

# Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method

*Mikko Hartikainen, Esa-Pekka Salonen and Markku Turunen*

Speech-based and Pervasive Interaction Group, Tampere Unit for Computer-Human Interaction  
Department of Computer Sciences, University of Tampere, Finland  
{mhartikainen, eps, mturunen}@cs.uta.fi

## Abstract

There is demand for subjective metrics in spoken dialogue system evaluation. SERVQUAL is a service quality evaluation method developed by marketing academics. It produces a subjective measure of the gap between expectations and perceptions in five service quality dimensions common for all services. We present how the method was applied to spoken dialogue system evaluation. In order to improve the suitability of the original method, we modified the test questionnaire and the test process. We demonstrate how the modified method was successfully used in an evaluation of a telephone-based e-mail application. The evaluation gave us directions for further development of the system. In addition, we found some interesting phenomena, such as the variation between genders. We present how the method can be further improved, for example, by dividing the questionnaire into two parts.

## 1. Introduction

The current methods for subjective evaluation of Spoken Dialogue Systems (SDSs) leave room for improvement [1]. While some subjective SDS evaluation methods are introduced [1, 2, 3, 4], none of them has been widely adopted by developers, who often choose between objective evaluation and no evaluation.

Subjective evaluation has numerous advantages over objective evaluation. First, subjective measures are the only way to find out the user's opinion of the SDS. Second, there are cases when there is no choice, if, for example, the interaction deals with sensitive data such as medical records or personal correspondence. Third, it is cost effective to skip the annotation and dialogue analysis part of evaluation – although objective evaluation is complementary to subjective evaluation. Fourth, the analysis of a straightforward subjective questionnaire requires no special skills.

Moreover, subjective measures are highly relevant with SDSs. There are many reasons for this. Most importantly, objective measures, such as performance, do not have as direct a link to the user satisfaction as with graphical user interfaces. For example, users may prefer system initiative dialogues over more efficient user initiative dialogues [5].

We see SDSs as services – it is hard to think of a practical SDS which is not a service. For example, most of the commercially launched SDSs are services that provide information similar to human operated services. Yet the service evaluation research has been mostly ignored by developers of subjective SDS evaluation methods. That is why we applied a service evaluation method to SDS evaluation.

Based on an extensive survey of service evaluation methods, we found the SERVQUAL method suitable for SDS evaluation. It is popular, taught in marketing courses, referred to in the marketing literature, and used in a variety of businesses in several branches [6]. It is generic and therefore easy to apply to a new field [7]. Since it has been designed to provide information that can be used when further developing the service, it is useful for iterative development of SDS development as well. As an example of its diagnostic power, the results could imply that it is more critical to concentrate on security issues than on performance issues. Such results are vital for making an SDS more acceptable.

Subjective measures based on SERVQUAL were used to evaluate the speech-based Dutch transport information system VIOS [8]. Unfortunately, the modifications to the method were only weakly justified. We applied SERVQUAL to SDS evaluation without major modifications.

We present how SERVQUAL was used to evaluate AthosMail [9], a phone based e-mail reading system, which was built in the EU-funded DUMAS project (IST-2000-29452). In order to use the method, we varied the test process and made necessary modifications to the test questionnaire.

Section 2 introduces the SERVQUAL method and its main principles. The third section presents our modifications to the method. In Section 4 we present results from a concrete SDS evaluation. Finally, we consider our findings in a general discussion of SDS evaluation.

## 2. The SERVQUAL Method

The marketing academics Parasuraman, Zeithaml and Berry developed and iterated the SERVQUAL method based on two principles: Service quality can be divided into dimensions, and measured as a difference of expectations and perceptions [7]. They defined five service quality dimensions.

Next we describe the SERVQUAL method. First we present the service quality dimensions. From the developer's point of view there are two tasks: collection of data and analysis of data. We present the SERVQUAL questionnaire and the analysis methods respectively.

### 2.1. Service quality dimensions

Service quality dimensions are *tangibles*, *reliability*, *responsiveness*, *assurance* and *empathy*. Based on an extensive number of audits, dimensions are claimed to be universal, i.e. applicable to any service [7, 10]. They can be measured with 21 [6] or 22 items [10]. However, authors agree that modifications to items may be necessary when using the method to evaluate different services [7]. We focus on the 22 item version, not to discard any factors. Descriptions of dimensions are presented in Table 1.

Dimension	Items	Description
<b>Tangibles</b>	4	Appearance of physical facilities, equipment, personnel, and other materials.
<b>Reliability</b>	5	Ability to perform the promised service dependably and accurately.
<b>Responsiveness</b>	4	Willingness to help customers and provide prompt service.
<b>Assurance</b>	4	Employees' knowledge and courtesy and their ability to inspire trust and confidence.
<b>Empathy</b>	5	Caring, individualised attention given to customers.

Table 1: Service quality dimensions [6].

Items are statements, such as “service is fast”. Customers rate how well their expectations of this kind of service meet their perceptions of this particular service. Next we look more closely at giving ratings with the SERVQUAL questionnaire.

## 2.2. The SERVQUAL Questionnaire

The questionnaire can have one, two or three columns. The number of columns means how many questions respondent needs to answer. In the three column format a separate rating is asked for the acceptable level of quality, the desired level, and the actual perceived quality [11]. Rating can be done using any scale (e.g. 1...3, 1...7, 1...9). Concrete examples of the appearance of the questionnaire and templates can be found from the Internet appendix of this paper [12].

We found the three column format most useful, since it gives more descriptive data. One and two column formats give no insight whether expectations were low or high. Such data is useful, for example to find out which service quality dimension users find important. The mere exceeding of desired level means that quality is superior to expectations but this is not necessarily informative, if user expectations are low. In the next subsection we present the ways to analyse data collected with a three column format of SERVQUAL.

## 2.3. Analysing SERVQUAL Data

SERVQUAL authors propose some ways to analyse the data [6]. Some simply describe the gap between expectations and perceptions: One produces disconfirmation measures and others produce graphical presentations, such as the zone of tolerance. Otherwise, evaluations can also be plotted into an importance-performance grid. Analyses can be done per (i) each item, (ii) each dimension or (iii) as an overall measure.

Disconfirmation measures are the Measure of Service Superiority (MSS) and the Measure of Service Adequacy (MSA). The measures are computed by simple extraction formulas presented in Equation 1. An equivalent analysis is to present ratings in a line chart, which shows the relation of expectations and perceptions [6].

$$\begin{aligned} \text{MSS} &= \text{Perceived level} - \text{Desired level} \\ \text{MSA} &= \text{Perceived level} - \text{Accepted level} \end{aligned}$$

Equation 1: Disconfirmation measures.

The zone of tolerance presents perceptions relation to a zone based on expectations. The zone of tolerance presentation consists of a line and a dot, as depicted in Figure 1. In this example, the user has found the service to be slow. The presentation shows exactly what the expectations are and how perceptions correspond to them; in the example, perception is clearly under the acceptable level.

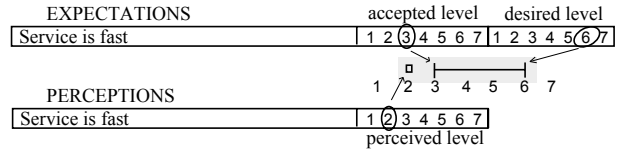


Figure 1: Construction of the zone of tolerance presentation.

Importance-performance grid and examples are presented in Section 4, where we analysed the data from a case study. Next we present the modifications made to the method in order to make it suitable for SDS evaluation.

## 3. Modified SERVQUAL Method

Some modifications were needed to apply the method to SDS evaluation. First, the questionnaire was modified to better correspond to the services provided by a machine which you cannot see but only hear. Second, when the evaluation takes place prior to the launching of a service, e.g. during a testing of a prototype system, it is meaningful to divide the questionnaire into two parts, expectations and perceptions.

### 3.1. Modified questionnaire

As discussed in subsection 2.1, individual items may be revised to better suit the service in question. Such revisions are needed in the case of SDS evaluation, too.

Items in the dimension *tangibles* need to be adapted for all SDSs. For example, item 3 was originally “employees have a neat, professional appearance”. We changed it to describe a more essential tangible in phone-based SDSs, “the service has a pleasant voice”.

The main reason for modifications was that statements about personnel had to be changed to statements about the system when evaluating SDSs. Modifications were made so that the items still represent the five dimensions of service quality. The original and the modified questionnaire can be found from the Internet appendix of this paper [12].

### 3.2. Evaluation process

Traditionally, service quality is evaluated after the launching of the service. Otherwise, service providers would be aware of the testing circumstances and try to provide better service than in the real situation. However, contrary to the reality in service businesses, SDS development usually includes several evaluation phases before finalising the product.

To make the most of evaluation, we propose dividing SERVQUAL questionnaire into two parts, so that expectations are collected before using the SDS and perceptions are collected afterwards. This way, expectations are not influenced by experiences from that particular system.

The process is depicted in Figure 2. There are two actors, the tester (which can mean several people) and the participant. There are three steps. In each step the tester presents an artefact (e.g., a questionnaire or a system) to the participant, and the actions of the participant lead to a result (e.g., a completed questionnaire or system data).

In the first step expectations are collected. The questionnaire contains the *accepted level* and the *desired level*. In the second step participant uses the system. It can be a fully implemented version, a partly working prototype, a Wizard of Oz setup or just a design presented with print-outs or recordings. In the third step, perceptions are collected.

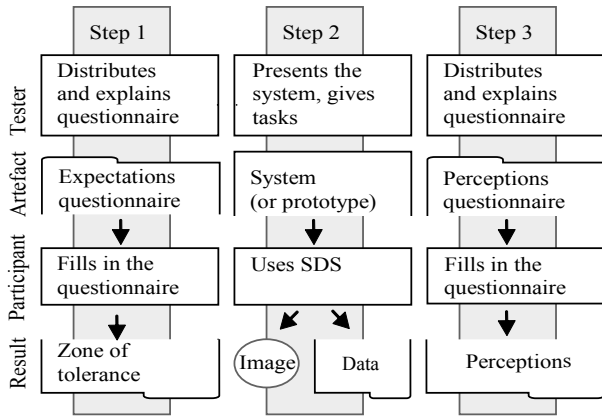


Figure 2: Process of using SERVQUAL in a test situation.

Next we present a case where we applied the SERVQUAL method to evaluate AthosMail, a phone-based e-mail reading spoken dialogue system in a task-based test situation.

#### 4. Case Study: AthosMail

AthosMail is a multilingual telephone-based e-mail application [9]. Inputs are given mostly by speech, with the exceptions of entering user codes and navigating in the message with DTMF keys. The system responds to natural language inputs, like “What messages do I have from Mary?” The system is built on top of robust and adaptive architecture [13] and it has a flexible dialogue management scheme [14]. AthosMail contains adaptive features, e.g. it groups messages, uses sender names dynamically and gives more help for novice users. It is personalised, e.g. it knows the user’s name and preferences. As a part of the development process we evaluated the system using the modified SERVQUAL method. Next we present the test setup and the main results from the viewpoint of the SERVQUAL method.

##### 4.1. Test setup

We evaluated a Finnish configuration of the AthosMail system using a task-based test setup. Each user used the system on two consecutive days. On both days they had two tasks, each with a different mailbox. Each task was a separate dialogue in a separate phonecall. The first day they practised with another SDS, a local bus timetable system. Users were given a description of AthosMail. Expectations were asked at the beginning of both days. Perceptions were collected after each task. All in all, we analysed ratings from 84 calls.

##### 4.2. Test results

Overall, the users were satisfied with the AthosMail system. Disconfirmation measures show how the mean of perceptions was not under the mean of accepted level in any stage of usage. However, the zone of tolerance presentation and importance-performance grid indicate differences between genders and quality dimensions, respectively.

###### 4.2.1. Disconfirmation measures

We computed disconfirmation measures (see Section 2.3) for each item, each dimension and an overall score. As Table 2 illustrates, overall MSS scores were negative and MSA scores were positive throughout the experiment. In other words, users found the system acceptable, yet not above the desired

level. Normally this would mean that the system is ready to be deployed, but the artificial task-based setup may have influenced the answers.

	After 1 <sup>st</sup>	After 2 <sup>nd</sup>	After 3 <sup>rd</sup>	After 4 <sup>th</sup>
MSS	-0,92	-0,81	-1,06	-1,03
MSA	0,71	0,81	0,57	0,60

Table 2: Overall disconfirmation measures after each task.

There were no deviances in the measures per dimensions either. Only on the item level some cases stood out. After the two first tasks, the item “non-speech audio” was found superior (MSS 0,09 and 0,09), and item “brochure” was found inferior (MSA -0,13 and -0,09). After subsequent tasks these cases blended in with other tolerant results, but we find it easy to believe that the audio signals really helped users to know when to speak, especially at the beginning, and that our black-and-white text-only brochure was not too attractive.

###### 4.2.2. Zone of tolerance

The zone of tolerance indicated differences in system acceptance between female and male users. In general, as Figure 3 shows, males had wider zones of tolerance.

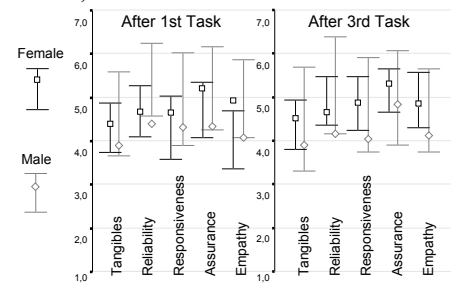


Figure 3: Male and female zones of tolerance.

Acceptance varied especially on dimensions *reliability* and *empathy*. While males found the system narrowly acceptable or not acceptable, females found the system clearly within or above the zone of tolerance. After the third task the situation evened out a little, mainly due to changes in expectations. It is possible that the experiences from using the system affected to the ratings of expectations. In any case, *reliability* and *assurance* were the most important dimensions.

###### 4.2.3. Importance-performance grid

The information about importance and performance can be presented by plotting the items in a grid. An example from AthosMail evaluation is depicted in Figure 4.

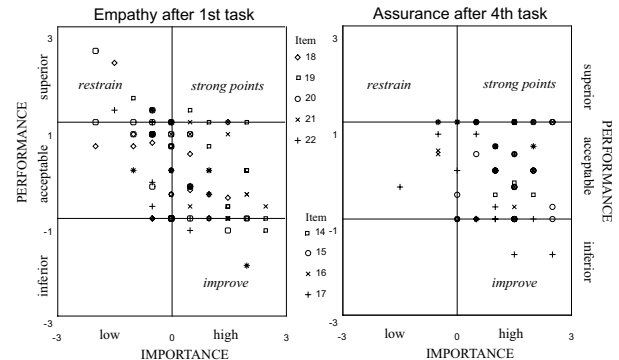


Figure 4: Importance-performance grids.

SERVQUAL authors present a 2 by 2 grid [6] but we extended performance into three degrees. We used a 2 by 3 grid in order to maintain correspondence between performance-scale and the zone of tolerance, i.e. the cases within accepted level (see Figure 4) are also within the zone of tolerance. The example indicates that assurance is more important than empathy. At its strongest, the grid can point out also the factors that should be improved or restrained and the factors that are the strong points of the system.

## 5. Discussion

In general, we found the SERVQUAL method applicable for SDS evaluation. The methods strong points are utilising service quality dimensions and the measuring of the difference between expectations and perceptions. However, the method should be further adapted for SDS domain. Our proposals for improvement concern count and content of the items, and questionnaire styles. We will also study different ways for analysing data, for example by further developing the importance-performance grid.

To get one round of ratings, a respondent needs to answer 66 questions. Not all users are motivated to accurately answer so many questions. Items may be found awkward, too. These difficulties can lead to hasty estimates.

We suggest two changes to the questionnaire. First, it should be more flexible, for example by allowing the count of items to be varied. We believe that even one or two items per dimension are adequate for certain evaluations. When necessary, item count can be increased to provide more detailed evaluation. Likewise there are cases when only some dimensions need to be evaluated. Second, the dimensions and hence the items should be revised to better suit SDS evaluation. The current items evaluate service quality on a detailed level; nevertheless they say little about factors that affect the quality. For example, the item "politeness" could be divided to specific prompt design issues. Such factors are studied previously [1, 2, 3, 4], therefore the question is of integrating those findings into the SERVQUAL method.

We believe that questionnaires could be edited to make evaluation more fluent and effective. Effectiveness can be increased with automated data collection. For optional questionnaire styles, we propose more intuitive and graphical ways of expressing opinions. For example, the zone of tolerance could be drawn as a line and the perceptions as dots. We will also study how to utilise the importance-performance grid in the questionnaire formatting.

Finally, SERVQUAL should not be a restriction of any sort. In general, service quality knowledge and methods could be applied more extensively for versatile SDS evaluation and development.

## 6. Conclusion

We have presented how a valid service evaluation method, SERVQUAL, can be applied for the subjective evaluation of spoken dialogue systems. We modified the test questionnaire and the test process to improve the suitability of the method in the area of speech applications. We demonstrated how the method was used in the evaluation of a telephone-based e-mail application. In order to make the method more usable, we discussed how it could be improved. The major challenge, and at same time the greatest potential, is to make the method

more flexible and to integrate its strong points with other subjective ways of evaluating SDSs.

## 7. References

- [1] Larsen, L.B., "Assessment of spoken dialogue system usability – what are we really measuring", *EUROSPEECH 2003 Proc.*, Geneva 2003.
- [2] Möller, S. and Skowronek, J., "Quantifying the impact of system characteristics on perceived quality dimensions of a spoken dialogue service", *EUROSPEECH 2003 Proc.*, Geneva 2003.
- [3] Walker, M.A., Litman, D.J., Kamm, C.A. and Abella, A. "PARADISE: A framework for evaluating spoken dialogue agents", *Proc. of the 35th Annual Meeting of the Association of Computational Linguistics*, 1997.
- [4] Hone, K. S. and Graham, R., "Towards a tool for the subjective assessment of speech system interfaces (SASSI)", *Natural Language Engineering*, 6, 3-4, Cambridge University Press 2000.
- [5] Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C., and Hindle, D., "What can I say?: evaluating a spoken language interface to Email", *SIGCHI Conference on Human Factors in Computing Systems Proc.*, 1998.
- [6] Zeithaml, V.A. and Bitner, M.J. *Services marketing: integrating customer focus across the firm*, 3<sup>rd</sup> ed., McGraw-Hill, 2003. p. 93, 135-155.
- [7] Parasuraman, A., Zeithaml, V.A. and Berry, L.L., "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality", *Journal of Retailing*, 64, 1, 1988.
- [8] van Haaren, L., Blasband, M., Gerritsen, M., and van Schijndel, M. "Evaluating quality of spoken dialogue systems: comparing a technology-focused and a user-focused approach", *First International Conference on Language Resources & Evaluation Proc.*, Granada 1998.
- [9] Turunen M., Salonen, E.-P., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M., Olsson, F., and Sahlgren, M., "AthosMail – a Multilingual Adaptive Spoken Dialogue System for E-mail Domain", *Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces Proc.*, Geneva, 2004.
- [10] Parasuram, A., Berry, L.L., Zeithaml, V.A.: "Refinement and Reassessment of the SERVQUAL Scale", *Journal of Retailing*, 67, 4, 1991.
- [11] Parasuraman, A., Zeithaml, V.A. and Berry, L.L., "Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria", *Journal of Retailing*, 70, 3, 1994.
- [12] Internet appendix available in the address <http://www.cs.uta.fi/hci/spi/SERVQUAL/>
- [13] Turunen, M., Salonen, E.-P., Hartikainen, M. and Hakulinen, J., "Robust and adaptive architecture for multilingual spoken dialogue systems", *ICSLP 2004 Proc.*, Jeju 2004.
- [14] Salonen, E.-P., Hartikainen, M., Turunen, M., Hakulinen, J. and Funk, J. A., "Flexible Dialogue Management Using Distributed and Dynamic Dialogue Control", *ICSLP 2004 Proc.*, Jeju 2004.