Speech quality of wide- and narrowband speech codecs: Object- and subject-oriented view

Alexander Raake[1], Sascha Spors[1], Hans-Joachim Maempel[2], Timon Marszalek[2], Simon Ciba[2], Nicolas Côté[1]

[1] *Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Email: alexander.raake@telekom.de*

[2] *Fachgebiet Audiokommunikation, TU Berlin*

Introduction

We describe an extensive auditory test series on the perceived quality of narrowband (NB, 300-3400 Hz) and wideband (WB, 50-7000 Hz) speech coders under different network conditions. Listening quality tests may be classified according to two different dichotomies [1]: (i) *analytical—utilitarian* refers to whether the perceptual features of transmitted speech are assessed (*analytical*), or its integral quality (*utilitarian*). (ii) *subject-oriented—object-oriented* refers to whether the perception process or the role of the human test subjects is under test (*subject-oriented*), or the speech transmission system (*object-oriented*). In the test discussed in this paper, integral speech quality was assessed (*utilitarian* test), taking both a *subject-* and an *object*-oriented perspective.

Accounting for the object-oriented view, the quality-impact of different conditions of WB and NB speech codecs were assessed, namely: (i) in single & tandem operation, (ii) under IP packet loss, and (iii) in the presence of background noise at send side. The subject-oriented view was addressed by recruiting a large number of test subjects from six different user groups (120 subjects; appr. 50% female, 50% male; age 17 to 80 years, close-to normal distribution). The grouping is based on a telephone pre-screening of the subjects using a questionnaire and targets a classification with regard to their assumed market behavior (Deutsche Telekom's so-called BBFN$^+$–segments: (1) IP-Experts; (2) Entertainment generation; (3) Demanding establishment; (4) Critical followers; (5) Cost-oriented laggers; (6) Conservative telephone users).

Test procedure

Overall, 114 test conditions plus 11 reference conditions were assessed, using source recordings from 4 speakers (2f/2m). The conditions included WB-codecs such as pure PCM, the AMR-WB (ITU-T Rec. G.722.2), the G.722, and the G.729.1, and NB-codecs such as the G.711, the G.729A, and the G.726. In addition to the single operation mode, both WB↔WB and NB↔WB codec tandems were tested. Most codecs in single operation were also tested with additional background noise at send side. Here, two types of noise were used (cafeteria & car, at two levels). Packet loss (uniform) was inserted for the majority of the WB-codecs, with loss-rates from the set $0, 1, 2, 4, 8\%$. After each test sample, quality ratings were collected using the 5-point ACR-scale (the "MOS-scale", see ITU-T Rec. P.800, 1996). After the actual rating phase, a questionnaire was presented to the subjects to collect some additional subject-oriented information.

Results: Object-oriented view

In order to make the object-oriented analyses usable for a WB-extension of the so-called E-model (ITU-T Rec. G.107, 2005), the MOS-data were transformed onto the E-model's new WB *R*-scale [0, 129] (see [2]), following a similar procedure as in [4].

The test is broader than typical codec tests – in terms of range of conditions and range of test subjects, yielding results that show some interesting deviations from the respective literature. In the following, we summarize some few examples of the test results:

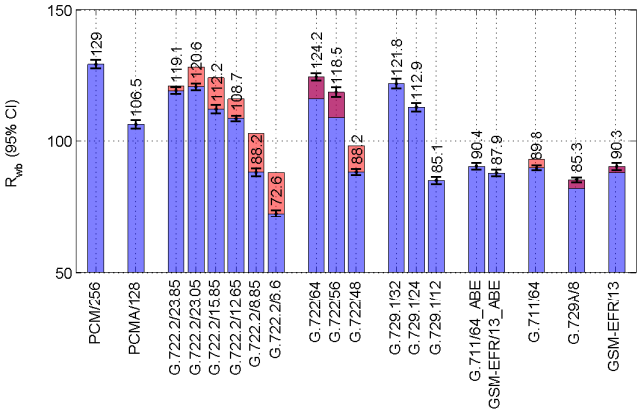

Figure 1: Transformed test results for single coding conditions. The error-bars show the mean *R*-values and their 95% confidence intervals. The underlying bars highlight the differences between the expected ratings according to [3, 2] and our test results. See text for details.

Fig. 1 shows a comparison of the test results for single-coding conditions. The deviations from the *R*-values expected based on [3] are also indicated, using lighter red stacked bars to indicate conditions yielding lower quality ratings than expected, and darker red bars those with higher quality than expected. The test confirms the quality advantage of WB over NB of more than 35 points on the 129-point scale. An unexpected result is the reversed quality rank-ordering of the G.722 at 64 kbit/s and the G.722.2 at 23.05 kbit/s (compared to [4, 3]; see bars 4 & 9). For the G.722.2, the quality decrease with decreasing bitrate is stronger than expected (bars 9 & 10). In contrast, the G.722 is rated better than expected, at least at the two higher bitrates.

Informal listening to the processed samples and a comparison with other source material processed using the same channel conditions reveal an influence of the room-acoustics at the recording site: An increasing amount of room reflections audible in the signal seems to lower the quality-impact of nonlinear coding distortions. This effect may account for the unexpected rank-ordering of the G.722 and G.722.2 at their best bitrates. The effect needs to be investigated in more detail in the future, also answering the question of how much of the real-life room-acoustics will be captured by handset- or headset-typical close-talking microphones

As a complement to the WB and NB conditions used in the test, we have included two otherwise clean conditions with an artificial bandwidth extension, both applied to G.711- and GSM-EFR-encoded NB speech. The test results reveal no quality advantage over the NB-case (4^{th} and 5^{th} bars from the right in Fig. 1). Moreover, the three rightmost bars in Fig. 1 reveal lower quality-differences between the NB-codecs than expected, showing that subjects did not well resolve NB coding distortions.

In case of codec tandeming of two codecs, the following observations were made:

1. The best codec tandem still reduces quality by more than 10 points on the 129-point scale, as opposed to only 2 points expected according to [2, 3].
2. For asymmetric WB↔WB tandems, quality does not depend on the order in which the two codecs are applied, in contrast to the results in [4].
3. In case of NB↔WB tandems, there is a strong dependency on the codec order.

Observation 3.) can be explained as follows: For the order NB→WB, a filter was applied that is typical of the sending characteristics of NB-handsets (the so-called Intermediate Reference System, IRS, a pre-emphasis of approx. 3 dB/octave); however in the case WB→NB, such a filter was not used in our test. As a consequence, the direction NB→WB is systematically rated higher than the other direction. Especially the latter finding is of practical relevance: For connections where a WB→NB transcoding is to be used, such an IRSsend-type pre-filter should also be applied in the user interface.

Results: Subject-oriented view

A mixed model analysis considering repeated measurement, using the fixed factors "test session", "condition", "subject age", "user segment", and the "overall hearing loss" (here defined as the sum of the hearing loss (in dB) over the last three tested bands for each ear) as fixed factors revealed significant effects for all factors. The most important factor was found to be the "condition" ($F = 146.2$, $p < 0.005\%$), followed by "age" ($F = 39.7$, $p < 0.005\%$) and "user segment" ($F = 18.1$, $p < 0.005\%$).

An example of the subject-dependency is shown in Fig. 2, left graph. Here, the average, untransformed test results for the clean, log. PCM NB channel and for the lin. PCM WB channel are depicted, using the segment-membership of the subject as a grouping parameter. The plot reveals the similar perception of the WB-advantage over NB by segments 1–5, and a deviating rating behavior by segment 6, the conservative telephone users. Since this user segment comprises the oldest among the participating subjects, the smaller perceived advantage may be due to the reduced hearing capability of some of the members.

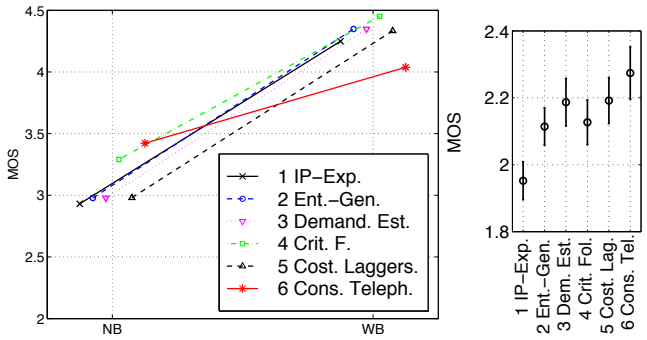

Figure 2: Non-transformed test results. Left: Clean NB (log. PCM) compared to clean WB (lin.PCM), grouped according to listener segments. Right: Ratings averaged over the packet loss rates $\geq 4\%$, plotted over the 6 segments.

Another example of a subject-dependent effect is shown in the right part of Fig. 2, depicting the untransformed ratings averaged over all conditions with a packet loss percentage $Ppl \geq 4\%$ plotted as a function of the user segment. Here, the IP-experts (1) show significantly lower ratings than the other segments, possibly owing to their prior experience with this kind of degradations. It has to be noted that the subject-dependent differences generally are much smaller than the differences caused by the different conditions.

Acknowledgment

The authors wish to thank Marcel Wältermann, Sebastian Möller and Robert Schleicher from Deutsche Telekom Laboratories for fruitful discussions and Lars-Alexander Mayer and Florian Köhler from trommsdorff + drüner, Berlin, for recruiting the test-subjects.

References

[1] Raake, A.: Speech Quality of VoIP – Assessment and Prediction. John Wiley & Sons Ltd, UK–Chichester (2006).

[2] ITU-T Rec. G.107 Appendix II: Provisional Impairment Factor Framework for Wideband Speech Transmission. ITU-T (2006), CH–Geneva.

[3] ITU-T Rec. G.113 Appendix IV: Provisional Planning Values for the Wideband Equipment Impairment Factor Ie,wb. ITU-T (2006), CH–Geneva.

[4] Möller, S., Raake, A., Kitawaki, N., Takahashi, A., Wältermann, M.: Impairment Factor Framework for Wideband Speech Codecs. IEEE Trans. Audio Speech and Language 14 (2006), 1969-1976.