Subjective Sound Quality Assessment of Mobile Phones for Production Support

T. Drascher
Siemens AG, ICM MP PD TI 4
Bocholt, Germany
E-mail: thorsten.drascher@siemens.com

M. Schultes
Institut für Mikrosystemtechnik
University Siegen, Germany
E-mail: martin-schultes@gmx.net

Workshop on Wideband Speech Quality in Terminals and Networks:
Assessment and Prediction
8th and 9th June 2004, Mainz, Germany

Abstract

Subjective perception of sound quality is an essential criterion in terminal assessment.
To guarantee consistent quality over terminal generations is a great challenge in day-to-day business. Normally each algorithm and feature is optimized on its own before a long lasting field test of the terminal as a whole is started. The aim of the work presented here is to develop a test protocol for testing mobile devices within a few days and to have available a first, overall quality assessment just before the release for unrestricted serial production.
The tests are carried out as conversational tests for single talk and double talk. Statistical relevant results can be generated within a few days by using short conversation test scenarios.
Test results are very sensitive to the questions proposed, regarding both number and content, and to the involved test subjects. Test subjects themselves are subdivided twofold: naive (or every day) users, who will mainly purchase the terminal at the end, will only have to answer up to three general questions; and expert users, who know exactly what to listen to and therefore are interrogated in more detail about one specific topic.

1 Introduction

Although there is no highly sophisticated definition, each person has his or her own intuitive feeling of quality. From a general point of view quality can be defined as *degree of satisfaction of a user* [1] or as *result of a judgement of a perceived constitution of an entity with regard to the desired constitution* [2]. Due to the high abstractness of those definitions, quality measurement is very difficult. Network quality can be foreseen approximately by using the E-model, afterwards quality measurements are performed by a standardised method, e.g. PESQ [4]. Using these techniques the quality perceived subjectively can be reproduced in satisfying approximation.
Development of a model for describing subjective end-to-end quality is in the focus of ITU-SG12. The first step in this development process is to collect data in subjective tests, which, after evaluation, can be used to build a theoretical framework for an objective quality estimation of subjective perceptions. Also for terminal manufacturers quality is a very important benchmark, since this might be the final criteria influencing the customer's decision of purchase. Hence it is in the manufacturer's focus of interest to get a feedback as soon as possible about the quality of his products.
This paper will concentrate on presenting how to obtain statistically significant results in a very short time.
The manufacturer's point of view is given in section 2, before the experimental setup is described in more detail in section 3, in which also the need for in-situ tests will be discussed. How to choose questions is the focus of section 4. A first shot of the developed test environment was presented on an internal fair and will be discussed briefly in section 5. In the last section a conclusion and a possible future works are presented.

2 Ancillary Conditions

After each terminal component and algorithm is tested separately, terminals are handed out to some test subjects for field tests. The results of these tests, typically a long term test protocol or periodically filled out questionnaires, are questionable

due to friendly subjects and owing to only a small number of tests the statistics of the results are not reliable. Although during the test period, sometimes several weeks, the final terminal performance is only estimated, the production is ongoing.
Manufacturers are therefore interested in obtaining statistically reliable results in a short time by neutral test subjects. An obvious time optimisation can be achieved by executing only the minimum number of tests required, which can be calculated by the error of the mean using basic statistics. The worst case can be approximated for a maximum standard deviation. The standard deviation is maximised if only supreme and inferior ratings are given with relative frequency of 50 In this worst case assumption, execution of about 30 tests guarantees therefore an error of the mean which will not exceed 10 % of the interval width of the ratings used.
Further time optimisations can be achieved by splitting tests into single talk, double talk, and comparison listening tests. Most phone conversations are single talk conversations, so the main part of the tests consists of single talk tests. Scenarios for these tests are proposed by the ITU [7] and by Moeller and Wiegelmann [5, 6]. These so called short conversation test scenarios (SCTs), whose general procedure is presented in figure 1, simulate some typical short conversation as e.g. booking a flight or ordering a pizza.
The fact that most test subjects regard the SCTs as natural and their shortness in time of about 2:30 min [5] lead to the conclusion that the most effective way of having single talk conversation tests are the SCTs.
A drawback of SCTs is a minor part of double talk, which is tested for this reason in a different test. As suggested by the ITU [8] a slightly different text is handed out to two subject. One will read the text aloud, the other in silence. If the second subject becomes aware of a difference in his text to the one read by the first, he immediately starts reading his own version aloud, till a common text passage is reached.
Single and double talk tests should be executed by naive listeners, who represent the largest customer group. If the tested terminal does not fulfil test criteria defined in advance, there will be a more detailed view on the terminal's most annoying properties.
For this more detailed view expert listeners are of more importance than naive listeners. To get information about an altered parameter set as fast as possible, comparison listening tests are carried out. Due to the fact that the two groups of tests are disjoint, comparison tests can only be used to evaluate a tendency, but will in no case give information about the impact of parameter alteration on the results of the naive user tests.

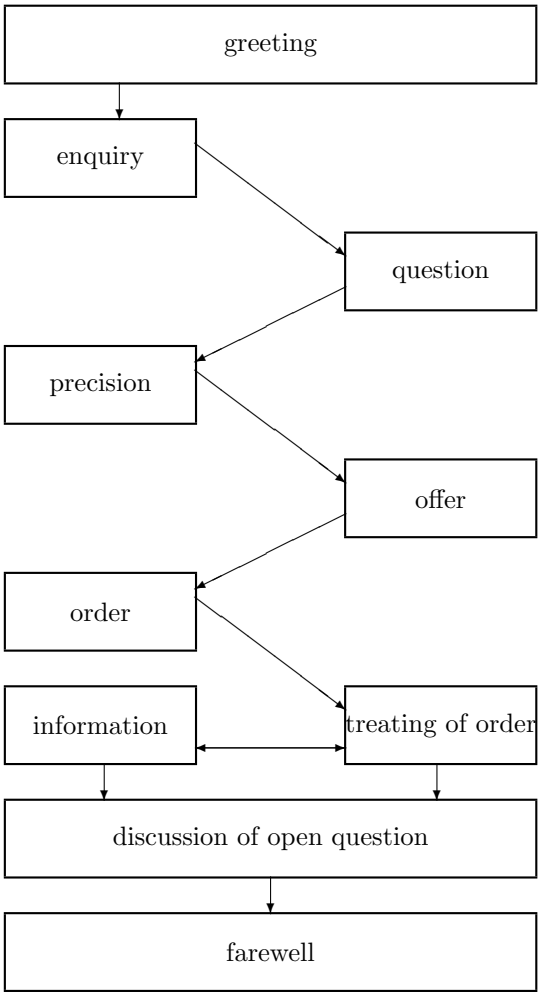

Figure 1: General procedure resulting from SCTs.

3 Environmental Conditions

In each test there will be one fixed network terminal in a silent environment, preferably a silent room as recommended and defined by the ITU [10] used as fixed reference point. The terminal in the silent room is referred to as terminal A, the mobile terminal under test is referred to as terminal B.
Due to their usage, it is mandatory to test mobile terminals in different environments, as e.g. in a crowd of people (babble noise), in train or car (car kit use only!), or a quiet office or home environment.
On the one hand there are less expensive laboratory tests under well known conditions, where reproducibility is guaranteed. On the other hand there are more expensive in-situ tests, which do not

require environmental simulation, and which, consequently, have higher reliability. Former tests of voice recognisers showed well accordance of laboratory and in-situ tests concerning recognition rates [9]. Because of slighltly different test conditions at least one in-situ test will be necessary for adjustment.
For both, in-situ and laboratory tests, there are lots of degrees of freedom, mainly microphone placing and sound level measurement of background noise, especially when testing car kits [11]. Mobile terminals will be held intuitively correct by test subjects, but for car kit tests a special setup is required. For symmetry purposes the microphone is placed in the front middle of the car dashboard, because the subject does not steer the car himself during in-situ tests and best possible comparability with laboratory tests is required. Typically background noise levels in cars are about 60 dB(A) at a velocity of 50 km/h, just below 70 dB(A) for 100 km/h, and just above 70 dB(A) for 130 km/h. Different levels of different noises can be applied in the laboratory as diffuse sound field, surrounding the test subject's head and the terminal or car kit microphone.
Mobile terminals are tested in babble noise, street noise, and - for reference purposes - white noise, car kits in natural compact car noise and artificial car noise [10]. Each background noise is used in at least three different levels, and to minimize network effects at least two different networks shall be involved in the tests. The number of terminal and car kit tests, N_t and N_c, are given by:

$$N_t = 2 \cdot (1 + 3 \cdot 3) \cdot 30 = 600$$
$$N_c = 2 \cdot (1 + 2 \cdot 3) \cdot 30 = 420$$

Using SCTs and assuming that answering the following questions will increase test duration to about 4 min, the whole series will last about 4100 min. Working each day around 10 h will allow the test series to be finished in two weeks, even in one week if two terminals can be tested in parallel.
It is recommended, that concentration periods should last about 20 min and in no case should exceed 45 min [10]. Assuming that after a break test subjects will be able to handle two concentration periods of 20 min each, about 200 test subjects are required in one test series.

4 The Questions

Questions (e.g. from [8, 10]) for test evaluation have to be selected case sensitive for different tests and test subjects, due to different designs and purposes. Naive users are able to answer up to three questions and only a few more can be answered by expert users. Too many questions will introduce a randomisation effect, especially for those asked at the end.
Naive user tests concentrate on questions concerning overall quality, while in tests for expert users, the focus is on a special performance task of the terminal. Technical terms are to be avoided in predefined answers of naive user tests, because of their expected lack of background knowledge in this field. A well known fact is the avoidance of extreme values by the test subjects. This means the five point MOS-scale (MOS: Mean Opinion Score) is effectively a little more than a three point scale. The whole range is used more effectively, when answers are presented on a linear scale, where also fractional and ratings beyond scale limitations ratings are allowed. An example of such a scale is presented in figure 2.

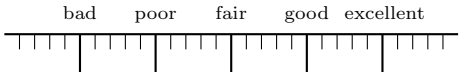

Figure 2: Continuous MOS-scale, which allows also fractional ratings and lowers the weighting of extreme ratings.

5 First Test Presentation

During an internal fair in Munich at the beginning of May the proposed test environment was presented for the first time. The test was carried out computer based using a Tcl/Tk script. Under test was a low end terminal, which development has been concluded and which was known for being too silent. It has to be stressed, that neither the test subjects nor the experimenter knew this fact. As background noise the environmental babble noise was used. The sound level was only estimated to be approximately 70 dB(A). People passing by were asked to carry out a test. The conversational task was booking a flight. The far end was served by a student using a fixed network phone. Since no expert user was expected, the naive user test was carried out. Its interview consisted of only two questions.
The first question was about the perceived overall quality. As a result a variation of the mean opinion score (MOS_v) was obtained. The answer was marked on scale [5] ranging from 0 to 120 where the MOS-labels bad, poor, fair, good, and excellent were placed at 20, 40, 60, 80, and 100, respectively. A slider was placed at "fair" at the beginning by

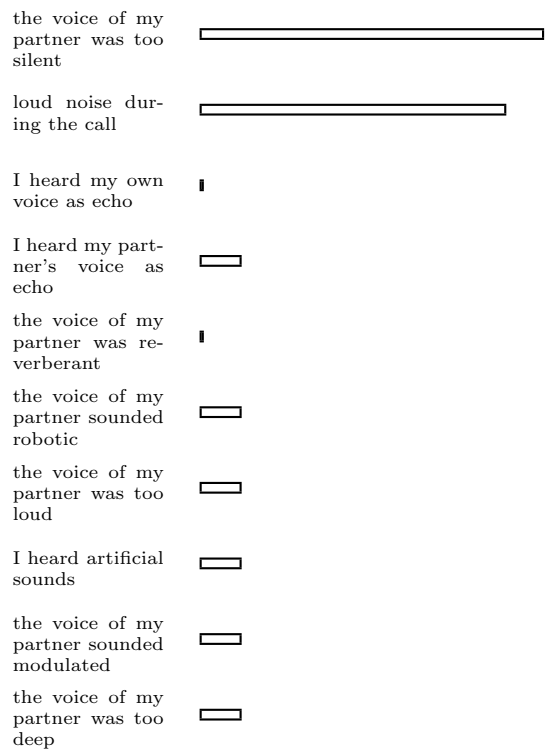

Figure 3: Most recognised annoyances. In this context "noise" had a double meaning: Most people considered the environmental noise as most annoying factor, only one heard noise over the connection. The last two annoying factors were added by test subjects during the test.

the script.
During this test 19 test subjects were interviewed. The mean overall quality was assigned to

$$MOS_v = 74 \pm 4$$

The 19 ratings are listed in detail in table 1. Shown are the ratings on the continuous scale from 0 to 120, and calculated MOS ratings under two different assumptions:

1. The labels displayed on the continuous scale of figure 2 lie in the middle of the subjective rating intervals. This means e.g. values from 50 to 69 are assigned to fair (MOS-rating 3)

2. The labels on the continuous scale of figure 2 lie at the lower rim of the subjective rating intervals. This means e.g. values from 60 to 79 are assigned to fair (MOS-rating 3).

In assumption two, maximum rating were only given twice, while in assumption one, it was given five times, which means that every fourth test person rated with maximum score. Since this is very unusual in former MOS ratings, regarding the labels as lower margins of a rating interval tends to be more reliable when scale ratings are compare to classic MOS. This can be seen more clearly, when averages of these three scales are calculated.

$$\begin{aligned} \langle MOS_v \rangle &= 74.16 \quad (62\,\% \text{ of interval width}) \\ \langle MOS_v \rangle &= \ 3.79 \quad (70\,\% \text{ of interval width}) \\ \langle MOS_v \rangle &= \ 3.32 \quad (58\,\% \text{ of interval width}) \end{aligned}$$

Due to placing the slider in the middle of the scale at the beginning of the test, 60 was the most given rate. In future tests the start position will be given randomly or on the first click.

TS	Rating	MOS_{centered}	MOS_{lowered}
1	38	2	1
2	103	5	5
3	95	5	4
4	60	3	3
5	60	3	3
6	82	4	4
7	81	4	4
8	60	3	3
9	67	3	3
10	72	4	3
11	90	5	4
12	74	4	3
13	103	5	5
14	73	4	3
15	93	5	4
16	38	2	1
17	60	3	3
18	82	4	4
19	78	4	3

Table 1: Detailed list of overall quality ratings (Rating) given by the 19 test subjects (TS). MOS_{centred} gives ratings on the MOS scale under the assumption that labels on the continuous scale mark centres of subjective rating intervals, analogous is MOS_{lowered} the corresponding MOS rating, assuming labels on the scale mark lower margins of the subjective rating interval.

The second question asked for the most annoying properties of the connection just being used. Here the test subjects were to mark some well known annoying properties during a phone call or to add an extra annoying factor. The results are presented in figure 3. Owing to the loud environment, in which the test took place, noise and a too silent voice of the conversation partner were regarded as most annoying. Most people who regarded noise as most annoying referred to the environmental noise,

which cannot be influenced at a certain place and time. Only one test subject was aware of noise in the connection.

6 Conclusion

In this paper, a test design for subjective audio quality was presented, the effort for executing the test was calculated. Using SCTs statistically reliable data can be acquired within one week.
Although not carried out under specified conditions and therefore not representative, a first presentation of the developed test environment detected a known annoyance factor, after 19 test subjects made the test. For a terminal in current production, detected annoyances will be eliminated to increase user acceptance. Some effort is required on the selection of questions and predefined answers and rating scores, respectively. To avoid misunderstandings and double meanings is of great importance in acquiring data from naive and expert test subjects. Therefore questions and rating scores have to be simple and clear. Some misunderstandings can be avoided by a detailed explanation of terms and test procedure in advance.
To avoid wrong parameter adjustment, sometimes a comparative listening test is carried out before there is a second turn of naive user tests to prove performance increase.

Acknowledgments

The work presented in this paper is a co-work of Siemens AG, ICM MP PD TI and ICM MP PD HW. The authors gratefully thank S. Möller and A. Raake from Ruhr-Universität Bochum for their helpful cooperation.

References

[1] ITU-T Rec.E.800, *Terms and definitions related to quality of service and network performance including dependability*, International Telecommunication Union (1994)

[2] U. Jekosch, *Sprache hören und beurteilen. Qualitätsbeurteilung von Sprechtechnologien als Forschungs- und Dienstleistungsaufnahme*, habilitation thesis, Universität/Gesamthochschule Essen (1998)

[3] ITU-T Rec.G.107, *The E-model, a computational model for use in transmission planning*, International Telecommunication Union (2003)

[4] ITU-T Rec.P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs*, International Telecommunication Union (2001)

[5] S. Moeller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Press (2000)

[6] S. Wiegelmann, *Einsatzmöglichkeiten und -grenzen der MNRU als Referenzverzerrung bei der Bestimmung von Sprachbertragungsqualitt im Telefoniebereich*, diploma thesis, Ruhr-Universität (1997)

[7] CCITT, *Handbook On Telephonometry*, (1992-2000)

[8] ITU-T Rec.P.832, *Subjective performance evaluation of hands-free terminals*, International Telecommunication Union (2000)

[9] T. Drascher, internal report, Siemens AG (2003)

[10] ITU-T Rec.P.800, *Methods for subjective determination of transmission quality*, International Telecommunication (1996)

[11] R. Aubauer, D. Leckschat, *Optimized second order gradient microphone for hands-free speech recording in cars*, Speech Communication **34** (2001)