in the definition used by the International Telecommunication Union for telephone services (ITU–T Rec. E.800, 1994). It includes service support, operability, security and serveability. Whereas these issues are necessary for a successful set–up of the service, they are not directly perceived by the user. In the following taxonomy, the focus is therefore put on the user side. The overall picture is presented in Figure 1. It illustrates the categories (white boxes) which can be sub–divided into aspects (gray boxes), and their relationships (arrows). As the user is the decision point for each quality aspect, user factors have to be seen in a distributed way over the whole picture. This fact has tentatively been illustrated by the gray cans on the upper side of the taxonomy, but will not be further addressed in this paper. The remaining categories are discussed in the following.

Walker et al. (1997) identified three factors which carry an influence on the performance of SDSs, and which therefore are thought to contribute to its quality perceived by the user: agent factors (mainly related to the dialogue and the system itself), task factors (related to how the SDS captures the task it has been developed for) and environmental factors (e.g. factors related to the acoustic environment and the transmission channel). Because the taxonomy refers to the service as a whole, a fourth point is added here, namely contextual factors such as costs, type of access, or the availability. All four types of factors subsume quality elements which can be expected to carry an influence on the quality perceived by the user. The corresponding quality features are summarized into aspects and categories in the following

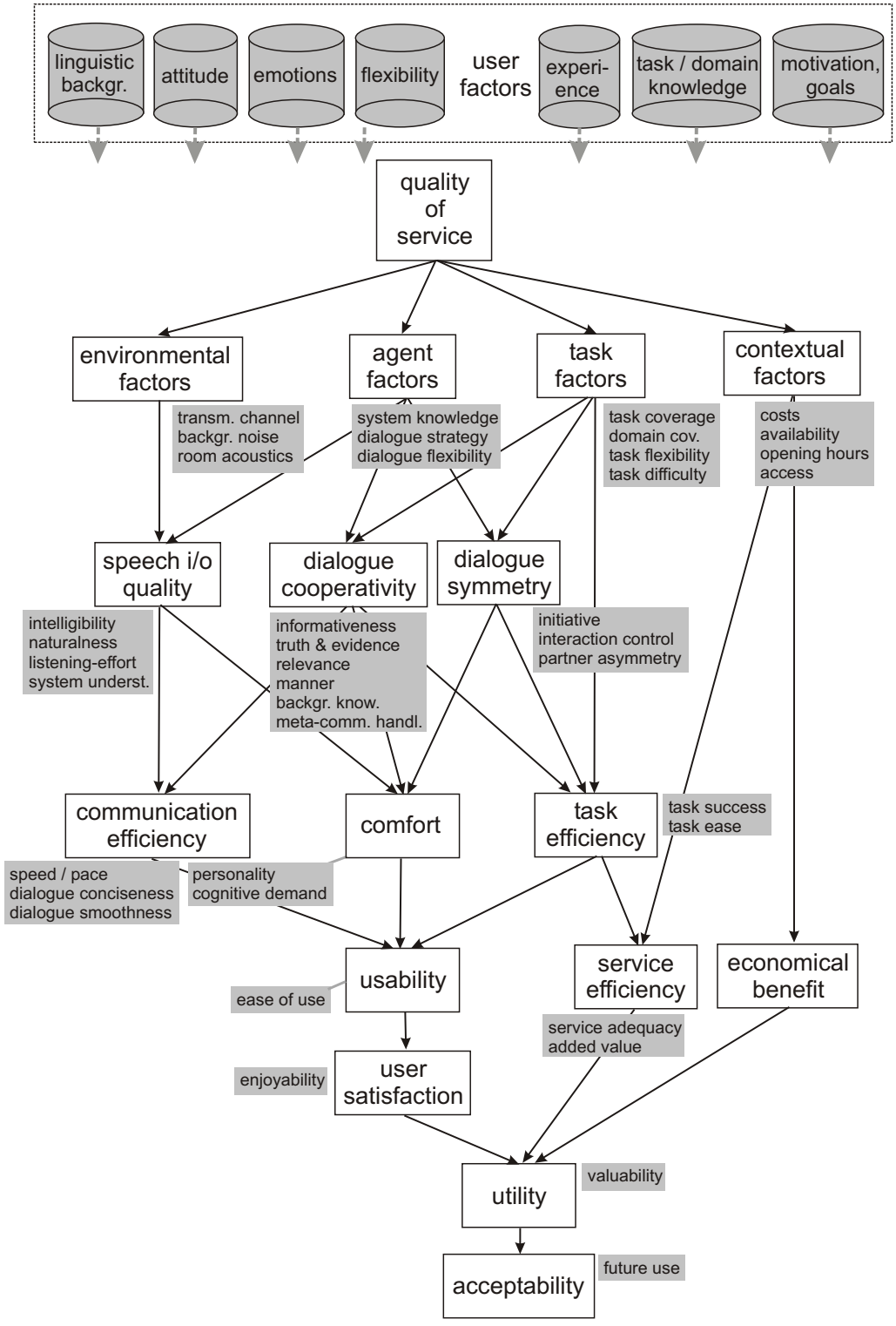

Figure 1: QoS schematic for task–oriented HCI.

lower part of the picture.

The agent factors carry an influence on three quality categories. On the speech level, input and output quality will have a major influence. Quality features for speech output have been largely investigated in the literature, and include e.g. intelligibility, naturalness, or listening–effort. They will depend on the whole system set–up, and on the situation and task the user is confronted with. Quality features related to the speech input from the user (and thus to the system's recognition and understanding capabilities) are far less obvious. They are, in addition, much more difficult to investigate, because the user only receives an indirect feedback on the system's capabilities, namely from the system reactions which are influences by the dialogue as a whole. Both speech input and output are highly influenced by the environmental factors.

On the language and dialogue level, dialogue cooperativity has been identified as a key requirement for high–quality services (Bernsen et al., 1998). The classification of cooperativity into aspects which was proposed by Bernsen et al., and which is related to Grice's maxims (Grice, 1975) of cooperative behavior in HHI, is mainly adopted here, with one exception: we regard the partner asymmetry aspect under a separate category called dialogue symmetry, together with the aspects initiative and interaction control. Dialogue cooperativity will thus cover the aspects informativeness, truth and evidence, relevance, manner, the user's background knowledge, and meta–communication handling strategies.

Adopting the notion of efficiency used by ETSI and ISO (ETSI Technical Report ETR 095, 1993), efficiency designates the effort and resources expanded in relation to the accuracy and completeness with which users can reach specified goals. It is proposed to differentiate three categories of efficiency. Communication efficiency relates to the efficiency of the dialogic interaction, and includes — besides the aspects speed and conciseness — also the smoothness of the dialogue (which is sometimes called "dialogue quality"). Note that this is a significant difference to many other notions of efficiency, which only address the efforts and resources, but not the accuracy and completeness of the goals to be reached. Task efficiency is related to the success of the system in accomplishing the task; it covers task success as well as task ease. Service efficiency is the adequacy of the service as a whole for the purpose defined by the user. It also includes the "added value" which is contributed to the service, e.g. in comparison to other means of information (comparable interfaces or human operators).

In addition to efficiency aspects, other aspects exist which relate to the agent itself, as well as its perception by the user in the dialogic interaction. We subsume these aspects under the category "comfort", although other terms might exist which better describe the according perceptions of the user. Comfort covers the agent's "social personality" (perceived friendliness, politeness, etc.), as well as the cognitive demand required from the user.

Depending on the area of interest, several notions of usability are common. Here, we define usability as the suitability of a system or service to fulfill the user's requirements. It considers mainly the ease of using the system and may result in user satisfaction. It does, however, not cover service efficiency or economical benefit, which carry an influence on the utility (usability in relation to the financial costs and to other contextual factors) of the service. Walker et al. (1998) also state that "user satisfaction ratings [...] have frequently been used in the literature as an external indicator of the usability of an agent." As Kamm and Walker (1997), we assume that user satisfaction is predictive of other system designer objectives, e.g. the willingness to use or pay for a service. Acceptability, which is commonly defined on this more or less "economic" level, can therefore be seen in a relationship to usability and utility. It is a multidimensional property of a service, describing how readily a customer will use the service. The acceptability of a service (AoS) can be represented as the ratio of potential users to the quantity of the target user group, see definitions on AoS adopted by EURESCOM (EURESCOM Project P.807 Deliverable 1, 1998).

From the schematic, it can be seen that a large number of aspects contribute to what can be called communication efficiency, usability or user satisfaction. Several interrelations (and a certain degree of inevitable overlap) exist between the categories and aspects, which are marked by arrows. The interrelations will become more apparent by taking a closer look to the underlying quality features which can be

associated with each aspect. They will be presented in the following section.

3 Classification of Quality Features

In Tables 1 and 2, an overview is given of the quality features underlying each aspect of the QoS taxonomy. For the aspects related to dialogue cooperativity, these aspects partly stem from the design guideline definitions given by Bernsen et al. (1998). For the rest, quality features which have been used in experimental investigations on different types of dialogue systems have been classified. They do not solely refer to telephone–based services, but will be valid for a broader class of systems and services.

By definition, quality features are percepts of the users. They can consequently only be measured by asking users in realistic scenarios, in a subjective way. Several studies with this aim are reported in the literature. The author analyzed 12 such investigations and classified the questions which were asked to the users (as far as they have been reported) according to the quality features. For each aspect given in Tables 1 and 2, at least two questions could be identified which addressed this aspect. This classification cannot be reproduced here for space reasons. Additional features of the questionnaires directly address user satisfaction (e.g. perceived satisfaction, degree of enjoyment, user happiness, system likability, degree of frustration or irritation) and acceptability (perceived acceptability, willingness to use the system in the future).

From the classification, it seems that the taxonomy adequately covers what researchers intuitively would include in questionnaires investigating usability, user satisfaction and acceptability.

4 Classification of Dialogue and System Measures

Experiments with human subjects are still the only way to investigate quality percepts. They are, however, time–consuming and expensive to carry out. For the developers of SDSs, it is therefore interesting to identify quality elements which are in their hands, and which can be used for enhancing the quality for the user. Unfortunately, only few such elements are known, and their influence on service quality is only partly understood. Word accuracy or word error rate, which are common measures to describe the performance of speech recognizers, can be taken as an example. Although they can be measured partly instrumentally (provided that an agreed–upon corpus with reference transcriptions exists), and the system designer can tune the system to increase the word accuracy, it cannot be determined beforehand how this will affect system usability or user satisfaction.

For filling this gap, dialogue– and system–related measures have been developed. They can be determined during the users' experimental interaction with the system or from log–files, either instrumentally (e.g. dialogue duration) or by an expert evaluator (e.g. contextual appropriateness). Although they provide useful information on the perceived quality of the service, there is no general relationship between one or several such measures, and specific quality features. The PARADISE framework (Walker et al., 1997) produces such a relationship for a specific scenario, using multivariate linear regression. Some generalizablility can be reached, but the exact form of the relationship and its constituting input parameters have to be established for each system anew.

A generalization across systems and services might be easier if a categorization of dialogue and system measures can be reached. Tables 3 and 4 in the Appendix report on the classification of 37 different measures defined in literature into the QoS taxonomy. No measures have been found so far which directly relate to speech output quality, agent personality, service efficiency, usability, or user satisfaction. With the exception of the first aspect, it may however be assumed that they will be addressed by a combination of the measures related to the underlying aspects.

5 Comparison to Human-Human Services

It has been stated earlier that the QoS taxonomy for telephone–based spoken dialogue services has been derived from an earlier schematic addressing human–to–human telephone services (Möller, 2000). This schematic is depicted in Figure 2, with slight modifications on the labels of single categories from the original version.

In the HHI case, the focus is placed on the categories of speech communication. This category (re-

Table 1: Dialogue–related quality features.

	Aspect	Quality Features
Dialogue Cooperativity	Informativeness	– Accuracy / Specificity of Information – Completeness of Information – Clarity of Information – Conciseness of Information – System Feedback Adequacy
	Truth and Evidence	– Credibility of Information – Consistency of Information – Reliability of Information – Perceived System Reasoning
	Relevance	– System Feedback Adequacy – Perceived System Understanding – Perceived System Reasoning – Naturalness of Interaction
	Manner	– Clarity / Non–Ambiguity of Expression – Consistency of Expression – Conciseness of Expression – Transparency of Interaction – Order of Interaction
	Background Knowledge	– Congruence with User's Task/Domain Knowl. – Congruence with User Experience – Suitability of User Adaptation – Inference Adequacy – Interaction Guidance
	Meta–Comm. Handling	– Repair Handling Adequacy – Clarification Handling Adequacy – Help Capability – Repetition Capability
Dialogue Symmetry	Initiative	– Flexibility of Interaction – Interaction Guidance – Naturalness of Interaction
	Interaction Control	– Perceived Control Capability – Barge–In Capability – Cancel Capability
	Partner Asymmetry	– Transparency of Interaction – Transparency of Task / Domain Coverage – Interaction Guidance – Naturalness of Interaction – Cognitive Demand Required from the User – Respect of Natural Information Packages
Speech I/O Quality	Speech Output Quality	– Intelligibility – Naturalness of Speech – Listening–Effort Required from the User
	Speech Input Quality	– Perceived System Understanding – Perceived System Reasoning

Table 2: Communication–, task– and service–related quality features.

	Aspect	Quality Features
Communic. Efficiency	Speed	– Perceived Interaction Pace – Perceived Response Time
	Conciseness	– Perceived Interaction Length – Perceived Interaction Duration
	Smoothness	– System Feedback Adequacy – Perceived System Understanding – Perceived System Reasoning – Repair Handling Adequacy – Clarification Handling Adequacy – Naturalness of Interaction – Interaction Guidance – Transparency of Interaction – Congruence with User Experience
Comfort	Agent Personality	– Politeness – Friendliness – Naturalness of Behavior
	Cognitive Demand	– Ease of Communication – Concentration Required from the User – Stress / Fluster
Task Efficiency	Task Success	– Adequacy of Task / Domain Coverage – Validity of Task Results – Precision of Task Results – Reliability of Task Results
	Task Ease	– Perceived Helpfulness – Task Guidance – Transparency of Task / Domain Coverage
Service Efficiency	Service Adequacy	– Access Adequacy – Availability – Modality Adequacy – Task Adequacy – Perceived Service Functionality – Perceived Usefulness
	Added Value	– Service Improvement – Comparable Interface
Usability	Ease of Use	– Service Operability – Service Understandability – Service Learnability

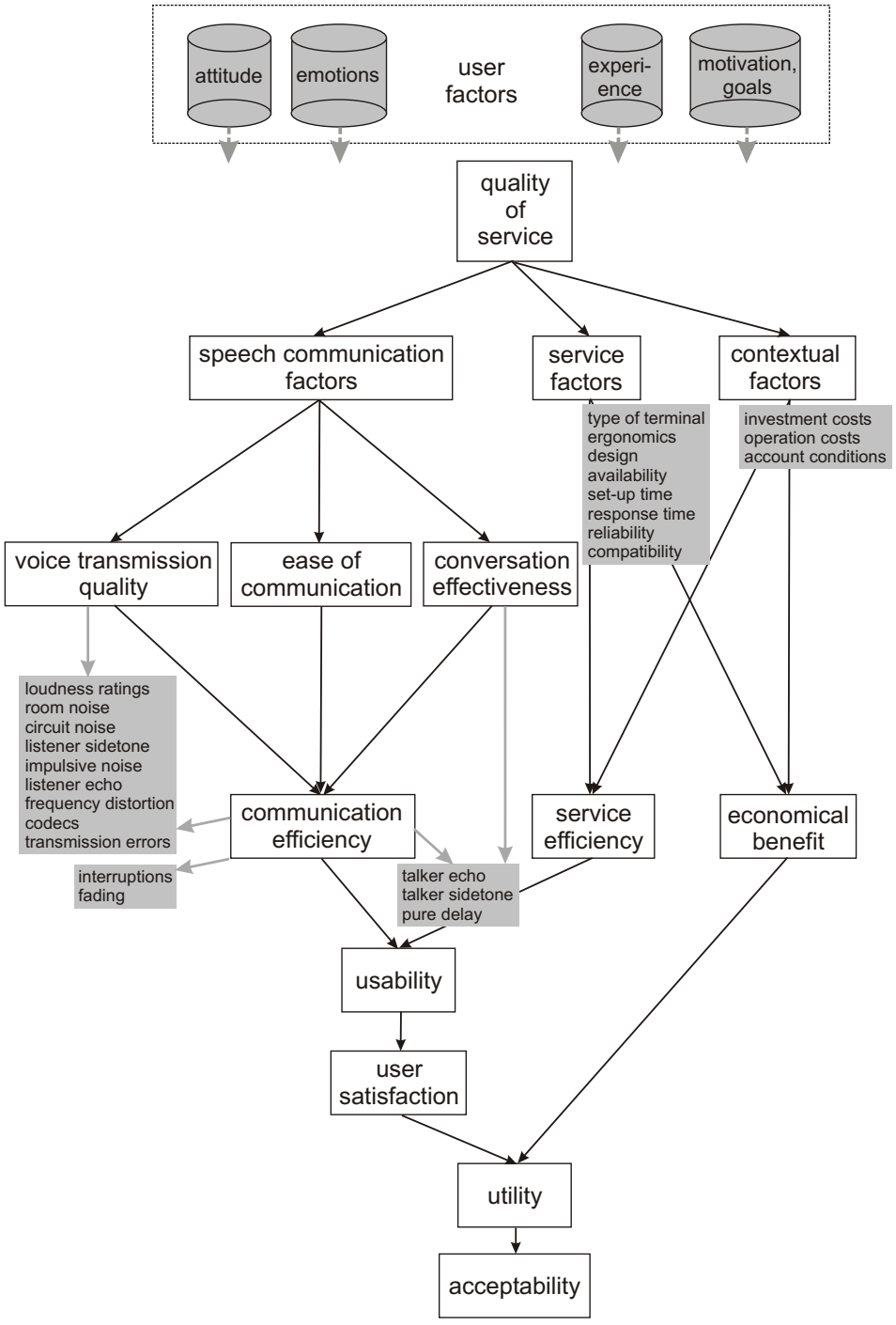

Figure 2: QoS schematic for human–to–human telephone services.

placing environmental and agent factors of the HMI case) is divided into a one–way voice transmission category, a conversational category (conversation effectiveness), and a user–related category (ease of communication; comparable to the category "comfort" in the HMI case). The task and service categories of the interaction with the SDS are replaced by the service categories of the HHI schematic. The rest of the schematic is congruent in both cases, although the single aspects which are covered by each category obviously differ.

The taxonomy of Figure 2 has fruitfully been used to classify three types of entities:

- quality elements which are used for the set–up and planning of telephone networks (some of these elements are given in the gray boxes of Figure 2)
- assessment methods commonly used for measuring quality features in telecommunications
- quality prediction models which estimate single quality features from the results of instrumental measurements

Although we seem to be far from reaching a comparable level in the assessment and prediction of HMI quality issues, it is hoped that the taxonomy of Figure 1 can be equally useful with respect to telephone services based on SDSs.

6 Discussion and Conclusions

The new taxonomy was shown to be useful in classifying quality features (dimensions of human quality perception) as well as instrumentally or expert–derived measures which are related to service quality, usability, and acceptability. Nonetheless, in both cases it has not been validated against experimental (empirical) data. Thus, one cannot guarantee that the space of quality dimensions is captured in an accurate and complete way.

There are a number of facts reported in literature, which make us confident that the taxonomy nevertheless captures general assumptions and trends. First of all, in his review of both subjective evaluations as well as dialogue- or system-related measures, the author didn't encounter items which would not be covered by the schematic. This literature review is still going on, and it is hoped that more detailed data can be presented in the near future.

As stated above, the separation of environmental, agent and task factors was motivated by Walker et al. (1997). The same categories appear in the characterization of spoken dialogue systems given by Fraser (1997) (plus an additional user factor, which obviously is nested in the quality aspects due to the fact that it is the user who decides on quality). The context factor is also recognized by Dybkjær and Bernsen (2000). Dialogue cooperativity is a category which is based on a relatively sophisticated theoretical as well as empirical background. It has proven useful especially in the system design and set–up phase, and first results in evaluation have also been reported (Bernsen et al., 1998). The dialogue symmetry category captures the remaining partner asymmetry aspect, and has been designed separately to additionally cover initiative and interaction control aspects. To the authors knowledge, no similar category has been reported. The relationship between the different efficiency measures and usability, user satisfaction and utility was already discussed in Section 2.

In the PARADISE framework, user satisfaction is composed of maximal task success and minimal dialogue costs (Walker et al., 1997), — thus a type of efficiency in the way it was defined here. This concept is still congruent with the proposed taxonomy. On the other hand, the separation into "efficiency measures" and "quality measures" (same figure) does not seem to be fine–graded enough. It is proposed that the taxonomy could be used to classify different measures beforehand. Based on the categories, a multi–level prediction model could be envisaged, first summarizing similar measures (belonging to the same category) into intermediate indices, and then combining the contributions of different indices into an estimation of user satisfaction. The reference for user satisfaction, however, cannot be a simple arithmetic mean of the subjective ratings in different categories. Appropriate questionnaires still have to be developed, and they will take profit of multidimensional analyses as reported by Hone and Graham (2001).

References

Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1998. *Designing Interactive Speech Systems: From First Ideas to User Testing*. Springer, D–Berlin.

Morena Danieli and Elisabetta Gerbino. 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, USA–Stanford CA*, pages 34–39, AAAI Press, USA–Menlo Park CA.

Laila Dybkjær and Niels Ole Bernsen. 2000. Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6(3-4):243–271.

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

EURESCOM Project P.807 Deliverable 1, 1998. *Jupiter II - Usability, Performability and Interoperability Trials in Europe*. European Institute for Research and Strategic Studies in Telecommunications, D–Heidelberg.

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

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

Kate S. Hone and Robert Graham. 2001. Subjective Assessment of Speech–System Interface Usability. *Proc. 7th Europ. Conf. on Speech Communication and Technology (EUROSPEECH 2001 – Scandinavia)*, pages 2083–2086, DK–Aalborg.

ITU–T Rec. E.800, 1994. *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability*. International Telecommunication Union, CH–Geneva, August.

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

Candance A. Kamm and Marilyn A. Walker. 1997. Design and Evaluation of Spoken Dialogue Systems. *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, USA–Santa Barbara (CA)*, pages 14–17.

Candance Kamm, Shrikanth Narayanan, Dawn Dutton, and Russell Ritenour. 1997. Evaluating Spoken Dialogue Systems for Telecommunication Services. *Proc. 5th Europ. Conf. on Speech Communication and Technology (EUROSPEECH'97)*, 4:2203–2206, GR–Rhodes.

Diane J. Litman, Shimei Pan, and Marilyn A. Walker. 1998. Evaluating Response Strategies in a Web–Based Spoken Dialogue Agent. *Proc. of the 36th Ann. Meeting of the Assoc. for Computational Linguistics and 17th Int. Conf. on Computational Linguistics (COLING-ACL 98), CAN-Montreal*.

Sebastian Möller. 2000. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publ., USA–Boston.

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue. 1992. Experiments in Evaluating Interactive Spoken Language Systems. *In: Proc. DARPA Speech and Natural Language Workshop*, pages 28–33.

Patti J. Price, Lynette Hirschman, Elizabeth Shriberg, and Elizabeth Wade. 1992. Subject–Based Evaluation Measures for Interactive Spoken Language Systems. *In: Proc. DARPA Speech and Natural Language Workshop*, pages 34–39.

Andrew Simpson and Norman M. Fraser. 1993. Black Box and Glass Box Evaluation of the SUNDIAL System. *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2:1423–1426, D–Berlin.

Helmer Strik, Catia Cucchiarini, and Judith M. Kessens. 2001. Comparing the Performance of Two CSRs: How to Determine the Significance Level of the Differences. *Proc. 7th Europ. Conf. on Speech Communication and Technology (EUROSPEECH 2001 – Scandinavia)*, pages 2091–2094, DK–Aalborg.

Marilyn A. Walker, Diane J. Litman, Candance A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, pages 271–280.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1998. Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. *Computer Speech and Language*, 12(3).

A Classification of Dialogue and System Measures

Table 3: Classification of measures (1). #: average number of ... per dialogue. For references, see caption of Table 4.

	Aspect	Dialogue / System Measure
Dialogue Cooperativity		– CA: contextual appropriateness (SF93, F97)
	Informativeness	– # user questions (P92) – # help requests from the user (W98)
	Truth and Evidence	– # questions correctly/incorrectly/partially/failed to be answered (P92) – DARPA score, DARPA weighted error (P92)
	Relevance	– # barge-in attempts from the user (W98)
	Manner	– # system turns (W98) – no. of words per system turn
	Background Knowledge	– # help requests (W98) – # cancel attempts from the user (W98) – # barge-in attempts from the user (W98) – # time-out prompts (W98)
	Meta–Comm. Handling	– # system error messages (Pr92) – # help requests (W98) – # cancel attempts from the user (W98) – CR: correction rate (SCR) (F97, SF93) – IR: implicit recovery (DG95)
Dialogue Symmetry	Initiative	– # user questions (P92) – # system questions – CR: correction rate (SCR, UCR) (F97, SF93)
	Interaction Control	– # barge-in attempts from the user (W98) – # help requests (W98) – # cancel attempts from the user (W98) – CR: correction rate (UCR) (F97, SF93) – # time-out prompts (W98)
	Partner Asymmetry	– # barge-in attempts from the user (W98) – # time-out prompts (W98)
Speech I/O Quality	Speech Output Quality	–
	Speech Input Quality	– word accuracy, word error rate (SF93) – sentence accuracy, sentence error rate (SF93) – number or errors per sentence (S01) – word error per sentence (S01) – HC_{U1}, HC_{U2}, HC_{S1}, HC_{S2} (K97) – UER: understanding error rate – # ASR rejections (W98) – IC: information content (SF93) – # system error messages (Pr92)

Table 4: Classification of measures (2). #: average number of ... per dialogue. References: DG95: Danieli and Gerbino (1995); F97: Fraser (1997); K97: Kamm et al. (1997); P92: Polifroni et al. (1992); Pr92: Price et al. (1992); SF93: Simpson and Fraser (1993); S01: Strik et al. (2001); W98: Walker et al. (1998).

	Aspect	Dialogue / System Measure
Communic. Efficiency	Speed	– TD: turn duration (STD, UTD) (F97) – SRD: system response delay (Pr92) – URD: user response delay (Pr92) – # timeout prompts (W98) – # barge-in attempts from the user (W98)
	Conciseness (Litman et al., 1998: dialogue efficiency)	– DD: dialogue duration (F97, P92) – # turns (# system turns, # user turns) (W98)
	Smoothness (Litman et al., 1998: dialogue quality)	– # system error messages (Pr92) – # cancel attempts from the user (W98) – # help requests (W98) – # ASR rejections (W98) – # barge-in attempts from the user (W98) – # timeout prompts (W98)
Comfort	Agent Personality	–
	Cognitive Demand	– # timeout prompts (W98) – URD: user response delay (Pr92)
Task Efficiency	Task Success	– TS: task success (DG95, F97, SF93) – κ: kappa coefficient (W98) – task solution (P92) – solution correctness (P92) – solution quality
	Task Ease	– # help requests (W98)
Service Efficiency	Service Adequacy	–
	Added Value	–
Usability	Ease of Use	–
User Satisfaction		–