Speech Quality Testing – Part I: Auditory Assessment of Communication Scenarios

Magnus Schäfer[1], Jan Reimes[1], Alexander Kling[2]
[1] HEAD acoustics GmbH, 52134 Herzogenrath, Germany
[2] Technische Universität Köln, Germany

Abstract
Assessing the user experience is paramount for testing and developing modern communication systems which are expected to provide high speech quality in all use cases and in all acoustic environments. A wide range of challenging acoustical conditions is encountered particularly by hands-free devices. Noticeable reverberation and background noise are among the most common degradations for this use case.

This contribution provides an auditory assessment of the influence of different degradations on the speech quality. Acoustical measurements of realistic telecommunication scenarios were carried out to compile a comprehensive and meaningful set of test conditions for two listening tests. These conditions comprise smartphones in different positions as well as vehicle and desktop hands-free devices. Two speech quality tests according to the guidelines in ITU-T P.800 and ITU T P.863 were conducted.

An analysis of the results of the listening tests shows that the design can assess such a wide variety of test conditions and that differences between these conditions lead to plausible differences between their respective scores. The resulting datasets provides a valuable resource to investigate the capabilities of instrumental assessment methodologies in a companion contribution.

Introduction
Testing of speech quality is important to ensure that communication devices deliver the performance that the customers expect. The auditory assessment of speech quality usually serves one of two purposes: Either the results of the listening test are directly used to quantify the performance of a set of devices or they are used to develop or validate an instrumental assessment method. One prominent example of such an instrumental assessment for speech quality is [1].

As both the communication devices as well as the environments in which they are used are continuously changing, it is of high interest to ensure that existing instrumental methods are stil usable for contemporary communication scenarios. This contribution describes two listening tests that were conducted to evaluate the capabilities of different instrumental assessment methods. This evaluation can be found in the companion contribution [2].

Audio recordings
To conduct a listening test, suitable audio material is needed. In a communication scenario, this is almost always going to be speech. The German speech recordings from Annex E to ETSI TS 103 281 [3] are used in this investigation. They are sequences of sentences spoken by eight speakers in total (four female, four male speakers). Each speaker utters a pair of sentences.

Two different types of recordings were made with the devices that are used. A communication scenario consists of (at least) two devices with one being on the side of the talker (denoted sending side or SND in the following) and the other being on the side of the listener (denoted receiving side or RCV in the following). The interface between the two sides is located in the communication network and called the point of interconnect (POI).

This implies that there are three different possibilities to make recordings in this situation:

- **Sending direction**
  From the talker to the POI, this case is visualized in Figure 1. The speech signal is emitted by an artificial mouth, picked up by the communication terminal, transmitted through the network and electrically recorded at the POI.

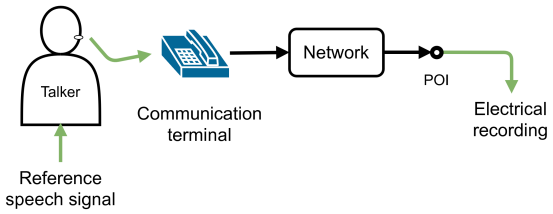

Figure 1: Setup for the recordings in sending direction

- **Receiving direction**
  From the POI to the listener, this case is visualized in Figure 2. The speech signal is fed into the network, emitted by the communication terminal and then binaurally recorded at the listener position.

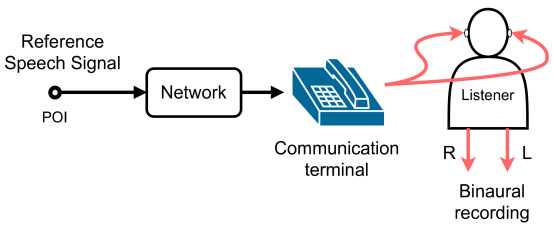

Figure 2: Setup for the recordings in receiving direction

- **End-to-end recording**
  From the talker to the listener, this case is the combination of the two aforementioned cases connected at the POI. This case obviously features two communication terminals, then. For most applications,

it is only relevant to quantify the performance of one terminal. Accordingly, no recordings were made for the end-to-end case in this investigation.

Three types of communication terminals are present in this investigation: modern smartphones, vehicle hands-free devices and desktop conference phones. All terminals encounter different degradations in their respective use cases. Thus, the measurements differed as well. The smartphones, e.g., were measured both in the handset [4] and the handheld hands-free [5] position. The measurements in the cars were done without background noise and with the fan being active.

Listening Test

General Setup

The listening test follows the absolute category rating paradigm with the opinion scale on speech quality from [6]. The scale is given in Table 1 with the corresponding numerical scores.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 1: Opinion scale on speech quality from [6]

Two different listening tests were prepared – one using the material for the sending direction, one using the material for the receiving direction.

Specific Requirements for Anchor and Test Conditions

In accordance with Appendix II of [1], there are certain requirements for the anchor and test conditions. The twelve anchor conditions are actually well-defined in that appendix. Besides the unmodified reference condition, they consist of different noise scenarios, signal attenuations, bandpass filterings and packet losses.

There are also constraints on the test conditions due to certain design rules that are specified. At least 30 % of the conditions have to be noisy. The target for the acoustic bandwidths is given in Table 2.

Bandwidth	Share
Super-wideband	>10 %
Wideband	>30 %
Narrowband	>30 %

Table 2: Design rules for the acoustic bandwidths

The final set of design rules considers the presentation level relative to nominal level. They are given in Table 3.

The constraints for noisiness and acoustic bandwidths could be easily met with the available material by choosing suitable recordings. The presentation levels required

Range of presentation levels	Share
−20 dB to −12 dB	>10 %
−12 dB to −3 dB	>10 %
−3 dB to +3 dB	>60 %
+3 dB to +6 dB	>10 %

Table 3: Design rules for the presentation levels

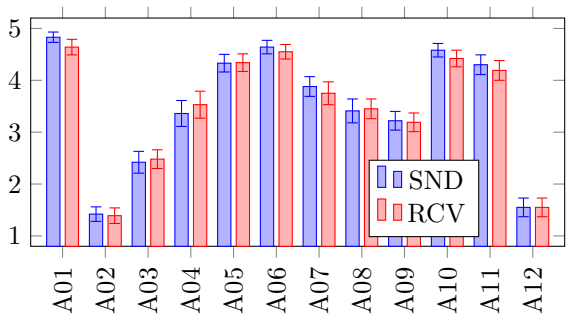

Figure 3: Overview of anchor conditions of both listening tests

some adjustments to fulfill the requirements from Table 3. When adjusting levels, care was taken to change the levels for all recordings of one device (or one operating mode) by the same amount. Thereby, the impact of the different degradations is still quantifiable from the auditory results.

For the listening test in the sending direction, the signals of the smartphones could be left unaltered, the signals of the cars had to be amplified and the signals for the conference phones needed slight adjustments (small amplifications for one device, small attenuation for the other device). For the listening test in the receiving direction, the signals for all smartphones and all cars had to be attenuated, the signals for the conference phones had to be attenuated.

Conducting the test

With these preparations of the signals, eight different versions of both listening tests were prepared (corresponding to the eight speakers in the sequences that were used) with each anchor and test condition present in all tests. The individual sentence pairs were assigned randomly to the eight tests. The signals were presented to the test subjects using equalized playback equipment. The electrical recordings in the test for the sending direction were presented diotically while the recordings for the receiving direction were played back binaurally.

In total, 48 test subjects participated in the tests. As some of the subjects rated more than one version of the test, 64 votes were collected per condition. The results are the Mean Opinion Scores (MOS) for all conditions, the corresponding confidence intervals are reported as well to illustrate the spread of votes.

Auditory Results

The listening tests provide a large amount of results and facilitate both interesting analyses of subgroups of test

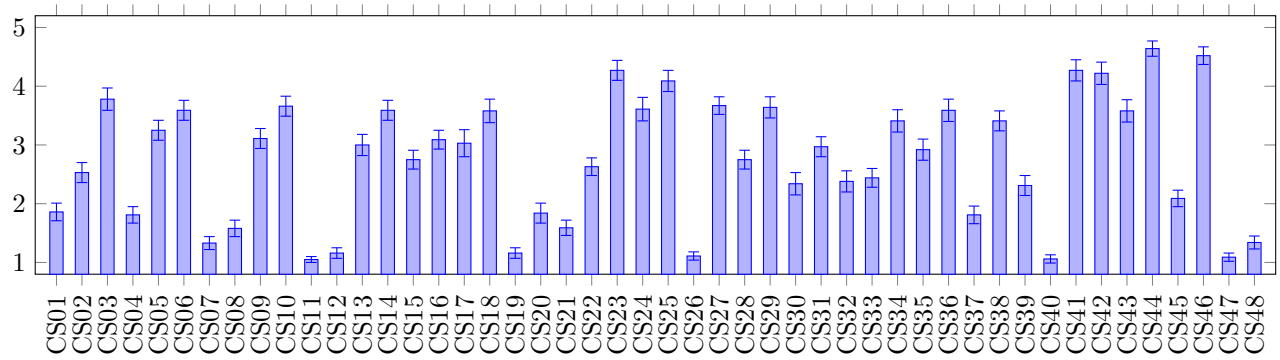

Figure 4: Overview of all test conditions of the listening test for the sending direction

conditions as well as a comparison of auditory results with the output of instrumental assessment methods (see [2] for more details on this). In this section, the results for the anchor conditions and the test conditions are presented. For both tests, one subgroup of the results is analysed in more detail to show the impact of different signal degradations on the perceived speech quality.

The anchor conditions are completely identical between the two tests. Accordingly, one should expect that the results are at least similar. They are depicted in Figure 3 with the results for the sending side in blue and the results for the receiving direction in red. The whiskers around the ends of the bars show the confidence intervals for the conditions.

As expected, the results are very similar. The best result is A01, the clean reference without any degradations. This condition has a score of 4.83 MOS and 4.64 MOS in the test for the sending and receiving direction, respectively. The lowest score (1.42 MOS and 1.39 MOS, respectively) can be observed for A02, one of the four noisy cases among the anchor conditions.

An overview of the results for the sending direction can be seen in Figure 4. Test conditions CS01 to CS22 are cars, CS23 to CS32 are conference phones and CS33 to CS48 are smartphones.

The highest score among the test conditions for the sending side is CS44 with 4.64 MOS. This is smartphone in handset position with a super-wideband transmission and without any additional degradations. The lowest score of 1.05 MOS can be observed for CS11. In this condition, the talker is positioned on the passenger seat of a car, a narrowband transmission is used and there is fan noise.

It can be observed that a large portion of the available rating scale is used in this listening test. Even very low average scores can be observed for some conditions. It is quite difficult to demonstrate the impact of individual degradations in this overview. An excerpt of the results of the listening test for the sending direction is shown in Figure 5.

This excerpt consists of the results for all smartphone conditions in the test, sorted from best (CS44, score

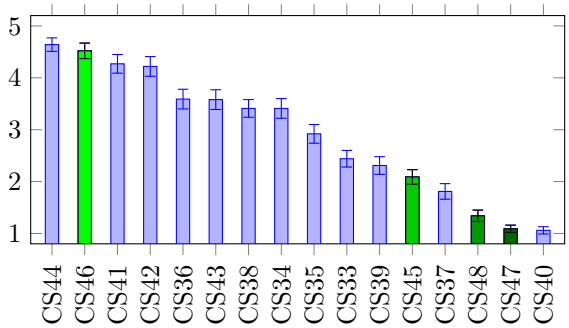

Figure 5: Overview of all smartphone conditions of the listening test for the sending direction

4.64 MOS) to worst (CS40, score 1.06 MOS). These conditions feature two different phones, two different acoustic bandwidths, two different positions of the phone. In addition, background noise [7] and/or reverberation [8, 9] is present in some conditions. The impact of these two degradations can be quantified from the auditory results.

There are four conditions marked in the figure in increasingly darkening shades of green: CS46, CS45, CS48 and CS 47. These four conditions all are the same smartphone in the same position (handheld hands-free) with the same acoustic bandwidth (super-wideband). The difference between the conditions is the presence or absence of background noise and reverberation.

The highest score of 4.52 MOS can be seen for the case without any degradations. Background noise lowers the score to 2.09 MOS, the impact of reverberation is even higher with a score of 1.34 MOS. And the combination of the two degradations almost reaches the bottom end of the rating scale with a score of 1.09 MOS.

Figure 6 shows a complete overview of the results for the listening test for the receiving direction. Here, conditions CR01 to CR16 are smartphones, CR17 to CR42 are cars and CR43 to CR48 are conference phones.

Compared to Figure 4, it is obvious that the spread between the worst and the best scores is not as large in this test. The best case here is CR35 with a score of 4.23 MOS. This condition has the listener on the passenger seat of one of the cars with a wideband transmission and without any degradations. The worst case is CR36

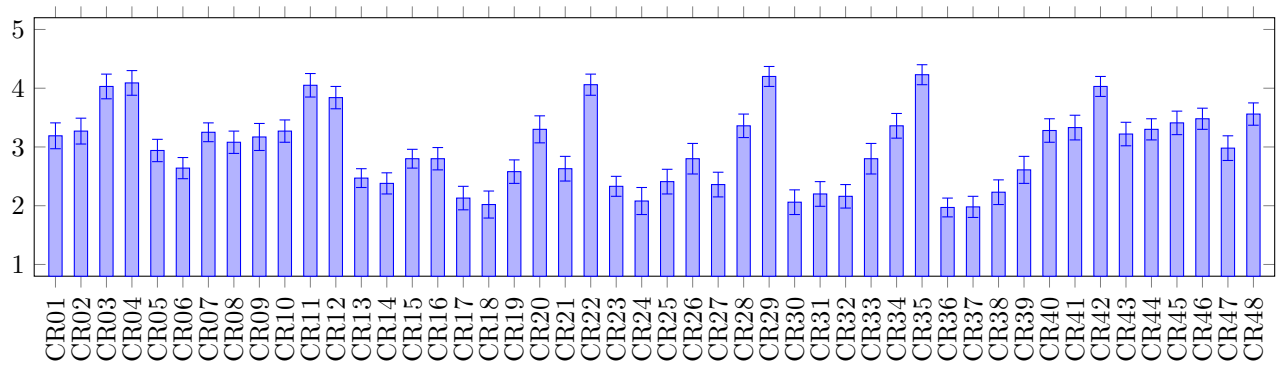

Figure 6: Overview of all test conditions of the listening test for the receiving direction

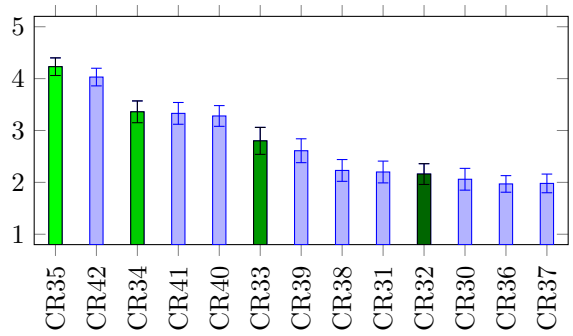

Figure 7: Overview of all conditions for one car of the listening test for the receiving direction

(score of 1.97 MOS). The listener here is on the driver seat in the same car with a narrowband transmission, fan noise and with the hands-free device set to maximum loudness. As an excerpt from this test, only the results for one of the cars are shown in Figure 7.

The four marked conditions in this diagram (CR35, CR34, CR33 and CR32) all have the listener on the passenger seat of the same car and use the same acoustic bandwidth (wideband). The differences here are the presence or absence of fan noise and the loudness setting of the hands-free device (nominal or maximum loudness).

The condition without fan noise and with the hands-free device on nominal loudness achieves the highest score of 4.32 MOS. Switching the hands-free device to maximum loudness introduces noticeable distortions that have a detrimental effect on the perceived speech quality leading to a score of 3.36 MOS. The fan noise lowers the score to 2.80 MOS and the combination of both degradations is even worse at 2.16 MOS.

Conclusion

An overview of two comprehensive listening tests is given in this contribution. The speech quality tests focus on modern communication devices: smartphones, conference phones and vehicle hands-free frontends. Recordings of realistic usage scenarios with realistic degradations were made for all devices. From these recordings, two listening tests according to the design rules from [1] were prepared and conducted.

A brief analysis of the results shows that the anchor conditions, which are identical in both tests, also get very similar scores in the two tests. The results for the test conditions exhibit a plausible impact of the different degradations as shown by analyses of subsets of the results.

The listening tests are very well suited for a comparison of the auditory results and the corresponding output of instrumental methods, this comparison is done in the companion contribution [2].

References

[1] ITU-T Recommendation P.863. *Perceptual objective listening quality prediction*, Mar. 2018.

[2] Jan Reimes, Magnus Schäfer, and Alexander Kling. Speech Quality Testing – Part II: Performance of Instrumental Assessment Methodologies. In *Fortschritte der Akustik - DAGA 2021*. DEGA e.V., Berlin, 2021.

[3] ETSI TS 103 281 V1.2.1. *Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*, January 2018.

[4] ITU-T Recommendation P.64. *Determination of sensitivity/frequency characteristics of local telephone systems*, June 2019.

[5] 3GPP TS 26.131. *Speech and video telephony terminal acoustic test specification*, Sep. 2018.

[6] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*, August 1996.

[7] ETSI TS 103 224 V1.3.1. *A sound field reproduction method for terminal testing including a background noise database*, July 2017.

[8] ETSI TS 103 557 V1.1.1. *Methods for reproducing reverberation for communication device measurements*, December 2018.

[9] Hans W. Gierlich, Magnus Schäfer, Torsten Ungerechts, Benedikt Koppers, Jan Reimes, Shahd Al Hares, and Christoph Nelke. Reproduction of Reverberation. In *Fortschritte der Akustik - DAGA 2019*. DEGA e.V., Berlin, 2019.