A new tool for evaluation of sound quality in carkit telephony

Thorsten Drascher[1], Virginie Gilg[2], Martin Schönle[3]

[1] *BenQ mobile GmbH & Co OHG, 46395 Bocholt, Germany, Email: thorsten.drascher@benq.com*

[2] *Siemens AG, 81379 Munich, Germany, Email: virginie.gilg@siemens.com*

[3] *Siemens AG, 81739 Munich, Germany, Email: martin.schoenle@siemens.com*

Introduction

The evaluation of sound quality in mobile telephony applications has gained increasing interest during the last years. Target of existing tools like PESQ [1] is to quantify audio quality by MOS values. Thereby only global information can be obtained, important individual subjective impressions are neglected. For the assessment of some typical audio scenarios like double-talk or residual echo these tools are less suited. We present a new flexible tool allowing the comprehensive assessment of sound quality in a given configuration, e.g. in carkit telephony. The goal is to obtain an end-user assessment with reproducible and statistically reliable results in a relatively short time. The tool provides an ITU-conform test procedure for subjective tests delivering MOS values along with their related statistics. In addition it gives information on most annoying properties that allow a precise diagnostic of sound quality factors. A major advantage of the tool is the wide coverage of audio characteristics, some of them strongly contributing to the overall quality in the carkit use case. Example results obtained during in-situ tests of a carkit as well as first results of doubletalk tests are provided.

Evaluation Tool

Our tool is based on the subjective evaluation of audio quality by human test subjects. It can be tailored to the assessment of different audio systems and use cases. This is achieved by setting up individual tests from a pool of more than 100 questions extracted from various ITU-T recommendations. The focus of this contribution is on the evaluation of conversation quality, which will be presented in more detail in the following sections.

To obtain a simple and convenient user interface the tool is programmed in Tcl/Tk. Several basic GUI components are used, as for example radiobuttons or a slider element, which can be shifted continuously to obtain a higher resolution of the MOS [2]. Its range is linearly mapped to a numeric scale between 0 and 100, with "bad" at 10 and "excellent" at 90. The headroom left for assessments worse than 'bad' or better than 'excellent' encourages the test subjects to use these extreme assessments if needed, thereby increasing the evaluation range.

Checkbox elements can be used for the assessment of different audio quality aspects. In combination with the MOS values provided by the sliders a very detailed analysis of possible degradations can be achieved.

Test Setup

The experimental setup is shown schematically in Figure 1. A fixed network telephone, which is used as reference terminal, is served by one test subject in a silent office environment, while the device to be tested is served by the second test subject in a different location. This can either be an acoustic laboratory, where various background noises with different levels can be simulated, or a real environment, a car for example. Both locations have to be equipped with a PC/laptop.

Test Design

The tests are carried out as conversational tests followed by an interrogation of the test subjects. Designing a new test just requires editing a test design file, where conversation tasks and background noises are defined, and a question file, where the questions for the interrogation are specified. ITU-comform conversation tasks [3, 4] as well as assessment questions can be adapted to the actual test. Graeco-Latin squares are used by the tool for the construction of a unique order of background noises and conversation tasks for each test subject. Thereby adaptation effects to different noisy environments can be avoided. After this preparatory work, test subjects and experimenters are guided through the test automatically.

Statistical Reliability And Time Effort

By taking an adequate number of test subjects test results with a high statistical significance can be obtained. The amount of reliability can be expressed by confidence intervals. Their size depends on the standard deviation and on the number of test subjects. According to our experience a number of 20-25 test subjects is sufficient, if expert users are participating, this number can be reduced to 10-12. The time effort for the tests presented in the next sections was two days for the doubletalk tests and 5 days for the carkit tests.

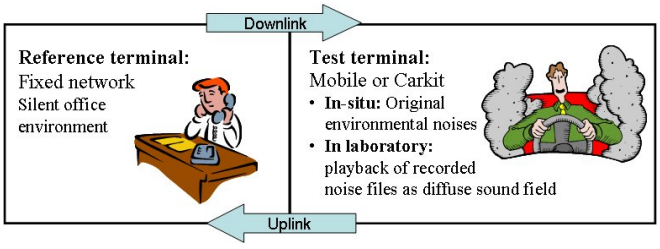

Figure 1: Schematic Test Setup

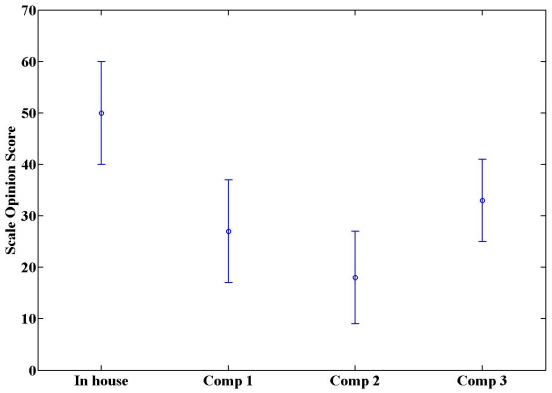

Figure 2: Capability to talk back and forth at the reference terminal: MOS values for each device under test and the corresponding 90% confidence intervals.

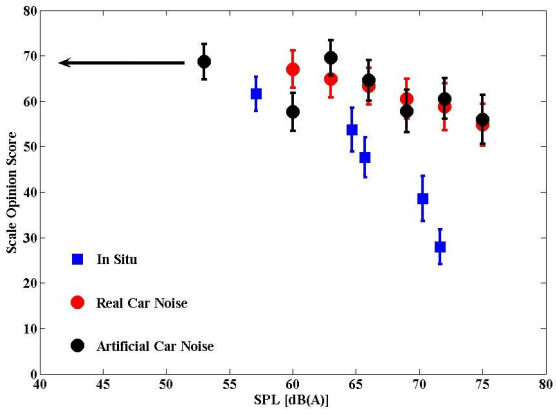

Figure 3: Overall quality at the reference terminal in laboratory and in-situ carkit tests: MOS values for each situation and the corresponding 1σ intervals. If the laboratory values (circles) are shifted to the left, all measurements form a smooth curve.

Doubletalk Benchmark

Doubletalk benchmark tests between four mobile devices were carried out as blind (referring to the test subject at the reference terminal) laboratory tests by twelve audio experts. Several questions and simultaneous reading tasks dedicated to double-talk quality were assembled in the test set-up. Figure 2 shows an example tool output and reveals the statistically significant better performance of the in-house device regarding the capability to talk back and forth at the reference terminal.

Carkit Assessment

In contrast to the double talk benchmark, carkit tests were carried out as in-situ tests with unexperienced users. The absolute noise levels were measured near the carkit cradle. The same tests were repeated in laboratory using a car mock-up. An ITU artificial car noise [5] as well as a real car noise were played back as diffuse sound field. The noise levels were adjusted in a way that in-situ and laboratory results can be compared. Figure 3 shows the results for the overall quality as assessed at the reference terminal for the in-situ and the laboratory situation. For both situations, the performance decreases with increasing noise level but the drop-off happens more rapidly in the in-situ test. It is interesting to note that in the laboratory tests, real and artificial car noises yield very close results. If the laboratory curves are shifted to the left, all measured points form a single curve, being this a systematic effect is still hypothetical. This deviation may be caused by the difficulty for simulation of some effects in the laboratory, as for example hand-over between two base stations. The additional impacts of speed and an environment moving rapidly relative to test subjects seem to decrease the subjective assessment in comparison to laboratory tests at constant noise level. Figure 4 gives an overview of the most annoying properties, which were announced by the participants during the tests. It shows the large amount of audio effects the tool can cover.

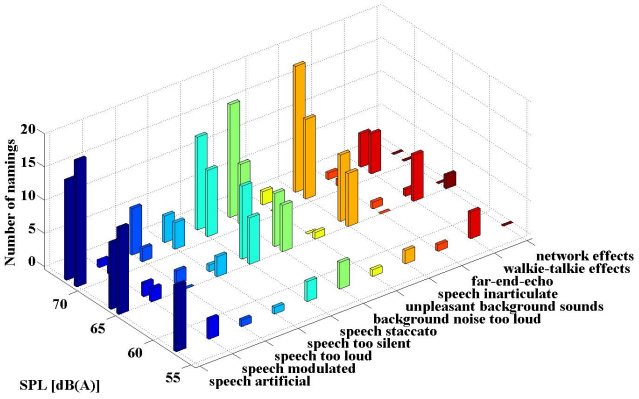

Figure 4: Example tool output for most annoying properties during carkit in-situ call at the reference terminal.

Conclusion

The flexible tool presented here can be used to assemble and execute statistically reliable ITU-conform subjective tests of arbitrary audio systems in short time. In difference to automatic evaluation tools a comprehensive assessment of audio systems is achieved.

References

[1] ITU-T Rec.P.862 (2001)

[2] ITU-T Rec.P.851 (2003)

[3] Assessment and Prediction of Speech Quality in Telecommunications, Kluwer Academic Publishers, Boston, 2000

[4] ITU-T Rec.P.832 (2000)

[5] ITU-T Rec.P.800 (1996)