3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

# How to talk about speech and audio quality with speech and audio people?

*Alexander Raake*[1], *Marcel Wältermann*[1], *Falk Schiffner*[1], *Ulf Wüstenhagen*[2], *Bernhard Feiten*[2]

[1]Deutsche Telekom Laboratories, TU Berlin, Berlin, Germany
[2]Deutsche Telekom Laboratories, Deutsche Telekom AG, Berlin, Germany
alexander.raake@telekom.de

## Abstract

We present results of two speech and two audio quality listening tests to compare the two test methods ACR and MUSHRA for different content types. These comparisons have two primary goals: (i) Find relations for transforming test results obtained with one method onto the scale of the other, and (ii) refine audio quality results obtained using the ACR method by using MUSHRA-results for the upper quality regime, where MUSHRA typically shows a sightly better resolution than ACR. The aim is to contribute to the harmonization of speech and audio assessment methods considered meaningful in the light of the convergence between speech and audio coding and transmission.

**Index Terms**: speech, audio, quality, test methods

## 1. Introduction

Speech and audio services as well as their technical components like the respective codecs are more and more converging: Today's users employ the same mobile devices for music listening and telephone conversations, the initially narrowband telephone speech (300-3400 Hz) is getting audio-typical bandwidths - making it wideband (50-7000 Hz), super-wideband (50-14000 Hz) or even fullband (20-22000 Hz) -, audio codecs are becoming low-delay, speech codecs need to well transmit music, and they more and more exploit not only speech production but also audio perception properties. Last but not least, with IP-based transmission, speech and audio services share the same transport mechanisms. In spite of this bi-directional convergence, auditory as well as instrumental quality assessment techniques are only slowly converging, also since the speech and audio communities have traditionally been two separated ones.

The motivation for conducting speech or audio quality tests are varied, for example: Selection of a winning candidate during a codec standardization process; benchmarking of codecs, components or entire service technologies during the design or planning of the service; development of instrumental quality prediction methods and models. For any benchmarking task with an outcome that remains valid and usable across the borders of a given experiment, a common way of expressing the results is required. For speech quality, first attempts have been made to express the quality of narrowband (NB), wideband (WB), super-wideband (SWB) speech on a single quality scale, the $R$-scale of the so-called E-model (E-model: [1]; extensions: [2, 3, 4]). Based on these previous analyses, different speech codecs can directly be compared on this scale. So far, a test-independent comparison of different audio codecs or of audio with speech codecs cannot readily be made.

As a practical way forward, this paper provides an example of how ratings obtained with the two most important listening test methods from the audio and speech quality assessment domains can be compared and possibly converted into one another. The method frequently used for intermediate audio quality assessment is MUSHRA (MUlti Stimulus with Hidden Reference and Anchors, [5]). For speech quality assessment, up to wideband speech, the Absolute Category Rating is typically used (ACR, [6]).

The paper analyzes and compares the results of four listening tests, where we have assessed on both scales (a) speech quality in a VoIP context with narrowband up to fullband speech, using both speech and low-delay audio codecs, and (b) audio quality in an IPTV context. The two audio tests (b) have been described and analyzed earlier in [7]. In the previous paper, the focus was on using the ACR method for covering a wide range of quality-levels, and the MUSHRA test for achieving more detailed ratings in the higher quality range. A linear relation between the ratings on the two scales could be observed. The ACR-test involved stimuli with a relatively high dynamic range of quality levels, and the MUSHRA-test a subset of these stimuli explicitly covering a smaller range of quality levels. Results show that – for audio – the accuracy of the MUSHRA test in the higher quality range is higher. The present paper summarizes the main findings of tests (b), attempting to make the MUSHRA-results usable in the ACR-context (see Section 4.

Another question the paper addresses is how the two methods compare when the *same* set of stimuli is being used, and the set covers a wide range of quality levels like in the ACR-case of test (b). In two recent speech quality listening tests (a), we have assessed an identical set of mixed narrowband, wideband, super-wideband and fullband stimuli once using ACR with the classical 5-point scale, and once using a MUSHRA test (see Section 3).

The paper is laid out as follows: Section 2 gives a brief overview of existing auditory and instrumental speech and audio quality assessment methods; Section 3 discusses the two speech quality tests (a) conducted using the ACR and the MUSHRA methods; Section 4 discusses the possible usage of MUSHRA-results within the ACR-framework, based on the results for tests (b) reported earlier in [7]; finally, we provide conclusions and an outlook in Section 5.

## 2. Overview auditory test methods

It is common practice that different methods are being used when it comes to assessing audio or speech in auditory tests, and the choice is based on the range of qualities to be expected, but also related with the practice of what has been used in the respective field in the past. As pointed out also in [8], there are three methods that are most commonly used in the context of audio quality evaluation:

**ACR:** In ACR-tests, users are sequentially presented with


10.21437/PQS.2010-13

single test stimuli, and are asked to judge their integral quality, typically on the 5-point category rating scale (the so-called "MOS"-scale, mean opinion score [6], see Figure 1(a) for an example implementation of the scale). In case of speech, source sentences typically stem from 4-6 speakers of half female, half male sex. These are processed with the test conditions and typically presented in different randomized playlists to the subjects, who rate quality after each item. This test method has been the method of choice in the majority of speech quality tests for bandwidths up to wideband.

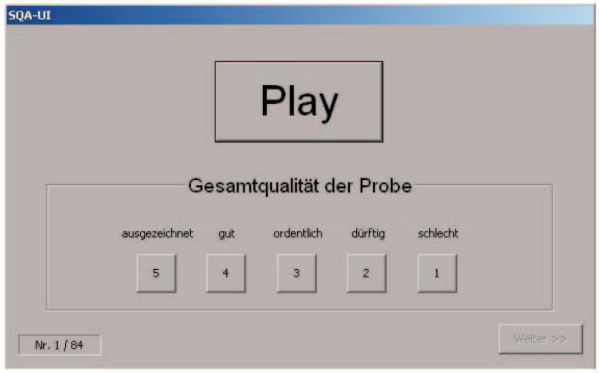
**MUSHRA:** Tests according to [5] are recommended for intermediate levels of audio quality degradations and differences. Here, users are presented with a GUI that allows listening to a set of $N$ stimuli, typically with $N \leq 12$. The original, undistorted signal is used as explicit reference. Three stimuli are used as anchors hidden among the test stimuli: An additional instance of the reference, a low-quality anchor, typically NB-filtered, and a medium-quality anchor, typically un-coded WB. The stimuli are looped, and the subjects can switch between the stimuli, giving their judgments on a continuous 100-point scale relative to the known reference. For assessing the entire set of test conditions, it is typically required to employ several test sets. The GUI for an example implementation of the MUSHRA-method is shown in Figure 1(b). The MUSHRA method is the one most commonly used for audio quality evaluation.

**BS.1116:** This method is targeted primarily at the evaluation of small audio impairments [9]. It employs a triple-stimulus with hidden reference paradigm: Three stimuli are presented in one set, where stimulus 'A' is the explicit reference, while stimuli 'B' and 'C' are the test stimulus and an additional hidden instance of the reference, with random assignment to 'B' and 'C'. The test subjects provide judgments of audio quality for 'B' and 'C' relative to the reference using a continuous quality scale. Interestingly, instead of the MUSHRA method often used in audio coding evaluation scenarios, the BS.1116 method and its derivative according to [10] have recently been selected as the methods of choice for the evaluation of FB speech codec standardization candidates [11]. In spite of this latest development, the work in this paper focuses on the comparison of ACR [6] and MUSHRA [5].

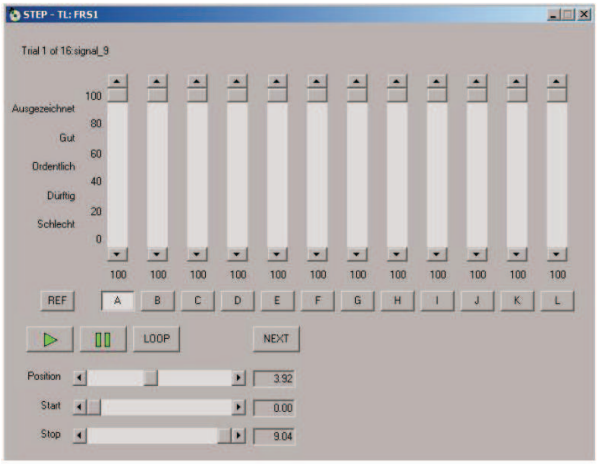
## 3. Speech – test (a)

Two listening tests were carried out, one using ACR, the other using the MUSHRA method. The same test conditions and speech material were used in both tests. The goal was to cover the entire range of qualities to be expected in today's telephone networks, from degraded NB-conditions up to clean FB. Apart from different NB, WB, SWB and FB codecs, we have used a number of conditions with bandpass filters, a few conditions with coding under packet loss, two conditions with codec tandeming, and two conditions with background noise at send side. The test conditions are summarized in Table 1. Note that one of the initially 39 conditions for each of the two tests has been omitted here and from the further comparison, since there was a small bug in the processing chain.

As source material, 30 German shortened sentences from the EUROM database were used [21], 15 recorded with two female and 15 with two male speakers. All sentences have been processed with all test conditions. For each of the ACR-playlists, a randomized selection of the available sentences from the four speakers was used, so that no fixed combinations of sentences–conditions were used in the test. For the MUSHRA-




(a) ACR.




(b) MUSHRA.

Figure 1: *Graphical User Interfaces used in the ACR (top) and MUSHRA tests (bottom).*

test, for each subject, each of the 16 sets of 12 stimuli was created with a different sentence–speaker combination, in total covering all speakers for each of the subjects.

The graphical user interfaces of the two software tools used for the ACR and MUSHRA tests are depicted in Figure 1.

The stimuli were presented diotically over Sennheiser HMD-410 headsets at 73 dB(A). The test room is an acoustically treated lab space at Deutsche Telekom Laboratories complying to [6]. The tests were conducted with two independent panels of naive listeners, who were paid for their test participation. For the ACR test, there were 24 subjects (12 female, 12 male), of an average age of 25.7 years. Twenty subjects took part in the MUSHRA test (11 female, 9 male), on average 25.8 years old. All subjects were recruited from the university campus of Technical University Berlin. All subjects were screened using pure tone audiometry, and were found to be normal hearing.

### 3.1. Test results

For Figure 2, the ACR and MUSHRA results were averaged over the test subjects. The graph shows the highly linear relationship between the mean ratings obtained on the two scales, with a linear correlation of 0.98. The highly linear relation between the results imply that any bias due to a perceptually non-

| # | Codec | Bitrate [kbit/s] |
|---|---|---|
| 1 | AAC-LD | 32 |
| 2 | AAC-LD | 48 |
| 3 | AAC-LD | 64 |
| 4 | AMR-WB+ | 13 |
| 5 | AMR-WB+ | 36 |
| 6 | AMR-WB+ | 6 |
| 7 | CELT | 32 |
| 8 | CELT | 40 |
| 9 | CELT | 48 |
| 10 | CELT | 64 |
| 11 | G.711 | 64 |
| 12 | G.722.2 | 12.65 |
| 13 | G.722.2 | 23.05 |
| 14 | G.722.2 | 6.6 |
| 15 | G.722 | 64 |
| 16 | G.729A | 8 |
| 17 | G.722.1 Ann. C | 24 |
| 18 | G.722.1 Ann. C | 32 |
| 19 | G.722.1 Ann. C | 48 |
| 20 | SILK | 12 |
| 21 | SILK | 32 |

| # | Bandpass (BP) | lower cut-off [Hz] | upper cut-off [Hz] |
|---|---|---|---|
| 22 | NB | 300 | 3400 |
| 23 | WB | 50 | 7000 |
| 24 | SWB | 50 | 14000 |
| 25 | FB (original) | 20 | 22000 |
| 26 | BP | 300 | 7000 |
| 27 | BP | 50 | 10000 |
| 28 | BP | 50 | 12000 |
| 29 | BP | 30 | 3400 |
| 30 | BP | 50 | 8000 |
| 31 | BP | 600 | 7000 |
| | Noise | type | Level [dB(A)] |
| 32 | WB | car | 55 |
| 33 | WB | car | 70 |
| | Packet loss | Bitrate | Ppl [%] |
| 34 | G.729A | 8 | 8 |
| 35 | G.722 | 64 | 2 |
| 36 | G.722 | 64 | 8 |
| | | Tandems | |
| | Codec 1 | Bitrate 1 | Codec 2 | Bitrate 2 |
| 37 | G.722 | 64 | G.722.2 | 12.65 |
| 38 | G.729A | 8 | G.729A | 8 |

Table 1: Test conditions used in both the ACR and MUSHRA tests. AAC-LD: Advanced Audio Coding – Low Delay [12]; AMR-WB+: Extended Adaptive Multirate – Wideband [13]; CELT: Constrained Energy Lapped Transform [14]; G.711: logarithmic PCM, a-law [15]; G.722.2: Adaptive Multi-Rate – Wideband [16]; G.722: see [17]; G.729: see [18]; G.722.1 Ann. C: see [19]; SILK: see [20].

linear scale [8] either equally applies for both scales (unlikely) or simply is not present. This effect is assumed to be due either to (i) the semantically equidistant labels in case of the German scale [8] or (ii) the fact that numeric labels and/or a graphical arrangement of the scale or buttons neutralize the effect of non-linearly distributed labels.

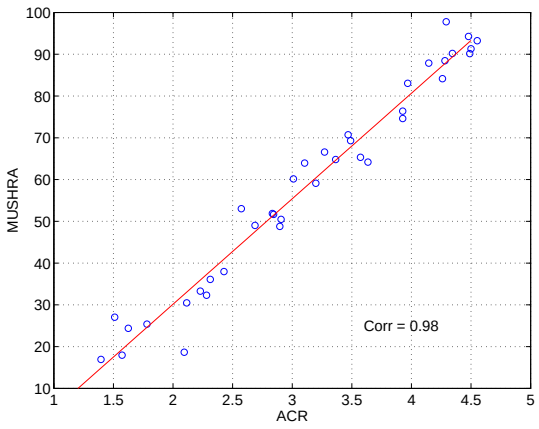



Figure 2: *Comparison of ACR and MUSHRA speech quality test results.*

Using least-square curve fitting, the following linear conversion between ACR and MUSHRA was obtained:

$$MOS_{MUSHRA} = 25.26 \cdot MOS_{ACR} - 20.38 \qquad (1)$$

Based on this relation, the ACR-results were converted to the MUSHRA-range. Note that the confidence intervals (CIs) for the mean ACR-ratings are scaled accordingly. As a result, Figure 3 shows the ratings from both tests including confidence

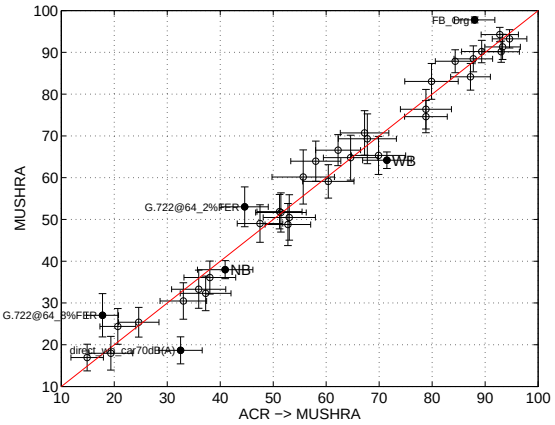



Figure 3: *Comparison of ACR and MUSHRA speech test results, with the MOS-results transformed to the MUSHRA-range using Equation (1).*

intervals. It can be seen that deviations from the correlation line seem to mainly occur for degradations perceptually different from the codecs mainly studied in this work: Cases of packet loss are rated more positive for MUSHRA than for ACR, while cases of background noise are rated better with ACR than with MUSHRA. One possible reason may be the direct comparison with a reference in case of MUSHRA, instead of the indirect referencing with previous stimuli and the subject's internal reference as in ACR. Similarly, it can be further noted that the clean FB-reference is not rated as the best condition in the ACR-test, while it is rated best in MUSHRA. Again, this effect may stem from an internal rather than explicit reference in case of ACR, yielding a more 'absolute' rather than a relative, and even directed rating.

An analysis of the confidence interval sizes for both methods is shown in Figure 4. As expected, the CIs are smallest towards the scale end-points, and largest in the middle of the scale. In case of MUSHRA, the two anchor conditions obtain the smallest CIs, due to the higher number of evaluations with the repetition in each set of 12 stimuli.

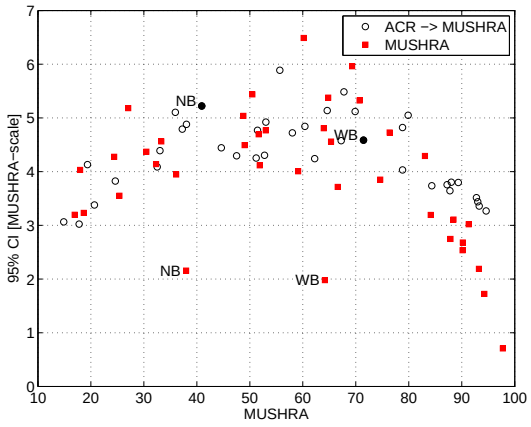



Figure 4: *Comparison of confidence intervals for ACR (expressed in MUSHRA units) and MUSHRA speech quality test.*

It can further be observed from Figure 4 that for speech, the two methods lead to very similar CIs, with more expressed differences only for the two hidden MUSHRA-anchors and at higher quality levels. The latter finding is supporting the hypothesis that the resolution of the MUSHRA method in the higher quality regime is better than for ACR. The parabolic shape of the CI-curves implies a non-linear contraction bias [8].

If the two scales were to be combined for the higher quality range – to enhance the ACR-resolution in this range – no significant rank-order changes are permissible. In spite of some visible rank-order differences, none of these was found to be statistically significant.

A univariate ANalysis Of VAriance (ANOVA) on the ACR- and MUSHRA-data was carried out, with the 'condition' as fixed and the 'test subject' as random factors. Note that for some of the conditions, the MUSHRA-ratings were not normally distributed, which limits the validity of the ANOVA (normal distribution tested using the Kolmogorov-Smirnov test). The ANOVA revealed highly significant effects for both 'condition' and 'subject', as well as for the interaction 'subject*condition', as expected based on the test results (ACR: Condition $F = 93.42$, $p \leq 0.001$; subject $F = 28.46$, $p \leq 0.001$; condition*subject $F = 2.18$, $p \leq 0.001$. MUSHRA: Condition $F = 102.381$, $p \leq 0.001$; subject $F = 29.31$, $p \leq 0.001$; condition*subject $F = 3.32$, $p \leq 0.001$). A post-hoc analysis of the data using a Bonferroni test showed that 586 of the possible 39*38/2=741 condition-pairs can be differentiated based on the ACR-test, and 611 in case of the MUSHRA test. For the speech tests, both the ANOVA and post-hoc results support the observation that MUSHRA is slightly but not much more sensitive than ACR.

## 4. Audio – test (b)

Three audio quality listening tests were carried-out to assess the quality of audio in the context of IPTV (Internet Protocol TeleVision) applications, as reported in [7]. The first test was a pure ACR test, the second and third test an ACR and a MUSHRA test, respectively. In the present paper, we will focus only on the first ACR test ("ACR-audio"), and the MUSHRA test ("MUSHRA-audio").

| Codec | Bitrates [kbit/s] | Ppl [%] |
|---|---|---|
| **MPEG-1 L II** | **96, 128, 192** | **0**, 1, 4, 8 |
| **MPEG-1 L III** | **64, 96, 128** | **0**, 1, 4, 8 |
| **MPEG-2 AAC** | **48, 64, 96, *128*** | **0**, 1, 4, 8 |
| **MPEG-4 HE-AAC v2** | **32, 48, 64, *96*** | **0**, 1, 4, 8 |

Table 2: Test conditions used in the audio tests: ACR (all conditions, apart from *italic*), and MUSHRA (highlighted in **bold**).

Note that as a matter of the initial context of the audio tests, the 11-point labeled but continuous scale according to [22] was used instead of the 5-point category scale. The ACR test comprised conditions with four different audio codecs at three different bitrates, and per codec-bitrate combination four conditions with uniform packet loss, in total 4*3*4=48 test conditions. The codecs, bitrates and loss percentages were chosen to represent current audiovisual distribution services, at the same time covering a sufficiently large quality range. The conditions are summarized in Table 2.

As can be seen from Table 2, the MUSHRA test comprised only the loss-free conditions, extended by a few lower-quality anchors to ensure that the test was balanced in terms of quality range and degradations, and two additional high-quality conditions. In order to nonetheless cover a wide range of quality-levels and -dimensions in both tests, a set of seven anchor conditions was used: (1) The clean reference, (2) NB-anchor, (3) WB-anchor, (4) signal-correlated, multiplicative white noise at 12 dB, (5) heavy coding: AAC at 48 kbit/s, (6) packet loss: MP2 with 8% packet loss, (7) low-quality speaker impulse response. In the MUSHRA test, the hidden reference (1), and the hidden anchors (2), (3) and (5) were repeated in each test set.

As source material, five audio samples from different genres were chosen: (i) Classical music, (ii) pop music, (iii) news (speech), (iv) advertisement (speech on soft classical music), (v) sports (soccer crowd). All stimuli had a duration of 16 s. This requirement is related with the respective IPTV video tests, where scenes with a length of 16 s were used. High quality loudspeakers were used for audio presentation. A large television screen was placed in front of the listeners between the speakers to give the impression of a standard TV viewing situation. For further details on the test design including test session lay-out please refer to [7].

### 4.1. Test results

As can be seen from Figure 5, showing the mean results over all subjects and contents, the correlation between MUSHRA and ACR is lower than in the case of the speech quality tests (test a). Further, a clustering of results in the low, medium and high quality ranges can be observed for both methods.

As for speech, a linear relation can be found to relate the results of the ACR and MUSHRA tests. However, due to the different quality ranges tested in the ACR and MUSHRA test, and the fact that an 11-point rather than a 5-point scale was used for ACR, the curve-fitting parameters are quite different from the ones found for speech (Eq. 1):

$$MOS_{MUSHRA} = 16.6 \cdot MOS_{ACR} - 46.1 \qquad (2)$$

This effect also is due to the range equalization bias, leading to

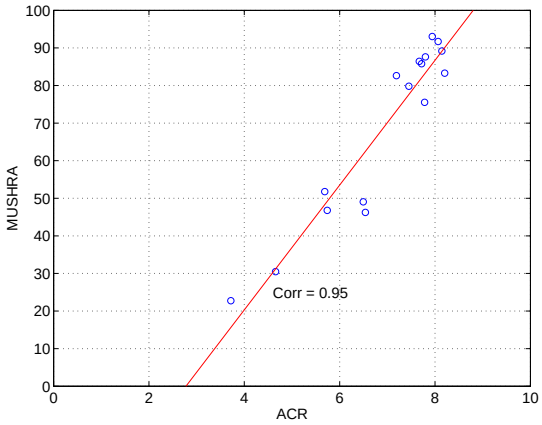



Figure 5: *Comparison of ACR and MUSHRA audio quality test results.*

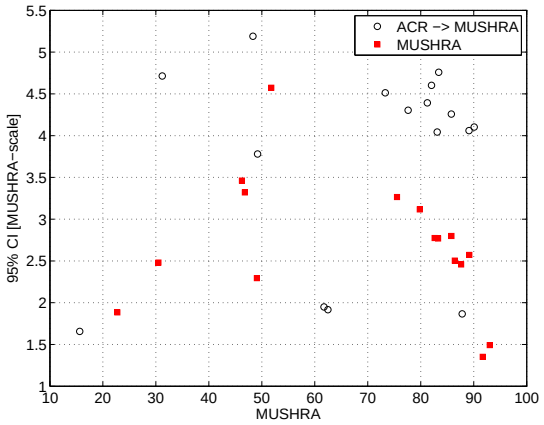



Figure 7: *Comparison of confidence intervals for ACR (expressed in MUSHRA units) and MUSHRA test.*

a more or less full coverage of the test scale, regardless of the actual ranges of qualities presented in the two tests (see [8]). This can well be observed in Figure 5, where no data points are shown for the lower range of the ACR-scale with $MOS < 4$, since the low-quality conditions were used only in the ACR-test and are thus not contained in the plot.

In Figure 6, the converted ACR-data according to Equation (2) is drawn against the MUSHRA-data. Obviously,

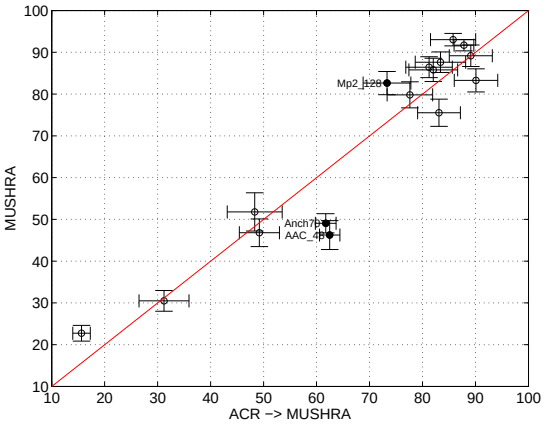



Figure 6: *Comparison of ACR and MUSHRA auio quality test results, with the MOS-results transformed to the MUSHRA-range using Equation (2).*

the relative confidence intervals for ACR are higher than for MUSHRA, as also shown in Figure 7. This is thought to be due, among other reasons, to the highly different audio contents presented in random order for the ACR-test, and as fixed set for each of the condition-sets in the MUSHRA test. In case of speech, no comparable effect is observed for ACR, since the variability between speakers does not compare with the variability between audio contents.

In spite of the less noisy results obtained for MUSHRA in the high-quality regime, the results for a number of the respective conditions are not statistically different, so that they do not allow the ACR-data to be refined accordingly. In addition, some possible rank-order reversals can be observed between MOS and MUSHRA. This may be due to the packet loss conditions being presented only in the ACR test. Hence, different perceptual quality dimensions and quality-levels were contained in the two tests, probably yielding differently directed foci of attention.

## 5. Conclusion and Outlook

The results of test (a) show an only slightly higher discrimination power of the MUSHRA- than of the ACR-method for speech, and a linear relation between the mean ratings. In Section 3, we have derived a respective transformation between the two scales. The results of the audio test (b) show a linear relation between the results obtained with MUSHRA and ACR, too. However, the relationship is less closely linear, which is assumed to be due to the different quality dimensions and quality ranges assessed in the ACR and MUSHRA audio tests. In principle, our results support the idea to convert MUSHRA-ratings into ACR-ratings to increase the resolution of ACR at higher quality levels: The smaller confidence intervals observed for the MUSHRA than for the ACR audio test are a prerequisite. However, the non-linear relation between the ACR- and MUSHRA-results in the high-quality regime makes it difficult to provide such a conversion with our data.

Future work is planned to further analyze whether the two scales can be complemented in certain quality regimes, by conducting a respective audio quality listening test. Additional tests will reflect the recent usage of the double-blind triple-stimulus with hidden reference method [9] for super-wideband and full-band speech quality tests reported in work-groups of ITU-T [11], and the relation of the results with those from ACR and MUSHRA. Future work will need to also address the relative importance of the test method and the employed listening panel: For BS-1116-type and MUSHRA-tests, experienced subjects are usually recommended, while for ACR-type tests naive subjects are used, to reflect a real-life situation.

## 6. References

[1] ITU–T Rec. G.107, *The E–Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union, CH–Geneva, 2009.

[2] A. Raake, *Speech Quality of VoIP – Assessment and Prediction*. Chichester, West Sussex, UK: John Wiley & Sons Ltd, 2006.

[3] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and

M. Wältermann, “Impairment factor framework for wideband speech codecs,” *IEEE Trans. Audio Speech and Language*, vol. 14, no. 6, pp. 1969–1976, 2006.

[4] M. Wältermann, I. Tucker, A. Raake, and S. Möller, “Extension of the e-model towards super-wideband speech transmission,” *In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, vol. March 14 - 19, USA–Dallas, 2010.

[5] I.-R. BS.1534-1, *Method for the Subjective Assessment of Intermediate Quality level of coding systems*, International Telecommunication Union, CH–Geneva, 2003.

[6] I.-T. R. P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, CH–Geneva, June 1996.

[7] B. Feiten, A. Raake, M.-N. Garcia, U. Wstenhagen, and J. Kroll, “Subjective quality evaluation of audio streaming applications on absolute and paired rating scales,” *In: Proc. 126th AES Convention*, vol. May 7 - 10, D–Munich, 2009.

[8] S. Zieliński, F. Rumsey, and S. Bech, “On some biases encountered in modern audio quality listening testsa review,” *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[9] I.-R. BS.1116-1, *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, International Telecommunication Union, CH–Geneva, 1997.

[10] I.-R. BS.1285, *Pre-selection Methods for the Subjective Assessment of Small Impairments in Audio Systems*, International Telecommunication Union, CH–Geneva, 1997.

[11] C. Quinquis and P. Usai, “Fullband conversational codec: What testing methodology?” *In: Proc. EUSIPCO 2008, CH–Lausanne*, 2008.

[12] ISO/IEC 14496-3:2009, *Information technology – Coding of audio-visual objects – Part 3: Audio*, International Organization for Standardization, CH–Geneva, 2009.

[13] 3GPP TS 26.290, *Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions*, 3rd Generation Partnership Project (3GPP), F–Sophia Antipolis, December 2009.

[14] J.-M. Valin, “The celt ultra-low delay audio codec,” *http://www.celt-codec.org*, 2010.

[15] ITU–T Rec. G.711, *Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunication Union, CH–Geneva, November 1988.

[16] ITU–T Rec. G.722.2, *Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB)*, International Telecommunication Union, CH–Geneva, January 2002.

[17] ITU–T Rec. G.722, *7 kHz Audio-Coding Within 64 kbit/s*, International Telecommunication Union, CH–Geneva, November 1988.

[18] ITU–T Rec. G.729 Annex A, *Reduced Complexity 8 kbit/s CS-ACELP Speech Codec*, International Telecommunication Union, CH–Geneva, November 1996.

[19] ITU–T Rec. G.722.1, *Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss*, International Telecommunication Union, CH–Geneva, May 2005.

[20] K. Vos, S. Jensen, and K. Soerensen, “Silk speech codec – draft-vos-silk-01,” *http://tools.ietf.org/html/draft-vos-silk-01, (08.03.2010)*, 2010.

[21] D. Gibbon, *EUROM.1 German Speech Database*. ESPRIT project 2589 report (SAM, Multi–Lingual Speech Input/Output Assessment, Methodology and Standardization), Universität Bielefeld, D–Bielefeld, 1992.

[22] ITU-T Rec. P.910, *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, Geneva, 1999.