Quantifying the Impact of System Characteristics on Perceived Quality Dimensions of a Spoken Dialogue Service

Sebastian Möller, Janto Skowronek

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

Abstract

Developers of telephone services which are relying on spoken dialogue systems would like to identify system characteristics influencing the quality perceived by the user, and to quantify the respective impact before the system is put into service. A laboratory experiment is described in which speech input, speech output, and confirmation characteristics of a restaurant information system were manipulated in a controlled way. Users' quality perceptions were collected by means of a specifically designed questionnaire. It is based on a recently developed taxonomy of quality aspects, and aims at capturing a multitude of perceptually relevant quality dimensions. Experimental results indicate that ASR performance affects a number of interaction parameters, and is a relatively well identifiable quality impact for the user. In contrast, speech output affects perceived quality on a number of different levels, up to global user satisfaction judgments. Potential reasons for these findings are discussed.

1. Introduction

When spoken dialogue systems (SDSs) are used in real-life application scenarios, system components which operate with the acoustic realization of spoken language are of primary importance. On the one hand, the degree to which the user feels to be understood by the system will depend on the performance of speech recognition and language understanding components. On the other hand, the speech output component may largely affect quality dimensions which relate to the system utterances. In between these two interface components, the dialogue manager will be responsible for the overall interaction behavior of the system, and hereby significantly influence the quality perceived by the user.

Although there are a number of comparative evaluations of different system versions reported in literature, only little information is available that would explicitly quantify the impact of the named components. The lack of data will be partly linked to the relatively complex interaction of system components of a SDS, which makes it difficult to pick out individual characteristics in a controlled way. In addition, most investigations try to quantify the overall impact, in terms of an integral quality or usability judgment. Without denying the usefulness of such integral judgments for the system designer, it is important to first address the question *which* quality dimensions perceived by the user are impacted, and then to quantify how these quality dimensions impact overall quality.

The present paper describes a Wizard-of-Oz (WoZ) experiment which has been designed in order to investigate the impact of speech recognition performance, confirmation strategy, and speech output strategy on different quality dimensions perceived by the user, and on overall quality judgments. A prototype system for providing information about restaurants in the area of Bochum, Germany, has been taken as an example (Bochum Restaurant Information System, BoRIS). This system is integrated in a test environment which facilitates the collection of a large number of interaction parameters during the experiment. In order to obtain analytic descriptions of quality perceived by the user, a specific questionnaire has been designed. It is based on theoretical considerations which are illustrated by a taxonomy of Quality of Service (QoS) aspects, see Section 2. During the experiment, test subjects communicated with several system versions differing with respect to the named system characteristics. The dialogue system and test set-up are described in Section 3. Results will be analyzed with respect to the impact of speech recognition (Section 4), speech output (Section 5), and confirmation strategy (Section 6), both on the interaction parameter values and on the quality judgments given by the test subjects.

2. Quality of Spoken Dialogue Services

The quality of an SDS-based service is determined by the perceptions of its users. It turns out as the result of a perception and a judgment process, in which the perceiving subject establishes a relationship between the perceptive event, and what s/he expects or desires from the service. This user-centric point of view is reflected in the definition of quality given by Jekosch [1]:

> "Result of appraisal of perceived constitution of a unit with respect to its desired composition."

Because it is the user who judges on quality, user factors like attitude, emotions, experience, task/domain knowledge, etc., will determine the perception of individual quality aspects, and of global aspects like usability, user satisfaction or acceptability.

In an earlier publication, Möller [2] organized the most relevant quality aspects in terms of a taxonomy showing the influencing factors on quality, the aspects they carry an influence on, and the interrelationship of aspects. Apart from the user factors, four types of factors were identified: *Agent factors* describing the characteristics of the machine agent as an interaction partner; *environmental factors* covering the physical (acoustical) environment of the user, including any transmission channels involved in the interaction; *task factors* resulting from the task which can be carried out with the help of the SDS-based service; and *contextual factors* describing the non-physical context of use (e.g. price, opening hours, comparable interfaces). The taxonomy is shown in Figure 1.

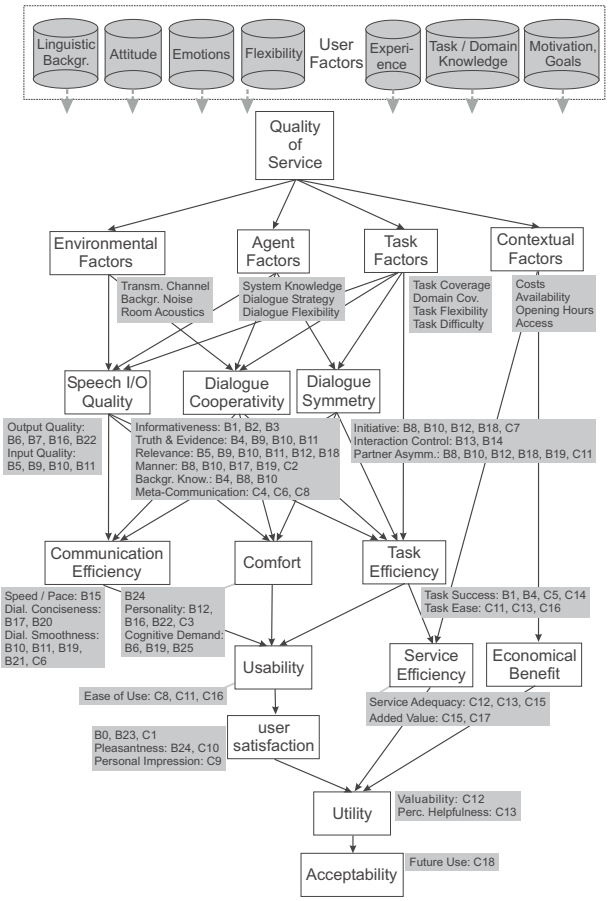

Figure 1: QoS Taxonomy. B*n* and C*n* refer to the questions of Part B and C of the questionnaire.

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 [3]. Although these principles have been developed with the focus on human-to-human interaction and are not meant to be strictly adopted in each conversation, they have been successfully applied in the design and evaluation of spoken dialogue systems [4]. Cooperativity includes the aspects informativeness, truth and evidence, relevance, manner, background knowledge, and meta-communication handling, see Bernsen and Dybkjær [4]. The partner asymmetry aspect (differences in interaction behavior to be attributed to the asymmetry of the interaction partners) has been split into a new 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 have to be catered for: 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* here.

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 ways for obtaining the same information). Usability, service efficiency, and *economical benefit* result in *utility* of the service, and finally in its *acceptability*.

Two types of information can be collected in order to describe the mentioned quality aspects. On the one hand, subjective judgments obtained from the users – usually via questionnaires with pre-defined questions – are direct descriptors of perceived quality dimensions. On the other hand, system- and dialogue-related parameters can be logged automatically during the interaction. With the help of expert transcriptions and annotations, interaction parameters can be extracted from the log files. These parameters cover a range of system performance measures, but they are no direct indicators of quality.

The taxonomy of QoS aspects can be used to efficiently set up evaluation schemes, involving both subjective judgments and instrumentally or expert-derived interaction parameters. A large number of interaction parameters has been classified in the QoS taxonomy, see [2]. The majority of these parameters will be extracted during the experiment. For measuring the perceptive quality dimensions, a questionnaire has been developed. It consists of three parts which are compiled by the test subjects before the test starts (Part A), after each interaction with the dialogue system (Part B), and after the whole test session (Part C). A copy of the questionnaire is available from the author. The questions can be assigned to the different categories of the QoS taxonomy, as it is indicated in Figure 1. Because Parts A and C refer to the test session as a whole and not to the individual system configuration, they are not discussed in the following.

3. BoRIS Dialogue System and Test Set-Up

For the experiment, a spoken dialogue system providing information about restaurants in Bochum and its surroundings has been set up at IKA. The system is able to search for restaurants with respect to five criteria: Type of food, location of the restaurant, price range, day of the week, and the time the user wants to eat out (opening hours of the restaurant).

At the speech input side, the system either uses a commercial speech recognizer, or a recognizer simulation. The simulation is based on the transcription of a human wizard, on which controlled "recognition errors" are generated according to a previously measured confusion matrix. The confusion matrix has been determined for the target vocabulary of the recognizer, and it is scaled in an exponential way in order to simulate arbitrary recognition rates. Dialogue management is implemented as a finite state machine, using the CSLU tool-

kit [5]. Different confirmation strategies can be selected: Explicit or implicit confirmation, summarizing confirmation at the end of the information-gathering part of the dialogue, or no confirmation at all. Speech output is implemented either via pre-recorded messages from two non-professional speakers (1 male, 1 female), or via a text-to-speech (TTS) system. Both speech output options can be combined, using one option for the fixed system messages, and another for the variable restaurant information parts. TTS consists of the symbolic pre-processing unit SyRUB and the synthesizer IKAphon [6], concatenating units of different length which have been recorded from a professional male speaker. The individual system modules can be combined in order to generate system configurations which differ in well-defined characteristics. Table 1 shows the list of system configurations used in the experiment.

Table 1: System configurations used in the experiment.

No.	Recog. Rate [%]	Speech output		Confir-mation
		fixed	variable	
1	100	female	female	--
2	100	male	female	--
3	100	female	TTS (male)	--
4	100	TTS (male)	TTS (male)	--
5	70	female	female	--
6	100	female	female	explicit
7	90	female	female	explicit
8	80	female	female	explicit
9	70	female	female	explicit
10	60	female	female	explicit

40 subjects (11 f, 29 m) participated in the test. They were between 23 and 51 years old, and were paid for their service. The majority of subjects did not have any experience with spoken dialogue services, but most of them knew the town of Bochum and some of the local restaurants.

Test subjects communicated with the spoken dialogue system over the phone. At the beginning of the experiment, they had to compile Part A of the questionnaire. According to experimental tasks which have been described in terms of open or closed scenarios, the subjects then had to carry out five interactions with the system, and write down the restaurants which were found by the system. Finally, they judged their impression after the whole interaction experience.

During each interaction, a log file was produced by the system. This file has been annotated by an expert using a specifically designed Tcl/Tk tool. The tool extracts a large number of interaction parameters related to speech input (word accuracy, word error rate, concept accuracy, understanding error rate, parsing errors), cooperativity (contextual appropriateness, no. of user/system questions, correctness of system answers, DARPA measures) meta-communication (help requests, ASR rejections, system error messages, barge-ins, cancel attempts, system/user correction turns, implicit recovery), the overall dialogue and communication situation (dialogue duration, system/user turn duration, no. of system/user turns, words per system/user turn), and task success (binary and ordinal task success, kappa coefficient).

4. Impact of ASR Performance

The simulated recognition rate (configurations 6-10 in Table 1) shows a statistically significant influence on all speech-input-related parameters (ASR and speech understanding performance measures). This is obvious, because the recognition performance is the variable parameter. On the other side, the system and user correction rate, the cooperativity of system utterances, and the task success (kappa) are also significantly impacted. These parameters are linked to the cooperativity category (contextual appropriateness is a direct measure of cooperativity, and the correction rate is related to the system's meta-communication capabilities), to the communication efficiency (correction rates), and to the task success categories.

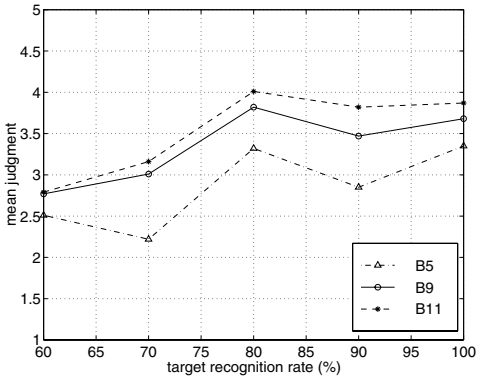

Figure 2: Impact of simulated recognition rate.

The effect on the subjective judgments is depicted in Fig. 2. Questions B5 ("How well did you feel understood by the system?") and B11 ("How often did the system make mistakes?") were significantly affected, and B9 ("In your opinion, the system processed your specifications correctly...incorrectly.") close to the significance level (Kruskal-Wallis test). Apparently, the test subjects were able to localize the source of interaction problems relatively well. Subjective ratings are relatively constant down to a recognition rate of approx. 80%; below this threshold they decrease significantly.

5. Impact of Speech Output

Only two interaction parameters are affected by the choice of the speech output component: The system turn duration (probably due to the slow speaking rate of the TTS system), and the response delay of the user. The latter finding shows the high cognitive demand put on the user. In contrast to the restricted effects on interaction parameters, subjective ratings are affected quite drastically, see Figures 3 and 4. Besides the obvious effects on listening-effort (B6: "You had to concentrate in order to understand what the system expected from you.") and intelligibility (B7: "How well was the system acoustically intelligible?"), also the clarity of the information (B3: "The information was clear...unclear."), the naturalness (B12: "The system reacted in the same way as humans do."; B18: "You perceived the dialogue as natural...unnatural."; B22: The system's voice was natural...unnatural."), the friendliness (B16: "The system reacted in a friendly...unfriendly way."), the smoothness (B21: "The course of the dialogue was smooth...bumpy."), the

pleasantness (B24: "You perceived the dialogue as pleasant...unpleasant."), stress (B25: "During the dialogue, you felt stressed...relaxed."), and the overall impression of the system (question B0) are strongly affected. The perceptive degradation is significantly higher than it was observed for the speech recognition configurations.

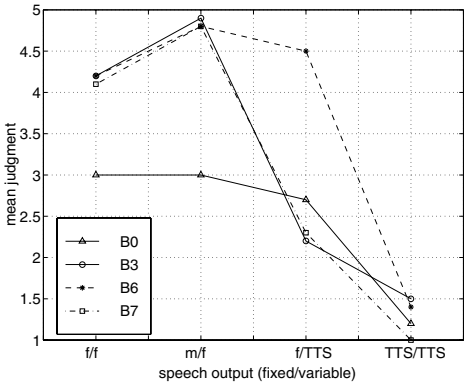

Figure 3: Impact of speech output configurations. f: female speaker; m: male speaker; TTS: male TTS voice.

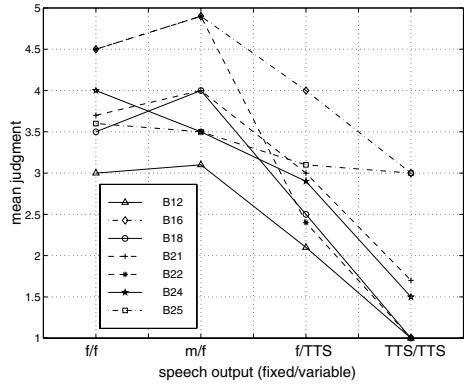

Figure 4: Impact of speech output configurations. f: female speaker; m: male speaker; TTS: male TTS voice.

6. Impact of Confirmation Strategies

The system configuration which differ with respect to the confirmation strategy are 1 vs. 6 (perfect recognition) and 5 vs. 9 (70% recognition rate), see Table 1. Using the explicit confirmation strategy, the number of system questions raised in both cases. With perfect recognition, also the implicit recovery rate significantly increased, whereas with lower recognition performance an additional decrease of the system turn duration, the system response delay, and the words per system turn could be observed (effects linked to the specific confirmation strategy implementation). Task success could in both cases not be increased by using confirmation. The reason might be that the confirmation was only applied in the information-gathering parts of the dialogue, and not in the final navigation through the system responses which seems to cause many errors.

Interestingly, the effect on the subjective judgments is relatively low. Only two questions showed a *negative* impact of applying the explicit confirmation strategy, and only for the 70% recognition rate: B6 (concentration required) and B8 ("You knew at each point of the dialogue what the system expected from you."). Apparently, transparency suffers from an inappropriately chosen confirmation strategy when the error rate is high.

7. Conclusions and Outlook

On both the speech input and the speech output side, significant influences on interaction parameters and subjective quality ratings were observed. Whereas the recognition rate mainly influenced interaction parameters related to the categories speech input quality, cooperativity, communication efficiency and task success, synthesized speech strongly affects perceptual dimensions which can be associated with different categories of the QoS taxonomy, down to the global overall impression of the user. Apparently, the system voice describes the system's "personality", and seems to act as a kind of "business card" of the system. This effect has been postulated earlier and could now be quantified with our experiment. Users seem to be less able to localize their perceptions with respect to speech output than to speech input. The confirmation strategy only showed weak effects on both interaction parameters and subjective quality judgments.

In order to test whether the findings are generic, it will be important to repeat the experiments with other dialogue systems. The taxonomy of QoS aspects proved to be a very useful tool for test design and result interpretation, and it will consequently be used in these future experiments. Because interaction parameters and subjective ratings are collected simultaneously under controlled conditions, it becomes possible to analyze and develop quality prediction models. This is a topic of ongoing work at IKA.

8. Acknowledgements

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

9. References

[1] Jekosch, U. *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, Habilitation thesis, Univ./GH, D-Essen, 2000.

[2] Möller, S. *A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems*, Proc. 3rd SIGdial Workshop, USA-Philadelphia PA, 2002, 142-153.

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

[4] Bernsen, N. O., Dybkjær, H., Dybkjær, L. *Designing Interactive Speech Systems: From First Ideas to User Testing*, Springer, D-Berlin.

[5] Sutton, S., Cole, R., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, M., Cohen, M. *Universal Speech Tools: The CSLU Toolkit*, Proc. ICSLP'98, AUS-Sydney, 3221-3224.

[6] Köster, S. *Modellierung von Sprechweisen für widrige Kommunikationsbedingungen mit Anwendung auf die Sprachsynthese*, Shaker Verlag, D-Aachen.